22 | Bamberger Beiträge zur Linguistik

Lexical anaphora

A corpus-based typological study of referential choice

Nils Norman Schiborr

University of Bamberg Press

Bamberger Beiträge zur Linguistik

hg. von Martin Haase, Thomas Becker (†), Geoffrey Haig,
Sebastian Kempgen, Manfred Krug
und Patrizia Noel Aziz Hanna

Band 22

University
of Bamberg
Press
2023

Lexical anaphora

A corpus-based typological study of referential choice

Nils Norman Schiborr

University
of Bamberg
Press
2023

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de/ abrufbar.

Diese Arbeit hat der Fakultät Geistes- und Kulturwissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.
Gutachter: Prof. Dr. Geoffrey Haig
Gutachter: Dr. Stefan Schnell
Tag der mündlichen Prüfung: 09.07.2021

Herstellung und Druck: docupoint, Magdeburg
Umschlaggestaltung: University of Bamberg Press
Umschlagbild: Nils Norman Schiborr
University of Bamberg Press, Bamberg 2023
https://www.uni-bamberg.de/ubp/

ISSN: 2190-3298 (Print) eISSN: 2750-7726 (Online)
ISBN: 978-3-86309-938-1 (Print) eISBN: 978-3-86309-939-8 (Online)

URN: urn:nbn:de:bvb:473-irb-597734
DOI: https://doi.org/10.20378/irb-59773

For my mother
and my little niece

May you have a future
worth looking forward to

Acknowledgements

I am eternally grateful to my supervisors, longtime collaborators, and dearest allies, Geoffrey Haig and Stefan Schnell. None of this would have been possible without their unwavering support.

I would also like to extend my thanks to Manfred Krug, who first introduced me to the study of language, and the many other teachers, mentors, and colleagues I have learnt from over the years, from the earliest days of my life to the present and into the future. Thank you all.

Lastly, I would also like to acknowledge the support of the DFG, who funded a large part of the research presented in this book (Grant no. HA 2839/6, awarded to Geoffrey Haig and Stefan Schnell).

Contents

1 | Introduction

Reference is an integral part of human language, and indeed of most systems of communication. When speakers refer to entities or events, they are tasked with making a series of decisions – which entity to refer to, how to embed it into syntactic structures, and which type of expression to use for the reference – in a process that is highly routinized. This study concerns itself with the last of these decisions, the selection of the linguistic exponents of reference, or referential choice. For any given entity that a speaker wishes to refer to, they must choose from a vast array of potential options for referring expressions:

(1) *Mother, she started the shop up herself,*
and ø applied for the post office
and ø got it. [mc_english_kent02_0739]

For any of the references in (1), the speaker had to select between either a pronominal noun phrase (*she*), a noun phrase headed by a lexical noun (*the post office*), or simply leave matters unexpressed (i.e. zero). These choices are not made randomly, but are subject to numerous soft and hard constraints (Huang 2000a; Ariel 1990; among others). For instance, the use of zero anaphors is infamously limited in English, with tight sequences of co-referential mentions across adjacent clauses as in (1) being the most common circumstances where they do occur with more than sporadic frequency. Referring expressions differ chiefly in terms of their information content and how specifically they identify the entity that is being referred to, and as is evident from the example

above, most references in natural discourse tend to be highly underspecified (Gundel 2010: 149).

Since the 1970s, the prevailing theories of reference have connected this form selection process to cognitive constraints on the interlocutors' memory and attentional states (Chafe 1976; Givón 1983a; Ariel 1990; Gundel et al. 1993a; Grosz et al. 1995; etc.). According to this view, speakers' choices depend on the salience of entities in discourse, which ebbs and tides as the discourse progresses. Speakers are hence tasked with selecting whichever expression is the most situationally appropriate: Most theories of reference contend that the use of an insufficiently informative expression is infelicitous if the speaker cannot assume that the addressee is already familiar with and aware of the referenced entity, as mentioned above; conversely, if the target entity is assumed to occupy a central position in the addressee's mind, the use of an expression that is too explicit is inefficient at best and potentially misleading at worst.

While the underlying mechanisms of selecting referring expressions are claimed to possess a degree of universality, the actual patterns of realization can differ substantially between languages (Torres Cacoullos & Travis 2019; Stoll & Bickel 2009). This has lead to the impression of stark contrasts between languages. Much attention has consequently been given to the choice between overt forms and zero (e.g. Travis & Lindstrom 2016), and here quite considerable cross-linguistic differences do obtain, as reflected in the traditional typology of pro-drop and non-pro-drop languages and work on the null subject parameter (Perlmutter 1971; Neeleman & Szendröi 2007; etc.).

But the focus on overtness in much of the literature on referential choice is in fact somewhat misleading. An alternative perspectives posits that the most fundamental referential choice is instead a binary one between reduced forms (pronouns, zero) and lexical forms (full NPs, Kibrik 2011; Torres Cacoullos & Travis 2019; Schiborr 2017). In this view, the question of referential choice in essence comes down to the reasons a speaker might have "for using a word that is leaner in semantic content rather than one that is fuller and vice versa" at a given point in the discourse (Bolinger 1979: 290), or put more simply, "why not lexical NPs?" (cf. Ariel 1990: 58). While the distinction between lexical and non-lexical expression is very much at the heart of most functionalist models of referential choice, there is a comparative lack of studies that target the selection criteria for lexical expressions specifically and in isolation (see e.g. Stoll & Bickel 2009; Kibrik et al. 2013, 2016), rather than as part of a

larger scale of referential expressions that also includes zero (e.g. Ariel 1990).

This study aims to address three key aspects of the selection of lexical expressions in natural discourse: One, how frequent are lexical anaphors across languages? Two, what are the selection criteria for lexical anaphors? And three, do the same criteria also determine the choice between specific types of lexical expression, such as proper names? It assumes a typological, variationist perspective, using spoken corpus data from ten languages, most of which are small and understudied. This puts it in stark contrast with most approaches to anaphora and referential choice, which tend to focus on specific well-represented languages such as English, Spanish, or Russian, and mostly or exclusively make use of written data (e.g. Ariel 1990; Kibrik et al. 2016; etc.). The data leveraged for this study are taken from the Multi-CAST collection (Haig & Schnell 2015), a freely available database of spoken language corpora, which this study was in part developed in parallel with.

An essential aspect of examing speakers' referential choices from a discourse-oriented perspective is identifying the properties of the contexts in which they occur. These properties should be quantifiable, operationalizable, and ideally founded on basic linguistic categories that can be measured without too great a reliance on theoretical preconceptions and assumptions. This study aims to find correspondences with structures that have been identified as salient in the literature (e.g. topicality, accessibility, recency, etc.) through the programmatic analysis of the data and the annotations applied to them, including the automatic calculation of anaphoric distances, mention frequencies, formal properties of antecedents, various semantic and syntactic properties, and more.

This book is organized as follows: Chapter 2 provides an outline of the research background, starting with the notions of discourse anaphora and referential choice, before moving on to the aforementioned cognitive theories of reference and concluding with quantitative approaches to the issue, including those in computational linguistics and natural language processing. Chapter 3 then briefly introduces the emerging field of corpus-based typology and the Multi-CAST project, and offers a description of the corpus data and its annotations. Chapter 4 deals with methodological issues relevant to the analyses presented in the subsequent chapters, including definitions of the tested factors and any categories pertinent to them.

Chapter 5 addresses the first of the research questions listed above, examining the overall rates of lexical expression in discourse across the ten cor-

pora. It differentiates these rates by syntactic position, and pays special attention to subject anaphors; lexical subjects are known to be quite rare, and hence constitute what Francis et al. (1999: 10) call a "highly anomalous" class. The following two chapters then explore the specific circumstances under which speakers select lexical expressions. This is achieved by testing a set of twelve discourse-contextual and referent-inherent factors for their association with lexicality. Chapter 6 first examines these associations individually and further explores a number of key interactions, Chapter 7 then describes a pair of predictive multifactorial models fit to the corpus data using the gradient boosting algorithm (Friedman et al. 2000; Friedman 2001, 2002), and Chapter 8 tests whether the same factors can also be used to predict the selection of specific types of lexical expressions – proper names, simple and more elaborate lexical NPs, and lexical NPs with demonstrative modifiers.

Chapter 9 then brings the findings from the previous chapters together. While part of this study is about evaluating how well the multifactorial models can predict the selection of lexical expressions, making accurate predictions is not its primary aim; instead, the models are leveraged chiefly for identifying and then explaining patterns of usage: In this vein, Section 9.1 discusses which of the tested factors appear to matter the most, under which circumstances, and in combination with which other factors. Here the central position of anaphoric distance is noted, and two extreme contexts in which multiple convergent factors align – one at very short, one at long distances – are outlined, the former of which in particular plays a key role in establishing and maintaining the coherence of discourse (Givón 2017). Section 9.2 then gives an account of the overall rates of lexical expression observed in data, using the aforementioned extrema as a starting point and leading to an explanation centered around considerations of the cohesiveness and information content of discourse, and the idea that perspective shifts in discourse – and hence lexical subject expression – tend towards a cross-linguistically stable median. Lastly, Chapter 10 provides summary of the key findings of this study and their implications, as well as an outline of a number of potential avenues for future research that further develop the methods and ideas presented in this book.

2 | Research background

This chapter provides background on the notions of anaphora, informativity, and referential choice (Section 2.1), with a focus on theories of reference that consider the expression of referential information in discourse to be largely a matter of cognitive mechanisms and constraints (Section 2.2). It also touches on a number of computational approaches to the matter, especially those that constitute a synthesis of cognitively-oriented considerations and the methods of natural language processing (Section 2.3). Additionally, a section is dedicated to listing a number of key factors that have been identified as affecting the selection of referring expressions (Section 2.4), and another to the status of lexical expression in various theories of reference, including various subtypes such as proper names (Section 2.5). Lastly, this chapter also provides an outline of the research questions that guide this study (Section 2.6).

2.1 | Anaphora and referential choice

In natural discourse, speakers commonly make repeated references to one and the same entity or entities. In fact, this is an essential aspect of any narrative, as the recurrence of the same entities ties individual events together (cf. Givón 1983b):

(2) *Well, we started doing this here bell ringing.*
There was a man$_i$ by the name of Conell,
he$_i$ volunteered to learn us. […]
He$_i$ told one boy$_j$ to pull a little harder,
well he$_j$ pulled a little harder,
and he$_j$ slammed the bell off, and it broke the stay,
and he$_j$ went up along with it.
Down he$_j$ come. [mc_english_kent03_0491-0492;0500-0501]

This phenomenon of repeated reference is called anaphora, and its instantiations anaphors.[1] As the pronouns in (2) demonstrate, the interpretation of anaphors tends to be dependent on the prior mentions of the same entity, their antecedents (Kibrik 2011: 35; Huang 2000b). But the term "anaphora" should not be taken in the literal sense of 'carrying back', which might imply that speakers actively search through the preceding discourse to find a matching antecedent. As Garnham (2001) and Kibrik (2011) note, this is psychologically implausible, as speakers do not maintain a detailed record of past utterances. Instead, if an interlocutor "makes a search through anything when resolving an anaphor, it is [their] cognitive system" (Kibrik 2011: 49).[2] In this view, anaphors link up with the mental representations of entities in the interlocutors' memory (Johnson-Laird 1983; Jackendoff 2002), what Karttunen (1976) calls "discourse referents", rather than linguistic expressions. Anaphors are hence the concrete instantiations of abstract discourse referents, serving as largely underspecified pointers to specific mental representations (Kibrik 2011: 31; cf. Givón 2001: 438–439). Strictly speaking, the notion of anaphora does not capture the first time an entity is referred to in a text, that is the introduction of a new referent into discourse (e.g. *a man* and *one boy* above), but only subsequent mentions. However, the distinction between introductory and anaphoric references is not a strictly categorical one (Kibrik 2011: 49), as it can be blurred, for instance, in the case of references to the

1 The term anaphor is sometimes only applied to pronominal references (and/or zero anaphors), but I use it here for any repeated reference, irrespective of form.

2 But see operationalizations of co-reference in certain areas of natural language processing, discussed further below.

speaker and interlocutor or shared world knowledge, or in the case of long intervals between mentions.

Studies of anaphora revolve around the specific relationship between anaphor and antecedent, and tend to largely fall into one of two areas of investigation depending on where said antecedent is located. If anaphor and antecedent are found in the same clause, their relationship is considered to be a syntactic issue, as it is primarily subject to grammatical constraints (cf. binding principles, Chomsky 1982, 1993). If they are in different clauses, it is instead a matter of pragmatics, as the interpretation of the anaphor depends on the interpretation of its antecedent (Levinson 1987, 1991; but see Dixon 1979). Different constraints are at play for anaphors within and across clauses, and a holistic approach to anaphora has to account for both syntactic and pragmatic dimensions (Huang 1995: 1112; Ariel 1990). Most studies, however, are content to focus on one or the other, the present study being no exception; the remainder of this chapter hence concerns itself with the pragmatic, discourse-oriented side of the matter.

One of the central research issues in discourse analysis is how discourse referents receive linguistic expression. Reference is effected through the use of various types of referring expressions, such as lexical and pronominal NPs. Referring expressions chiefly vary in terms of their semantic content (Levinson 1987, 1991; Huang 2000a; Ariel 1990), and are commonly arranged into hierarchies of the kind in (3),

(3)

zero	<	pronouns	<	lexical NPs	<	complex lexical NPs
ø		*she*		*the woman*		*the woman in the green coat*

where from left to right, expressions become more content-rich, more specifically referring, as well as less phonologically attenuated and structurally more complex. Referential hierarchies (e.g. Ariel 1990; Givón 1983a; Chafe 1976; etc.) generally include substantially more detail than the simple scale in (3), often also involving other formal dimensions, such as proper names and demonstratives. But as Kibrik (2011: 42) notes, the more detailed a scale is, the less likely it is to be universally applicable, despite frequent claims to the opposite (e.g. Ariel 1990: 76–92). Instead, so Kibrik argues, the most fundamental dichotomy is between lexical and non-lexical expressions (cf. also Schiborr 2017; Krahmer & van Deemter 2012: 204).

Crucially, studies on anaphora are not interested in actual semantic content, but rather in the relative differences in information content encoded in expressions. While the information content of an expression is not dependent on the circumstances of the surrounding discourse (cf. Chafe 1976), how effective it is in resolving ambiguities and identifying the intended referent – its informativity (or "surprisal" in terms of information theory) – is almost entirely contextual (Stoll & Bickel 2009: 544). Informativity is largely a function of distinctiveness, and is hence tied to the predictability (or inferability) of the information encoded. In practical terms, this means that the most appropriate referring expression in a given context is one that meets "the immediate communicative needs of the interlocutors" (Féry & Krifka 2008: 123) as per Grice's (1975) maxim of manner, that is one that provides concise and clear information to the addresse, neither too much or too little. For instance, a pronominal reference to an entity not previously mentioned in the current discourse would likely not provide the addressee with sufficient information to identify the intended referent. Conversely, free pronouns are frequently omitted in favour of zero expression if their referents are readily identifiable (Van Valin & LaPolla 1997), but exactly how commonly this occurs is subject to substantial cross-linguistic variation (i.e. pro-drop, Perlmutter 1971; see also Torres Cacoullos & Travis 2019).

Considerations of this kind are at the core of the notion of information structure (Chafe 1976; Lambrecht 1994; Polanyi 1995; Polanyi et al. 2003; Ward & Birner 2004; Zimmermann & Féry 2010; Féry et al. 2006; Krifka & Musan 2012; Féry & Ishihara 2014; among many others), whose typological relevance is well supported (e.g. Fiedler & Schwarz 2010; Fernandez-Vest & Van Valin 2016; Song 2017; Riesberg et al. 2018; Adamou et al. 2018; etc.). Complementary evidence from psycholinguistic studies lends further support, showing that respondents' performance in various experimental tasks improves when given additional cues in ambigious environments, and worsens when provided with redundant cues in unambigious environments (Bates & MacWhinney 1989; Kail 1989; Caballero & Kapatsinski 2015; etc.); see also Gordon et al.'s (1993) "repeated name penalty".

The informativity of the basic types of expressions varies between languages (Stoll & Bickel 2009: 544); this is especially true for pronouns, which differ substantially in the amount of grammatical and conceptual information they carry. The informativity of pronouns can range from very low (e.g. those only encoding person and number, such as English *they*) to close that of lex-

ical NPs (e.g. those also encoding gender, degree of politeness, kinship relations, distance from the speaker, relative elevation, etc.), in the latter case lacking only in lexical content. There may also be pragmatic dimensions to the informativity of an expression; for instance, Japanese pronouns carry presuppositions of familiarity between the speaker and referent, and may hence narrow down potential antecedents to those with appropriate relationships (cf. Hinds 1978).

There are two classes of approaches to reference, one taking the perspective of the speaker (i.e. production) and one the addressee's (i.e. comprehension; MacDonald 2013; McDaniel et al. 2015), though they are not always fully differentiated (Nariyama 2003: 42). Approaches focused on the comprehension side tend to chart discourse-spanning processes, for instance addressees' awareness of all hitherto mentioned referents (reference tracking, Heath 1975; Du Bois 1980; Foley & Van Valin 1984; Comrie 1989; Fox 1996; Van Valin & LaPolla 1997; Nariyama 2003; van Gijn et al. 2014; etc.) or conversely their anticipation of the entire subsequent text (Himmelmann 1996; Lichtenberk 1996). Others concern themselves with the mechanisms of determining whether two referring expressions are co-referential or not, that is whether they denote the same entity (anaphora resolution in NLP, Mitkov 2000, 2002; Mitkov & Barbu 2001; Tetreault & Allen 2003; Tetreault 2005; etc.).

Conversely, the production of referring expressions is focused on individual instances of reference, and is best captured by the notion of referential choice (Clancy 1980; Kibrik 2011; Holler & Suckow 2016; etc.) Referential choice is about the selection of an appropriate form of reference in a given context, but says nothing about the selection of the referent itself. In other words, referential choices are not directly related to the likelihood with which a specific referent is mentioned, as reference and referential choice are related, but separate processes (Gatt et al. 2014: 903). Kibrik (2000: 72) explicitly distinguishes the notions of discourse anaphora and referential choice, with the latter also encompassing the introductions of new referents into discourse, rather than just anaphoric mentions. While I agree that this differentiation is sensible, I here use the two terms interchangeably, with the meaning of the former (as per Kibrik's definition, i.e. restricted to anaphors).

Early approaches to referential choice and related concepts (e.g. Chafe 1976, 1994; Givón 1983a; Fox 1987b; Tomlin 1987b; Ariel 1990; Gundel et al. (1993a); etc.) all assume some kind of cognitive property of referents that

informs speakers' choices, such as "givenness", "topicality", "accessiblity", among many others. According to these theories, the mechanisms that guide referential choices are primarily a matter of the properties of the surrounding discourse, which shape interlocutors' discourse models. These cognitively-oriented theories of reference will be discussed in detail in next the section. There is also a line of research that links referential choices to prosodic features (e.g. Fretheim 1996; Mithun 1996; Watson & Gibson 2004; Baumann 2006; Torres Cacoullos & Travis 2019), as well as on the effect of phonological priming (Tanenhaus et al. 1985); these fall outside the purview of this chapter.

2.2 | Cognitive theories of reference

Since the 1970s, a long line of research in the tradition of Chafe (1976) has sought to explain the structure of discourse by appealing to cognitive constraints on interlocutors' working memory. This line of thinking rests on the assumption that the working memory has limited capacity, and so can only keep track of a certain number of referents concurrently. Discourse structure, it is argued, is hence shaped primarily by the imposition of cognitive constraints and rules of reference.

These cognitive theories of reference (e.g. Chafe 1976; Givón 1983a; Ariel 1990; but also Gundel et al. 1993a; etc.) assume a categorical form–function mapping of referential expressions to cognitive states. They predict that phonologically heavier, more informative expressions are used in favour of reduced expressions in contexts where the speaker does not assume that the addressee can unambigiously identify the intended participant in a particular role, while the inverse is true for the selection of reduced over lexical expressions. In particular, zero anaphors are used for the most readily identifiable and lexical NPs for the least identifiable referents, while pronouns and demonstratives fall inbetween (Gundel et al. 1993a: 285; Ariel 2001: 31).

In this view, speakers' assessments of referent identifiability is generally associated with the degree of "accessibility" of the underlying mental representations in the speaker's memory. Higher accessibility results in greater ease of retrieval. This general notion bears many different names in the literature: "Activation" and "givenness" in Chafe (1976), "topicality" in Givón (1983a), "accessibility" in Ariel (1990) and Bock & Warren (1985) (not to be confused

with the syntactic "NP accessibility" of Keenan & Comrie 1977), "salience" in Osgood (1971) and Sridhar (1988), "focus of attention" in Grosz & Sidner (1986) and Grosz et al. (1995), as well as the various "cognitive states" in Gundel et al. (1993a), and others. Ultimately however, they are all conceptualizations of (certain aspects of) the same abstracted cognitive property, with mostly overlapping definitions, but often slightly different assumptions. Givenness and topicality, for instance, emphasize the importance of the (local) information structure of discourse: If a referent is the topic of the preceding clause, then it is likely to be highly activated and hence readily accessible (Givón 1983a). Others, such as focus of attention and the cognitive states of Gundel et al. (1993a), place the emphasis on cognitive capacity: A continuously mentioned referent is more accessible, as it is likely to be deliberately maintained in memory (Foraker & McElree 2007). I will here use "accessibility" as an umbrella term for the bundle of concepts listed above, that is as the broadly defined notion of 'ease of retrieval of referential information', rather than specifically in terms of Ariel (1990) and accessibility theory, unless explicitly pointed out as such. Likewise, I will use of the term "salience" to refer to the accessibility-affecting properties of the referent itself, rather than merely as a variant conception of accessibility, as in Osgood (1971) and Sridhar (1988).

More explicit forms are also used for introducing new referents into discourse, which are then generally followed by less explicit forms for subsequent mentions (Givón 1983a), resulting in a steady decrease in informativity from one mention to the next (cf. Cameron & Flores-Ferrán 2004: 50). Once a maximum level of accessibility is achieved, successive mentions continue to employ the least explicit form available. If the accessibility of the referent drops, for instance because is has not been mentioned for a while, a more explicit form is used once again. Use of expressions that are not appropriate to the local discourse context can lead to interpretations of non-co-referentiality (if highly explicit), or else to ambiguity and an inability to correctly identify the intended referent (if not).

Research on discourse anaphora has mainly focused on the influence of semantic, linguistic, and discourse-related factors such as animacy, recency, syntactic function and functional continuity (or lack thereof), the presence of competing referents (requiring disambiguation through more specifically referring expressions), and more (cf. Brown & Yule 1983). These properties can either increase or decrease referent accessibility, and certain contexts can over-

ride the influence of otherwise more influential properties, for instance for ambiguity avoidance (Arnold et al. 2000; Arnold & Griffin 2007; Fukumura et al. 2010). Crucially, it is generally agreed that no single factor determines speakers' choices, but that instead multiple factors converge on specific outcomes; referential choice hence "belongs to a large family of multifactorial processes, generally characteristic of language production" Kibrik et al. (2016: 2). The various factors that have been identified as influencing referential choices and their treatment in the literature will be discussed in Section 2.4 below.

Aside from production-oriented constraints, the selection of referring expressions has traditionally been considered to be a chiefly addressee-oriented process ("recipient design"; Clancy 1980; Clark et al. 1983; Contemori & Dussias 2016). Speakers are generally believed to choose expressions so as to aid the addressee in retrieving the intended referent, that is, design the structure of their discourse around the needs of the addressee (Ariel 1990; Prince 1981b; Chafe 1987, 1994; as per Grice's cooperative principles, Grice 1975). As the linguistic context is available to both speaker and addressee, speakers' choices are assumed to align with the discourse information shared by both interlocutors. More recent work has cast doubt on this view, suggesting instead that speakers pay attention primarily to the needs of their own discourse model rather than the addressee's (Fukumura & van Gompel 2009; Contemori & Dussias 2016: 5; cf. also Schnell et al. 2021b, as well as evidence from inner-speech and child language, e.g. Piaget 1955). Fukumura & van Gompel (2012), for instance, observe that while speakers produce a higher rate of pronominal (vs. lexical) references when the antecedent is in preceding sentence, as would be expected by accessibility-oriented frameworks, they do so even when aware that the addressee did not hear said antecedent.

Lastly, it should be noted that in the context of the frameworks of reference discussed in this section, accessibility is strictly about the status of non-linguistic mental representations in the speakers' short-term memory, and as such is distinct from the accessibility of specific lexical items in speakers' mental lexicon (cf. Arnold 2010; Riester & Baumann 2017), the selection of which constitutes another dimension of referential choice. Bock & Warren (1985: 50–52) hence distinguish conceptual accessibility, that is "the ease with which the mental representation of some potential referent can be activated in or retrieved from memory" from lexical accessibility, "the ease with which the representations of word forms can be recovered from memory".

The remainder of this section outlines the most important cognitively-oriented approaches to referential choice (Sections 2.2.1–2.2.4) as well as a number of related concepts and frameworks (Section 2.2.5).

2.2.1 | Activation states

Chafe's (1976) seminal study links the choice of referring expressions to the cognitive status of the entity being referred to anaphorically (see also Chafe 1994). Entities are associated with certain activation states in a speaker's current discourse model, which are primarily determined by whether and how recently said entities have been mentioned previously: "Active" entities have been recently referred to, and so are assumed by the speaker to reside in the addressee's current focus of attention; "semi-active" entities are conversely only present in the periphery of the addressee's consciousness, and as such are part of what Chafe (1976: 33–34) calls their "background awareness". The state of an entity may change as discourse progresses, so that active entities that have not been mentioned for a while eventually drift into a semi-active state from lack of sustained attention. Entities that are invoked inferentially from the textual context, or are otherwise known from the physical surroundings or shared world knowledge, can be (semi-)active without prior mention (Chafe 1996). Lastly, "inactive" entities are those that are neither active nor semi-active, that is those assumed to fall outside of the addressee's field of awareness.

Crucially, the three states are not strict categorical classifications, but rather thresholds on a continuum of activation that map unto discrete types of expression on the linguistic level, so that which activation threshold a referent has crossed determines the form of expression used for it. Active entities receive reduced forms, either pronouns or as zero anaphors, due to their central position in addressee's attention; semi-active entities are verbalized with fairly simple definite lexical noun phrases, while inactive entities require more informative definite lexical NPs or indefinite lexical NPs to correctly identify. According to Chafe (1976), then, there is an inverse relationship between the activation of a referent and the informativity of the expression used to refer to it, so that the more active a referent is at the moment of being referred to, the less coding material it requires, and vice versa. Kibrik et al. (2016: 2) reformulates this relationship into the following basic "law of referential choice":

(4) *a.* If the referent's activation in the speaker's working memory is high, use a reduced referential device.

b. If the referent's activation in the speaker's working memory is low, use a lexically full referential device.

This is the most fundamental claim of Chafe's theory of referential choice, and one shared by essentially all of the cognition-based approaches discussed in this section, whichever label they may use for their particular variant of activation.

2.2.2 | Givenness

The activation distinctions drawn by Chafe (1976) further correspond to a scale of givenness, which determines the "activation cost" a speaker has to invest in order to elevate an entity to an active state (Chafe 1974; 1994; see also Lambrecht 1994 for a later development of Chafe's model, and MacWhinney & Bates 1978 for earlier ground-laying work). Said cost is lower for entities that are already active or semi-active, which are "given", and higher for those that are inactive, or "new". Any information the speaker assumes they are introducing into the addressee's awareness through reference is hence new. The classification by newness is largely a binary one; any entity introduced into the current discourse for the first time is considered new, irrespective of how known or familiar to the addressee it may otherwise be. For an entity to be considered given information, it must have either been introduced into discourse previously or be part of the immediate physical surroundings, though as Prince (1981b: 229) notes, the status of inferentially related information can be ambigious in terms of givenness.

Chafe's activation-based account is only one of numerous conceptions of givenness, with distinct but often overlapping definitions. A synergetic approach is that of Clark & Haviland (1977) and Prince (1981b), for whom givenness (or "assumed familiarity" in Prince's terminology) is chiefly a matter of identifiability; here, a given entity is one the speaker assumes the addressee to maintain a mental representation of, even if said representation is not actively being maintained.

Givenness in both senses is generally marked by morphosyntactic and prosodic means, with identifiability roughly corresponding to the grammatical

category of definiteness Lambrecht (1994: 87). According to Chafe (1976), new information cannot be pronominalized, as reduced expressions are not informative enough to adequately introduce referents into discourse; as such, new referents are generally expressed via lexical NPs (cf. Ariel 1988). More recent studies have cast doubt on the notion of a strictly categorical link between newness and lexicality (e.g. Stoll & Bickel 2009), though the association is still a strong one.

The activation cost associated with introducing new information is claimed to be particularly high, so that introductions place substantial demands on discourse processing and comprehension (cf. Du Bois 2003b: 38). Chafe (1987: 32) hence proposes a "one-new-concept-at-a-time constraint", which is related to the Gricean maxim of quantity and similar quantity constraints (cf. e.g. Engelhardt et al. 2006). According to this constraint, speakers will structure discourse in such a way as to spread out new information as efficiently as possible, ideally one piece per clause (but see Schnell et al. 2021b for a contrary account). It has been argued that the introductions consequently involve different mechanisms than anaphoric references to given referents, as speakers deploy certain strategies adapted to the task (e.g. Du Bois 1987b; Lambrecht 1994; Durie 2003).

2.2.3 | Topic continuity

Givón (1983b, 1983c, and other contributions in Givón 1983a) seeks to explain the mechanisms underlying the selection of referring expressions as primarily motivated by considerations of discourse coherence (or "cohesion"), which is attained and maintained primarily through topic continuity (cf. also Halliday & Hasan 1976). According to Givón, coherent discourse is structured along strings of repeated references to the same topical entity or entities. As such, topic continuity is strongly tied to predictability, with the assumption being that referents in prominent positions in the immediately preceding discourse are expected to persist in the subsequent discourse.

Topic continuity is hence largely a function of recency and syntactic prominence (Givón 1995; see Sections 2.4.2.1 and 2.4.2.2): A continuous (or "accessible") topic is one that was recently mentioned and is likely to be mentioned again, and tends to be expressed via reduced expression (pronouns, zero anaphors). Discontinuous topics are conversely more likely to receive lexical forms, being either newly introduced to discourse (see Section 2.2.2

above), or otherwise not persistent. In addition to recency, syntactic prominence also plays an important role in determining discourse cohesion, with subjects, ceteris paribus, being more continuous than objects and oblique arguments (Givón 2017: 98). As such, the context with the highest degree of discourse coherence is delineated by tightly linked anaphoric chains, in which the same referent is mentioned in successive clauses, especially if in the same position (i.e. same-role chains). In English, this is notably the only circumstance in which zero anaphors are frequent. Crucially, Givón (2017) considers topicality and accessibility from referential continuity to be distinct properties, though they may appear to be strongly associated: "Most conspicuously, a referring indefinite NP is by definition maximally discontinuous anaphorically, but may be either highly topical/important cataphorically or non-topical/unimportant" (Givón 2017: 98).

Givón (1983b, 1976) conceives of topicality not as a categorical, but rather a continuous property. While the associations between topicality and discrete form choices are necessarily deterministic – though only relative to other forms within a given language, and hence may vary substantially cross-linguistically – the same broad associations hold true across languages: Zero anaphors always indicate higher topicality than pronouns, which in turn indicate higher topicality than lexical NPs. However, Givón (1983b) acknowledges that factors other than topicality also play a role in determining referential choices, but argues that the concrete, measurable factors that determine topicality (i.e. recency, predictability, etc.) can already explain a significant part of the variation between referring expressions.

Notably, the various contributions in Givón (1983a) are careful to account for cross-linguistic variation from the start, and are furthermore based at least in part on analyses of spoken connected discourse, rather than exclusively on constructed or elicited example data, like many other approaches discussed in this chapter. Later work on discourse coherence has refined and expanded on these ideas (e.g. Gordon & Chan 1994; Gernsbacher & Givón 1995; Kehler 2002; 2004; Christiansen 2011, among many others).[3] The interplay of coher-

3 The notion of coherence has also been extended to apply to cognitive processes in general (Givón 1995) and beyond (Givón 2020); we will here focus specifically on work relevant to observable linguistic structures, however.

ence and predictability in particular has been formalized in centering theory (Grosz et al. 1983), which we will briefly discuss below in Section 2.2.5.2.

2.2.4 | Accessibility theory

Where Chafe (1976) and Givón (1983a) seek to explain referential choices in terms of a single factor – activation, topicality – later developments such as accessibility theory (Ariel 1988; 1990; 2001) instead contend that speakers' decisions are guided by complex interplay of factors, whose relative influence is contextual; as such, so Ariel argues, the selection of referring expressions results from the interaction of multiple soft constraints converging on a specific choice of form.

Accessibility theory considers all referring expressions to function as so-called "accessibility markers", which code instructions for the identification of the intended discourse referent through their relative informativity, specificity of reference, and phonological attenuation. In a given discourse context, speakers select the referring expression that best reflects the "accessibility" of the underlying referent. Accessibility is influenced by a variety of discourse-related, structural, and semantic factors, such as topicality, recency, frequency, and predictability, as well as grammatical role and inherent salience, which is in turn strongly influenced by humanness. In this way, accessibility theory is very much a direct development of earlier accounts of referential choice, such as Givón (1983a), Sanford & Garrod (1981), and others.

Accessibility is considered to be a property of discourse referents that shifts as discourse progresses. At any point in discourse, the mental representations that underlie referents are not equally activated (cf. Chafe 1976; Section 2.2.1 above): Some may be highly activated, some less so, while others, though known, may not be activated at all, as they have yet to be introduced into discourse. Speakers may make reference to any entity, irrespective of its activation state, but must provide sufficient information for the addressee to be correctly and efficiently identify the intended entity.

According to Ariel (1990), this is achieved through the combination of the semantic content of the referring expression used and the degree of "accessibility" of the target entity implied by said expression. An expression with comparatively low informativity, for instance a pronoun, is taken to signal to the addresse that the intended referent is a readily accessible one, usually a topic

or otherwise highly salient referent such as one that that has been recently mentioned, and which an expression with low informativity is sufficient to identify. Conversely, a highly distinctive and specific expression, such as a lexical noun phrase with high information content, not only signals that the target referent is not easily accessible – that is not among those currently assumed active in the addressee's working memory – but also provides the necessary information for its identification through its semantics. Addressees are expected to evaluate each referring expression against all potential discourse referents known to them, and select the most likely target on the basis of their discourse model Ariel (1990: 56). Speakers are hence tasked with helping addressees do so by selecting the most appropriate referring expressions given the situation, that is, engage in audience design (see discussion above). Precisely which expression is the most appropriate for a given referent is dependent on context, and more or less informative expressions can be appropriate depending on the exact circumstances surrounding their use. For instance, an otherwise highly accessible referent might be referred to via a more informative expression than might otherwise be necessary, for instance in situations where there are other highly accessible referents with which the intended one might be confused, that is, if there is competition between candidate antecedents.

However, as Ariel (1990) notes, certain expressions inherently indicate a higher or lower degree of accessibility than others, irrespective of the context of their use, which allows types of referring expressions to be graded on a scale of relative "accessibility marking". The original accessibility marking scale (or "accessibility hierarchy") for English (Ariel 1990: 73 ex. 1) is reproduced here in Figure 2.1 with minimally adjusted terminology. This scale has in later descriptions been simplified somewhat (e.g. Ariel 2001), but the principles of its construction remain unchanged within the accessibility theory framework.

Certain expressions (or "markers" in Ariel's parlance) are specialized to indicate high degrees of accessibility (i.e. those at the bottom of the scale); this includes various pronouns as well as zero anaphors. Other expressions are largely restricted to low accessibility contexts (i.e. at the top of of the scale): These are mostly lexical NPs, including various constellations of proper names (Ariel 2004: 92). In addition to the basic zero–pronoun–lexical NP distinction, the accessibility scale further distinguishes expressions by their weight, use of demonstrative determiners, type of demonstrative (i.e. proximal vs. distal, etc.), as well as phonological stress and independence.

⬆	lower accessibility
(a)	full name with modifier
(b)	full name
(c)	long definite lexical NP
(d)	short definite lexical NP
(e)	surname
(f)	given name
(g)	distal demonstrative with modifier
(h)	proximal demonstrative with modifier
(i)	lexical NP with distal demonstrative
(j)	lexical NP with proximal demonstrative
(k)	stressed pronoun with gesture
(l)	stressed pronoun
(m)	unstressed pronoun
(n)	cliticized pronoun
(o)	extremely high accessibility markers (gaps, traces, reflexives, agreement)
⬇	higher accessibility

Figure 2.1 | The accessibility marking scale for English in its original form, reproduced with adapted terminology from Ariel (1990: 73, ex. 1).

Notably, Ariel chooses to include "gaps, *wh*-traces, reflexives, and agreement" at the extreme high end of the scale (1990: 73), that is, expressions that are structurally conditioned, rather than selected by pragmatic choices made by speaker. Ariel grounds this decision in the claim that since "Accessibility is a cognitively based concept guiding anaphoric choices, we should expect it to be functional at the sentence level too, perhaps even partially reflected in the grammar itself" (1990: 97), invoking binding conditions (Chomsky 1982) as a comparable system. According to accessibility theory, then, both anaphoric and intra-sentential references are subject to one and the same accessibility constraints; we will discuss why this is problematic and other issues with accessibility theory further below.

Ariel cites three coding principles for the accessibility marking scale: infomativity, rigidity (= specificity), and phonological attenuation. Informativity, as mentioned above, is a function of the semantic content of an expression

and a measure of how distinctive it is in a given context. Informativity is inversely proportional to accessibility, with low informativity indicating high accessibility; lexical NPs tend to be more informative than pronouns, which in turn are more informative than zero anaphors, as the latter have no explicit semantic content.

Rigidity (or specificity of reference) indicates how "uniquely referring" (Ariel 1996: 21) an expression is. Rigid expressions are preferred when accessibility is low, as they allow for more precise identification of the intended referent vis-à-vis alternatives. In general, lexical NPs refer more specifically than pronouns and zero, with proper names being the most rigid, although the precise gradation of different constellations of proper names (for people names at least) tends to be specific to culture: In the larger Western cultural sphere, surnames tend to be more uniquely referring than given names because there exists "a greater variety" of them (Ariel 2006: 16) compared to the relative paucity of surnames in, for example, certain East Asian cultures.

The third and last coding principle involves the phonological attenuation of referring expressions: All else being equal, forms with reduced phonological size are associated with higher degrees of accessibility (and vice versa), yielding a cline of stressed, unstressed, and cliticized pronouns; lexical NPs are the least attenuated, zero anaphors the most. These three principles overlap substantially (Ariel 2006: 16), as informative expressions tend to also be fairly rigid and unattenuated, and highly attenuated expressions conversely uninformative and not very specifically referring.

Although the accessibility marking scale does not list every possible type of referring expression (and combinations thereof), Ariel (1990: 76–92, 2006: 16) claims that its hierarchization is "virtually universal", but strictly speaking this applies only to the mechanisms from which it is generated, not the scale in Figure 2.1 itself: Ariel (1990: 92) asserts that "all languages employ the same three principles (Informativity, Rigidity, and Attenuation) in translating the concept of Accessibility into an actual linguistic marking system." However, there is no single accessibility scale that is applicable across languages; the hierarchizations are by necessity language-specific, as every language associates particular levels of accessibility with different expressions (Ariel 1990: 76–79). As such, the scale reproduced in Figure 2.1 only applies to English; similar scales can be constructed for any language, Ariel argues, but the specific arrangement of levels is expected to differ from language to language. Yet all languages, irrespective of their respective sets of

available referring expressions, are claimed to define the distribution of said expressions in terms of accessibility, so that "discourse in all languages is governed by the very same Accessibility principles" (Ariel 1990: 92–93). For this reason, it is possible to predict the degree of accessibility associated with an anaphoric expression "due to a principled association of specific forms with specific levels of Accessibility, an association which holds across numerous languages," and "although the precise Accessibility rate attached to a specific referring expression may vary from one language to another, no language having counterparts […] of the expressions listed in [the scale for English] can arrange them in a different order of Accessibility" (Ariel 1990: 75–76). Accessibility scales can predict accessibility only in relative terms, however, as languages tend to place extraneous constraints on certain expressions, by limiting their occurrence either grammatically or pragmatically. A language may hence employ certain levels on its scales under circumstances not predicated by considerations of accessibility, or may lack expressions that English has, or use others that English lacks in turn, but, so Ariel (1990: 76) asserts, "it is predicted not to violate the principle that degree of Accessibility dictates formal choices."

According to accessibility theory, accessibility is thus the primary determinant of referential choices in essentially all instances. This applies both to discourse-new referents and their introductions – which tend to have very low accessibility – as well as subsequent anaphoric references. Accessibility is determined by the interplay of multiple soft constraints on referential choice, which which fall into two broad categories – "salience" and "unity" – acting in unison. Certain referents, such as humans, protagonists, and the speech participants themselves, are inherently salient; other referents, most notably discourse topics, have contextual salience. Highly salient referents tend to also be highly accessible, but competition between multiple potential referents can reduce salience, which in turn adversely affects their accessibility (Ariel 1990: 28; see Section 2.4.2.4 below).

Unity is an umbrella term for various factors describing the nature of the relation between the anaphor and its antecedent (Ariel 1996: 22). Recency, that is the figurative textual "distance" between anaphor and antecedent, is claimed to hold by far the greatest influence on accessibility (see Section 2.4.2.1), with accessibility decreasing as distance increases, and the highest levels of accessibility found at very short distances. References in cohesively linked units (i.e. in the same clause/sentence/paragraph/etc.) are also overall more accessible

(cf. also Clancy 1980; Givón 1983a), as highly cohesive units are processed as more relevant to one another, so that potential antecedents in linked units are given preference. However, as Ariel cautions, no single factor can explain the observed variation in the use of referring expressions by itself; rather, it is their joint contribution to accessibility that drives referential choices, so that "accessibility marker selection is determined by weighing together a whole complex of accessibility factors, which together determine what the degree of accessibility of a given discourse entity is at the current stage of the discourse" (Ariel 2006: 17).

Therein lies one of the larger conceptual issues with accessibility theory: The accessibility scale is claimed to be continuous, which means that strictly speaking, all of its levels are associated with a single variable (i.e. accessibility) that functions as the sole determinant of selection. Yet accessibility itself is determined by complex interplay of various factors dependent on the discourse context, and so cannot be reduced to a single constituent property, as it is what Ariel (2006: 17) calls a "total concept". Where accessibility theory pulls short, then, is in giving a full account of how the various contributing factors interact and converge on a particular form choice. As such, the claims made by the accessibility scale are not readily falsifiable – if one factor influencing accessibility cannot explain a specific choice of form, then it has to be another, but which exactly? How do factors interact with each other? Which factor weighting is triggered under which circumstances? As the exact pathways that lead to a form predictions are largely inscrutable, this essentially renders the notion of accessibility akin to a blackbox in practice, which can lead to some explanations of divergent patterns (in Ariel 1990 in particular) to be resolved by what at times seems like a "yes, but also..." approach. See Siewierska (2004) and Huang (1994) for more in-depth criticism of accessibility theory along these lines. Admittedly, this weakness largely stems from it being a rather ambitious framework that is unquestionably a child of its time, from before large digitalized corpora allowed for robust grounding of claims in statistical observations; much of accessibility theory is based on theoretical considerations and qualitative assessments, with only minimal quantitative testing – in the original 1990 study on two small corpora of English and Hebrew, mostly written, some conversational – though later studies have tested its predictions on larger samples from other languages (e.g. Gipper 2016 for Yurakaré, Portele & Bader 2016 for German, among others).

The theoretical underpinning of accessibility theory are especially notice-

able with the finer distinctions included on the scale, such as the various constellations of proper names, as well as phrase complexity, and the distinctions of proximal and distal demonstratives in terms of accessibility marking. Many of these are not based on empirical observations, but on theoretical reasoning only, and have been called into question by later studies (e.g. Botley & McEnery 2001: 221–222). One preliminary study (Schiborr 2017: 53–58), testing the selection of heavier (vs. lighter) lexical NPs as well as of proximal (vs. distal) demonstratives in spoken English, each supposedly indicating higher and lower degrees of accessibility, finds that neither is tied to recency, the strongest determinant of accessibility according to Ariel (1990).[4] The lack of an observable association between phrasal complexity and highly significant discourse properties such as recency in particular casts doubt on the fundamental ordering of accessibility scales in terms of informativity and specificness of reference.

The purported fundamentality of accessibility also motivates the inclusion of expression types not selected based on pragmatic considerations but syntactic constraints, such as gaps and reflexives on the accessibility marking scale (Ariel 1990: 97–105). While such expressions are essentially outside the purview of referential choice in any case, their inclusion on the scale does merit some comment. Ariel asserts that considerations of accessibility motivative form selections for intra-sentential references in much the same way as cross-sentential anaphora; since accessibility is deemed to be a cognitive property, it holds in all instances of referential choice, and in essence underpins all constraints on form selection. How one and the same notion of accessibility can explain choices (or lack thereof) under such disparate conditions is not entirely cogent, however. Accessibility, as noted above, is supposedly by and large determined by the salience of the referent and the unity of the local discourse, neither of which should affect syntactically conditioned constraints on reference form.

All this being said, accessibility theory, for all its shortcomings, nevertheless provides a wealth of useful insights into the relationship between referring expressions, discourse referents, and their underlying mental representations.

4 We will repeat some of these observations with a more nuanced approach in Chapter 8, and again find that discourse properties have no bearing on the selection of these expressions, contrary to claims of accessibility theory.

In particular, the central notion of multiple soft constraints on reference form is crucial for the present study. Though this study does not work within an accessibility theory framework, it tests a number of claims made by it, which we will evaluate in Chapter 9.

2.2.5 | Related and other approaches

2.2.5.1 | Constraints on discourse reference

Du Bois (1987b, 2003b, 2017) identifies a number of pragmatic and grammatical constraints on possible realizations of discourse references. Specifically, Du Bois notes that in natural discourse, certain types of expressions are more or less likely to be found in certain argument positions, so that referential choices pattern consistently with regards to different argument structures. These and related observations lead to the formulation of four constraints on argument realization, two pragmatic (5) and two grammatical (6) (Du Bois 1987b: 829):

(5) *a.* avoid multiple lexical core arguments in a clause ("one-lexical-argument constraint")

b. avoid multiple new referents in a clause ("one-new-argument constraint")

(6) *a.* avoid lexically realized subjects of transitive clauses ("non-lexical-A-constraint")

b. avoid new information in subjects of transitive clauses ("given-A-constraint")

The constraints in (5) in particular are manifestations of Chafe's (1994) "one new idea at a time constraint". see also Lambrecht's (1994) "principle of the separation of role and reference".

Du Bois (1987b) derives these constraints from observations of a small data set of spoken Sakapultek (Mayan, Guatemala), a language with ergative morphology, though their individual relevance has been confirmed by studies on larger samples of diverse languages (e.g. Kumagai 2006 for English). For Sakapultek, Du Bois finds that the subjects of transitive clauses (A) are sub-

stantially less likely to be expressed lexically than objects (P) and subjects of intransitive clauses (S), which are roughly equally likely to be lexical (a finding that latter studies have found to be anomalous, see below). This association holds especially for information not previously mentioned in the current discourse (i.e. "new" in the sense of Chafe 1976) in particular: Like Chafe above, Du Bois considers the instantiation of new referents into discourse to be a particularly "cognitively demanding task" (Du Bois 2003a: 38), both for production as well as comprehension, as they tend to be longer and more complex, and require the creation of a new 'mental file' in the working memoery. Speakers thus tend to spread out the introduction of new information across multiple clauses to alleviate the figurative "information pressure" exerted by the introduction of new referents under conditions of elevated cognitive load, in particular where many different participants are involved concurrently in a discourse (Du Bois 2003a: 76). Speakers primarily utilize the S and P roles for this purpose, but avoid the A role. Intransitive clauses in particular are hence added chiefly for the sake of information flow, not just for their conceptual content or semantic one-placeness (Du Bois 1987b: 831). This results in what Durie (2003) calls "pragmatic linking": Certain positions (i.e. S and P) provide "a predictable locus for unpredictable work" (Du Bois 2003a: 47). All this leads to the formulation of a "preferred argument structure", according to which "the information distribution among argument positions in clauses of spoken discourse is not random, but grammatically skewed toward an ergative pattern" (i.e. A vs. S and P, Du Bois 1987b: 805); ultimately, Du Bois (1987b) is interested in explaining the diachronic emergence of ergative morphosyntax from the supposed ergative alignment of lexical NPs in discourse noted above (see also Du Bois 1987a on the notion of "absolutive zero" in agreement paradigms).

These claims have been conclusively challenged in recent years: The non-lexicality of A is better explained in terms of a general association between agentivity and humanness (and hence with discourse salience) than through information management (Everett 2009; Haig & Schnell 2016; cf. also Payne 1987; Kärkkäinen 1996); likewise, newness does not play a major role in determining the form of core arguments overall, and the discourse profiles of A and S in fact have more in common than those of S and P (i.e. "discourse accusativity" in Kumagai 2006; see also Haig & Schnell 2016); and lastly, the supposed costliness of introductions and consequent use of intransitive clauses as an "escape valve" fail to materialize in larger samples of discourse

in a meaningful way, with speakers instead introducing new referents primarily in relation to already established ones, irrespective of position (Schnell et al. 2021b; Schnell & Haig & Schiborr & Vollmer 2020).

The specific claims of preferred argument structure are only indirectly relevant to considerations of referential choice, as the motivators for actual form selection are not within their purview. Nevertheless, there are a number of take-aways from the constraints in (5) and (6), when taken as observational tendencies of natural discourse. The avoid-lexical-A constraint suggests that, all else being equal, anaphoric subjects of transitive clauses should be noticeably less likely to be realized lexically than subjects of intransitive clauses (and other positions; see Section 2.4.3.1). This augments claims regarding the general association between subjects (specifically A) and reduced forms (cf. Chafe 1976; Ariel 1990; and others), which ultimately stems from their above-mentioned association with human and topical referents (cf. Haig & Schnell 2016). Furthermore, as per the one-lexical-argument constraint, we should find that if a given clause already contains a lexical argument, others are less likely to be lexical as well; lexical object should hence be more common alongside non-lexical subjects, and vice versa (see Section 2.4.3.2).

2.2.5.2 | Centering theory

Centering theory (Grosz et al. 1983, 1995; see also Walker et al. 1998) aims to account for the more localized properties of speakers' attention in discourse production by formalizing the relationship between form of expression and focus of attention. This formalization revolves around two types of so-called "centres" of local discourse, a backward-looking centre and one or more forward-looking centres, which are present in every utterance in a given stretch of discourse (Grosz & Sidner 1986). The former represents the current focus of attention in a given clause, and is determined by the relative salience of the discourse entities mentioned in the previous clause (i.e. the forward-looking centres). In centering theory, salience is primarily determined by syntactic prominence, with subjects outranking objects, which in turn outrank oblique arguments and other constituents (Gordon et al. 1993; see also Section 2.4.2.2 below), as well as position in linear syntactic order, with sentence-initial mentions being particularly salient. Having been the backward-looking centre of a preceding clause also provides a boost to salience (cf. topic continuity; Section 2.2.3). In essence, the highest ranking forward-looking centre

is expected to be picked up again as the backward-looking centre of the subsequent clause. In this way, the two types of centres are claimed to together provide the mechanism for establishing local coherence in discourse.

Centering theory rests on the assumption that speakers aim to produce discourse that is maximally coherent, chiefly by minimizing the complexity of the inferences addresses have to make about the relations between consecutive clauses (Grosz et al. 1995). This is achieved primarily by avoiding shifts in the identity of the backward-looking centre from clause to clause, and in the case where a shift does occur, by employing certain types of referring expression to signal that the expected alignment of forward and backward-looking centres does not continue. Generally, the most salient referents are referred to via pronouns (or zero anaphors), so that in turn, any pronoun in a given clause is a likely candidate for its backward-looking centre, though additional referents may be pronominally referred to concurrently.[5] Conversely, if none of the previously mentioned referents are sufficiently salient, the backward-looking centre may be expressed lexically.

Experimental research has shown that pronominal references to more salient entities and with closer forward-looking centres are processed more quickly than others, and that sentences are read faster and deemed more coherent when the backward-looking centres are pronouns rather than lexical NPs (Gordon et al. 1993; see also Brennan 1995). Centering theory has also been used as a basis for computational models, both for anaphora resolution and referring expression generation (e.g. Brennan et al. 1987 and Poesio et al. 2004 among others; cf. also Webber 1988 and Section 2.3 below).

2.2.5.3 | Rhetorical structure theory

Rhetorical structure theory (Matthiessen & Thompson 1988; Mann et al. 1992; Taboada & Mann 2006) aims to describe the organization of discourse, originally with the aim of aiding the automatic generation of texts. It was developed on the basis of exhaustive analysis of written data, but has since also been applied to spoken discourse (e.g. Taboada 2008 for spoken Spanish). Rhetorical structure theory has found common use especially in computational linguistics, where it is used to parse and plan the structure of coherent text. The theory

5 Considerations of centering also apply to intrasentential references, see Kameyama (1998).

was deliberately developed without grounding in existing theoretical frameworks and few assumptions of how (written) texts are structured, so as to be flexibly adaptable to different applications and linguistic situations. Taboada & Mann (2006: 425) note that it is intended to complement other systems of text description, such as semantics and pragmatics, and that "[t]he most familiar kinds of linguistic description, about words, phrases, grammatical structure, semantics and pragmatics, all make contributions that are qualitatively distinct" from those of rhetorical structure theory.

The theory subdivides discourse into elementary discourse units (EDUs), and attempts to explain text organization through the various relations between these units. This yields a hierarchical, connected structure of texts, where every element of a text has a function relative to other parts. These "coherence relations" are established via specialized annotations of a text, from which relational diagrams that map discourse structure onto binary trees can then be constructed. In this way, rhetorical structure offers a different view of text organization than most theories of discourse, in that it "points to a tight relation between relations and coherence in text" (Taboada & Mann 2006: 428). According to rhetorical structure theory, coherence is created through two related mechanisms (cf. Halliday & Hasan 1976; Kehler 2002; Poesio et al. 2004): One engendered by entities forming chains in discourse ("entity-based coherence"), and one that stems from implicit or explicit relations between parts of a text ("relational coherence").

The insights gained from exploring rhetorical relations have also been applied to studies of anaphora and discourse structure (e.g. Fox 1987b; Poesio & di Eugenio 2001), as well as semantic roles (Stevenson et al. 2000), certain clause-internal structures such as gapping and VP ellipsis (Kehler 2002), text cohesion in natural language processing (McNamara & Kintsch 1996), and, most importantly for the present study, the prediction of referential choices (Kibrik & Krasavina 2005; Tetreault 2005).

One such synergetic development is veins theory (Cristea et al. 1998; 2000; Ide & Cristea 2000), which examines the effect of discourse structure on anaphora on the basis of centering theory (Grosz et al. 1995; Section 2.2.5.2). Veins theory uses rhetorical structures to classify discourse structure relations, in the process identifying so-called 'veins' over rhetorical structure trees. These veins define the "domain of referential accessibility" for each referring expression, which places operational constraints on its selection in a given discourse context. Rhetorical structure theory also has some

⬆	higher status
(a)	in focus
(b)	activated
(c)	familiar
(d)	uniquely identifiable
(e)	referential
(f)	type identifiable
⬇	lower status

Figure 2.2 | The givenness hierarchy, reproduced in its original form from Gundel et al. (1993a: 275, ex. 1). Higher states imply all lower states.

conceptual overlap with segmented discourse representation theory (see seminal work by Kamp 1981 and Kamp & Reyle 1993, and later developments in e.g. Lascarides & Asher 1993; Asher & Lascarides 2003; etc.), which has been used to explain a number of related phenomena, including anaphora, bridging inferences, implicatures, presuppositions, and others.

2.2.5.4 | Givenness theory

Gundel et al. (1993b, 1993a) propose a hierarchy of cognitive states tied to specific types of referential forms. This "givenness hierarchy", here reproduced from Gundel et al. (1993a: 275, ex. 1) in Figure 2.2, describes six discrete levels of givenness of mental representations in the addressee's memory. Givenness in this sense is distinct from the notion of the same name in Chafe (1976) and Prince (1981b) (where it constrasts with newness, see Section 2.2.2), as it is not an inherent property of discourse entities, but rather a psychological measure of which entities the addressee is currently focusing on, as surmised by the speaker (cf. audience design), irrespective of whether said entity has already appeared in the preceding discourse or not. In other words, specific types of expression serve as "processing signals" (Gundel et al. 1993a: 276; cf. Ariel 1988; Garrod & Sanford 1982), indicating to the addresse where to "look" for the target entity in their memory.

Gundel et al. account for both entities in the linguistic context as well as those without (i.e. in the physical surroundings), but I will focus here specific-

ally on the former. An entity is deemed "in focus" or "activated", respectively, if it was introduced in a syntactically prominent position or merely mentioned in the previous clause, and assumed "familiar" if it has either been mentioned at any point in the preceding discourse or else can be derived from shared knowledge. "Uniquely identifiable" entities are inferable from another recently activated referent (via bridging, cf. Huang 2000a), or otherwise construable "solely on the basis of information conceptually encoded in the phrase" (Gundel 2010: 154, ex. 5). Conversely, entities are merely "referential" if reference to them is primed by the discourse context, that is, if it is clear that the speaker intends to refer to particular object or class of object. And lastly, an entity is "type identifiable" if the addresse is assumed to be able access a representation of the type of object described.

Entities in focus are usually signalled by personal pronouns (or zero anaphora), activated entities mainly by demonstrative pronouns (e.g. *this* or *that*, but also *this book*), familar entities by distal demonstrative NPs (*that book*), uniquely identifiable entities by regular definite NPs (*the book*), referential entities by what Gundel et al. (1993b; cf. Prince 1981a) call "indefinite" demonstrative NPs (*this book*, as in *there's this book I've been reading...*), and type identifiable entities solely by indefinite NPs (*a book*).

The six states on the givenness hierarchy are discrete – unlike, for example, the continuous accessibility hierarchy in Ariel (1990) – so that each state leads to a specific set of outcomes. The states are furthermore implicationally related (Gundel et al. 1993a: 275–276): Each state also entails all lower states, and attainment of higher states does not render lower states inoperable. As such, an entity in focus is by definition also activated, and an entity that is activated is necessarily also familiar, and so on (Gundel 2010: 155). For example, a definite lexical NP with *the* in English may be uniquely identifiable, or it may have one of the higher states (i.e. familiar, activated, or in focus) instead. However, pragmatic constraints encourage speakers to be maximally informative, and avoid use of expressions that are more explicit than is necessary (cf. Grice 1975). While these associations given above make it quite clear that givenness theory was (at least in its initial form) based chiefly on considerations of English discourse, the theory does attempt to account for cross-linguistic differences in use of referring expressions, and lays claim to universality (Gundel et al. 1993a: 284–285; cf. Gundel et al. 2010).

Notably, Gundel (2010) stresses the fundamental conceptual differences between givenness theory and most theories of discourse structure, such as

accessibility theory (Section 2.2.4). Unlike accessibility theory's scale of referring expressions, the givenness hierarchy does not predict that referents of expressions that encode higher cognitive states are necessarily also more easily retrievable, as there are other factors that come to bear, such as the amount of conceptual content and competition between candidate referents (2010: 161). Since there is no one-to-one mapping between accessibility and cognitive status, the theory does not predict that being in focus or activated is a sufficient condition for being accessible, and vice versa. While specific cognitive states may constrain speaker's choices of expression and in this way indirectly provide information about referent accessibility, the givenness hierarchy is hence itself not an accessibility scale in the sense of Ariel (1990) or Givón (1983a). As such, since it is impossible in practice to assess (from an annotator's point of view) which cognitive state speakers assume their addressees to hold for a given discourse entity, givenness theory does not hold the same predictive utility as other models of referential choice, as it is in essence limited to explaining said choices ex post facto.

2.3 | Computational approaches

Where the approaches to discourse anaphora and referential choice discussed in the previous section have been largely theoretical and based on qualitative judgements, those outlined in this section instead take a fundamentally quantitative stance, using statistical and other computational models to describe and predict speakers' choices and the structure of texts. Webber (1988: 1) frames computational approaches to discourse as having two closely related goals: First, the identification of "those aspects of discourse understanding that require process-based accounts", and second, the characterization of "the processes and data structures they involve." For the production of references in discourse, the aforementioned theories of reference provide numerous insights into the potential mechanisms involved, but typically describe them in terms of categorical and deterministic entailments of discourse context to cognitive status, and from cognitive status to form choice.

A less abstracted approach to referential choice might consider all form choices in all positions and contexts – that is, those that allow alternation – to be fully continuous, and entirely a matter of statistical tendency (cf. Kibrik et al. 2016). In principle, this means that all valid options are always available

to speakers in all situations, with some being more or less likely, even if in practice certain forms may (almost) never appear in certain extreme contexts. This is represented, for example, in the implicational nature of the givenness hierarchy in Gundel et al. (1993b, 1993a; see Section 2.2.5.4 above).

This line of thinking is captured well by computational approaches to referential choice, which tend to not work in terms of categorical classification, but rather through working out the relative likelihoods of particular form choices as determined by a complex interplay of multiple quantitative factors (cf. the fundamental idea of how accessibility is determined in accessibility theory). In this way, computational models offer an alternative to hypothesis testing (as done, e.g. with the small scale statistical analyses used to substantiate claims in Ariel 1990), and allow for so-called "bottom-up" accounts of linguistic phenomena that are informed by, but do not start from, any theoretical framework.

Within linguistics and typology, this approach has become increasingly widespread in recent years – a trend this study aims to continue – but a similar way of thinking has been the status quo in the computational linguistics of reference and anaphora for decades. While this section is primarily focused on computational work within linguistics proper,[6] such as that by Andrej Kibrik and colleagues, and others (Section 2.3.1), we will also briefly outline the state of research in two pertinent areas of computational linguistics and natural language processing in Section 2.3.2 further below.

2.3.1 | Quantitative models of referential choice

Kibrik (1996, 1999, 2000; Grüning & Kibrik 2005) develops a "cognitive multifactorial approach" to predicting referential choices, based on the understanding that speakers' choices are governed by the in-situ activation of referents in working memory, or rather by speakers' assumptions about the state of the addressee's working memory (cf. Chafe 1994; Givón 1995). Kibrik argues that a psychologically adequate model of referential choice has to "look for explanations of referential choices in the cognitive structures of the speaker at the moment of speech" (2000: 72; i.e. cognitive), and has to take into account the potential for multiple aspects of said structure to weigh on referential choices concurrently (i.e. multifactorial, Kibrik 2011: 393–394). The

6 As in, "linguistics with computers" rather than "computational linguistics".

latter point in particular distinguishes this approach from theories that explain choices via association with a single notion, such as accessibility (Ariel 1990) or topicality (Givón 1983a).

The calculative approach proposed by Kibrik and colleagues quantifies referent activation as a scalar variable: Various "activation factors" – such as the distance between anaphor and antecedent (both in terms of linear textual distance and rhetorical distance; Fox 1987b, see Section 2.2.5.3), the syntactic position of the latter, as well as certain properties of the referent itself, including animacy and protagonisthood – are assigned specific numerical weights, and the sum of the weights for each factor yields an "activation score" for a referent at a given point in the discourse. If this score exceeds a certain threshold, then the speaker is likely to use a reduced expression for this referent, else a lexical one (Kibrik 2011: 403–411). The model distinguishes categorical from alternating referential choices through the use of soft and hard thresholds, with the latter necessitating a certain form, and the former merely enabling it as a possibility. Crucially, all factors contribute or detract to the final score in all cases rather than in an ad-hoc manner, and possible interactions between factors are accounted for. According to Kibrik, this approach aims to "imitate the cognitive interplay of activation factors during discourse production" (2000: 72).

The major shortcoming of the early calculative approach is that, in Kibrik's own admission, the specific weightings for each factor are obtained essentially through "trial and error" (2000: 75), rather than through automatic, data-driven means. Later developments in the cognitive multifactorial mindset (Khudiakova et al. 2011; Loukachevitch et al. 2011; Kibrik et al. 2013, 2016) address this issue by employing methods from the statistics and machine learning toolkit, adopted from work on referring expression generation (see Section 2.3.2.1). Kibrik et al. (2013) use models trained on corpus data annotated for almost two dozen features to predict choices between reduced and lexical expressions on the one hand, and pronouns, full NPs, and proper names on the other. They compare the predictive power of various modelling strategies, finding that their approach performs at accuracies well above random chance, with the best results obtained by classification tree ensembles (cf. Elith et al. 2008; Strobl et al. 2009; Ridgeway 2020).[7] The relationship

7 It should be noted, however, that the focus on predictive accuracy as an evaluation metric

between the numerous predictors in these kinds of models is highly complex: Most exert only small influence on the results, but all contribute in some way, so that removal of any one factor leads to an appreciable reduction in predictive accuracy (Loukachevitch et al. 2011).

While Kibrik et al. (2013) conclude that the use of machine learning algorithms is a viable way of modelling referential choices, they also argue that perfect prediction is likely impossible, irrespective of how comprehensive one's approach may be, because speakers' choices are not fully categorical. Kibrik et al. (2016) further develop this observation, noting that there are conditions under which speakers select from multiple equally appropriate expressions, and may hence choose differently on different occasions. They test machine learning models similar to those in the earlier study against the judgements of human raters in an experimental task, and find that the divergences between the predicted and elicited values are most noticeable in those cases where choices are not categorical. In particular, pronominal forms are a viable option in most contexts, either marginally or in fully-fledged free variation with other forms (Kibrik et al. 2016: 10, ex. 3). The non-deterministic nature of referential choice thus means there is a degree of randomness inherent in discourse data, which ultimately places limits on the predictiveness of statistical models of referential choices.

Given their focus on proving the viability of predictive models for studies on referential choice, Kibrik and colleagues devote fairly little space to explaining the specific characteristics of their models, and only in passing expound on the relative influence of the tested factors. Since the predictive approach already operates on the assumption that statistical patterns are indicative of linguistic and cognitive associations (or can even serve as stand-ins for them in principle, cf. Kibrik et al. 2016: 207), it is well within reason to seek more nuanced explanations from them, as is done in the present study. Lastly,

is not ideal in practice, since the selection of the various types of referring expressions in the corpus data used by Kibrik et al. is not balanced in terms of relative proportions, with lexical expressions outnumbering non-lexical by a factor of three (Kibrik et al. 2013: 3, tab. 1). The more common outcome hence dominates the results while also being the easiest (i.e. least interesting) to predict. We will address this issue for the present study in Section 7.2 below, where we face the inverse imbalance (i.e. more non-lexical than lexical mentions) due to differences in the types of texts sampled (spoken here vs. written in Kibrik et al. 2013).

it should be mentioned that both the earlier manual calculative approach and the more recent classification tree-based approaches discussed in this section have exclusively drawn from written data, chiefly from English and Russian, as well as Japanese (Efimova 2006). Kibrik et al. (2016: 16) argue that methods of the latter are in theory applicable to "a wide range of discourse types, including various genres, spoken discourse, conversation, and multimodal interaction," but I am not aware of the existence of any such study on spoken data specifically, much less any taking a typological perspective (but see, e.g. Torres Cacoullos & Travis 2019 for a related approach).

2.3.2 | Anaphora in computational linguistics and NLP

The myriad of approaches to discourse anaphora and referential choice in computational linguistics and natural language processing can be broadly classified into two groups by their aims and perspectives on the data: The first seeks to automatically generate appropriate referring expressions in a given context (Section 2.3.2.1), and is as such in principle similar to the predictive-explanatory angle taken in linguistics proper. The second instead takes a post-hoc perspective on reference, seeking to identify co-reference relations between the expressions in a text (Section 2.3.2.2). While earlier approaches in both fields are chiefly based on linguistic knowledge and its implementation in terms of algorithms, in recent years methods have drifted further and further towards knowledge-independent methods and the use machine learning, neural networks, and "data science", so that the inclusion of linguistic knowledge in recent models is now a rarity instead of the norm (cf. Mitkov 2000; 2002; Mitkov & Barbu 2001). State-of-the-art methods are inarguably more powerful and accurate, but are essentially black boxes. For this reason, this section is mostly focused on the earlier state of research in computational linguistics from the 1970s to the early 2010s, whose considerations and findings can still (if at times only indirectly) provide some useful methodological insights.

2.3.2.1 | Referring expression generation

Algorithms for refererring expression generation seek to simulate speaker's referential choices in computational terms as part of more general agenda of natural language generation (cf. Belz et al. 2010; Krahmer & van Deemter 2012; Gatt et al. 2014; see also contributions in Branco et al. 2005). Their chief aim is the production of naturalistic, human-like references in a given discourse context – in most cases in written English. In one of the earliest approaches, Appelt & Kronfeld (1987) propose a formal computational model of reference planning centered around speakers' goals in terms of communicative intent (cf. also Appelt 1985; Dale & Reiter 1995). They base this model on the process of referent individuation via "identification constraints", which derive from the relationships between an intensional representation and an entity of object that can be referred to. Given a high-level description of speakers' goals, their model generates referring expressions – pronouns, lexical NPs, and proper names – appropriate to the communicative situation.

Most models of referring expression generation are concerned with the form of definite NPs for introducing external entities (i.e. in the physical world) into discourse. Dale & Reiter (1995), for instance, evaluate the optimal information content of expressions to enable correct identification of a not previously mentioned entity (cf. Krahmer & van Deemter 2012). Krahmer & Theune (2002) extend this approach to anaphoric mentions (see also Tutin & Viegas 2000) by also assessing the salience (in terms of centering theory, Section 2.2.5.2) of the intended referent, rather than selecting forms that only minimally distinguish it from competitors. Their algorithm assigns each discourse entity a score, which is increased every time it is mentioned, but decreased with every utterance it is not. Mentions in syntactically prominent positions (e.g. as subject) increase this score more than less prominent positions. The algorithm selects the most appropriate form based on the score of the target referent, which can be an underspecified form if salience is high enough. Other approaches also take into account factors such as animacy and ontological class (Strube & Wolters 2000) and the global properties of the discourse (McCoy & Strube 1999; Callaway & Lester 2002).

More recently, many advances have come out of the various GREC challenges ('Generating Referring Expressions in Context'; e.g. Belz et al. 2008, 2009, 2010; Krahmer et al. 2008; Greenbacker & McCoy 2009; and many more), which aim to predict the form of referring expressions in written data.

These tend to employ a variety of machine learning techniques with a focus on different aspects of linguistic and discourse structure; Greenbacker & McCoy (2009), for instance, use measures of competition, formal and structural parallelism, and recency.

There have also been efforts to "bridge the divide" between computational and empirical approaches to reference; see, for instance, the contributions in van Deemter et al. (2009, e.g. Zulaica Hernández 2009 on anaphoric distance and the use of demonstratives) and Botley & McEnery (2000a), and the annotation schema for anaphora described in Poesio & Artstein (2008). In a similar vein, van Deemter et al. (2012) aim to bring psycholinguistic considerations into models of reference production. Lastly, Orita et al. (2015) find that their language production model performs better when taking into account discourse information, noting that "speakers' behavior can be modeled in a principled way by considering the probabilities of referents in the discourse and the information conveyed in each word," but also cautioning that while "the relationship between discourse salience and speakers' choices of referring expressions is well known, there is not yet a formal account of why this relationship exists" (2015: 1639).

2.3.2.2 | Anaphora resolution

Anaphora resolution largely employs the same data and many of the same methods as referring expression generation, but has in essence the inverse goal: Here, the aim is to automatically identify the antecedents of (chiefly pronominal) expressions, that is, to establish co-reference relations between mentions, both anaphoric and to text-external entities (cf. Mitkov 2000, 2002; Mitkov & Barbu 2001; Tetreault & Allen 2003; Tetreault 2005; etc.).

Early approaches to anaphora resolution rely on statistical models that assign probabilities to candidate antecedents on the basis of pre-defined sets of criteria (e.g. Lappin & Leass 1994; Ge et al. 1998; and others). The algorithm described in Lappin & Leass (1994) operates via a discourse model containing information on discourse as a whole and the referents contained therein. A text is scanned linearly from beginning to end, and the model is updated whenever a new referent is introduced. The antecedent for a given pronominal anaphor is chosen from the list of all potential candidate antecedents by weighing its recency effects against syntactic preferences. Lappin & Leass report 86% accuracy for their algorithm when trained on a corpus of English

computer manuals (1994: 554). The highest weight is given to antecedent distance, measured in orthographic sentences; other contributing factors, in descending order of influence, are whether the candidate antecedent is the subject of its clause, the predicate of a presentational construction, an object, an oblique argument, or some other constituent. As the text in question is scanned and the associated discourse model grows, the likelihood of all referents increases naturally; this is accounted for by halving all likelihoods every time a sentence is processed. In effect, this algorithm assumes that the probability of a candidate antecedent to be the correct choice to drop quadratrically with each sentence separating it from the anaphor in question, so that it assigns the greatest likelihood to candidate antecedents that are immediately preceding the anaphor.

More recent approaches to anaphora resolution are more powerful, but also considerably more complex than earlier models, using methods from the machine-learning and neural network toolkits, and as such tend to operate without involving linguistic knowledge. As they require massive amounts of training data, they are still largely limited to written language, mostly from English.

One such model constructs modularized clustering models (Haghighi & Klein 2007), a later development of which (Haghighi & Klein 2010: 386) in particular employs a matrix of "tree distance [i.e. hierarchical syntactic distance, as gleaned from a parse tree], sentence distance [i.e. linear, textual distance], and the syntactic positions (subject, object, oblique) of the mention and the antecedent" to establish co-reference relations. The model prefers close antecedents in specific in prominent syntactic positions, and operates fairly reliably for pronominal and common NP mentions, but for proper names Haghighi & Klein find that the choice is "governed more by entity frequency than antecedent distance". Other approaches reason over all possible co-reference relations as sets rather than individual antecedent-anaphor pairs (Culotta et al. 2007) – that is, all mentions of the same referent are considered together, rather than two at a time – with the latter being more intuitive for humans, but turning out to be less predictive in practice. For referential choice, this might suggest that it is not just the immediate context of an anaphor that determines its form, but that the entire profile of a referent, established across multiple mentions, also bears an influence on form selection, and that referents should differ noticeably in this way.

2.4 | Factors influencing referential choice

This section lists a number of the key factors influencing referential choices proposed in the literature, and discusses examples of their treatment. It goes without saying that this list is far from exhaustive, as it is limited by practical constraints to only a small selection of the many features that have been identified as relevant; the focus here is specifically on those factors that are most relevant to the present study.

The list is split into three broad categories: First, those properties that are inherent to referents and their underlying mental representations and as such apply globally in all instances of anaphora, such as animacy and protagonisthood (Section 2.4.1); second, those properties that are dependent on the local structure of the discourse, such as recency as well as syntactic, thematic, and discourse prominence (Section 2.4.2); and third and finally, various properties of referring expressions and the clauses they are embedded in (Section 2.4.3), some of which mostly become relevant from a typological angle.

2.4.1 | Inherent properties of the referent

2.4.1.1 | Animacy and humanness

Animacy is an inherent semantic property of referents, and as such not dependent on the specific contexts in which anaphors are placed (Yamamoto 1999). Since human referents occupy a central position in most types of discourse, animacy considerations in discourse analysis are commonly structured around the dichotomy between human and non-human referents, and hence conceptualized in terms of a binary variable of "humanness". It is generally understood that human referents are more cognitively salient than non-human referents, and hence tend to be more readily retrievable from memory (Ariel 2001: 69–70; Fukumura & van Gompel 2011; cf. also Fraurud 1996 and Dahl & Fraurud 1996). For instance, Givón (1983a) notes an overall higher persistence of human referents in connected discourse compared to non-humans. Furthermore, various evidence from psycholinguistics and psychology – in the form of memory tasks, tests of lexical access and attention, and so on – all point to a cognitively favoured status of human entities (e.g. Fukumura & van Gompel 2011 among many others). These factors contribute to making

human referents have an increased tendency to be referred to with reduced (pronouns, zero) expressions.

Since humans tend to be the primary actors in narratives, there is furthermore a strong correlation between subjecthood and humanness (via semantic role associations), a tendency that is especially pronounced for the subjects of transitive clauses (Dahl 2000; 2008). It is also reflected in Chafe's (1994) light subject constraint as well as in the grammar of inverse alignment systems and certain kinds of split-ergative alignment (Silverstein 1976). The synergy between the resulting three-way association between reduced forms and humanness, humanness and subjects, and subjects and reduced forms, make human subjects the overall least likely to be lexical, non-human non-subjects conversely the most likely.

2.4.1.2 | Finer ontological class distinctions

Of course, the binary humanness distinction is only the most fundamental, but not the only one that matters. Non-human animate referents (i.e. animals), for instance, are likely to be more salient than inanimate objects. Beyond that, inanimate referents can be further differentiated into various ontological categories, such as concrete and abstract entities, individuable and non-individuable entities, and so on, each of which is deemed more or less animate (Dixon 1994; Yamamoto 1999).

Among human (or more generally animate) entities there are likewise gradations in animacy; collectives are less animate than individuals, for instance (in addition to individuation being a factor of salience in general), and semantically plural entities hence tend to receive distinct treatment in many languages (Corbett 2000). One expression of this tendency is the general avoidance of zero expression of plural subjects (cf. Cysouw 2003); similarly, third person plural forms tend to have a special impersonal meaning (e.g. English impersonal *they*, also in Russian and Iranian languages). While I am not aware of any conclusive evidence for the effect of semantic plurality on referential choices in general, it is reasonable to assume that if there is such an effect, it involves semantically plural referents being treated as inherently less salient than singular referents, and hence more likely to receive more explicit forms.

2.4.1.3 | Protagonisthood

Protagonisthood in the sense used in studies on reference (e.g. Kibrik et al. 2016; cf. the "very important participant" status in Dooley & Levinsohn 2001, based on Grimes 1975) refers to a referent's narrative centrality to the discourse. Protagonists tend to be the primary actors and drivers of events, and are almost always human or human-like. They are also 'topical' in the sense that they register key properties of the discourse as a whole, being in essence what the text is "about" (cf. Biber & Conrad 2009). They consequently are highly frequent and highly salient, which makes them more likely to be realized pronominally than other, less narratively central entities (Karmiloff-Smith 1981; Morrow 1985; Linnik & Dobrov 2011). Whether or not an entity is protagonist is not dependent on local discourse structure, but is rather a matter of narrative content, and hence an inherent property of referents; speakers presumably know which entities are most central to their narrative ahead of time, and plan their use accordingly.

Since protagonists tend to be mentioned with comparatively high frequency, one way of assessing their status is by the comparing relative token frequencies of all referents mentioned within a given discourse. Protagonisthood in this sense of total token frequency is correlated with local discourse prominence, which has been argued to bear influence on how referents are established in discourse and subsequently mentioned (Lichtenberk 1996; Himmelmann 1997; see Section 2.4.2.3).[8] Another way of determining protagonisthood, used in Kibrik et al. (2016: 6) following models described in Linnik & Dobrov (2011), is on the basis of a referent's relative anaphoric chain lengths as a proportion of the total number of referring expressions in a text, that is, as a function of the occurrence of said referent in contexts with high topic continuity (as per Givón 1983a; see Section 2.2.3).

8 But see Bischoffberger & Schnell (2014) for a critical re-assessment of these claims and evidence to the contrary.

2.4.2 | Discourse-contextual properties

2.4.2.1 | Recency

Considerations of recency effects are a central component of most approaches to referential choice, both within linguistic research and in natural language processing (e.g. Clark & Sengul 1979; Givón 1983a; Ariel 1990; Kibrik 2000, Kibrik 2011; Kibrik et al. 2016; Arnold 2010; Lappin & Leass 1994; Ge et al. 1998). Recency is commonly quantified as the figurative "distance" between the anaphor in question and its antecedent, though exactly how this distance is measured differs from study to study (see Section 4.6.1.1 for an overview). The greater this distance, it is has been noted, the higher the likelihood of speakers preferring a more informative form of reference, such as a lexical NP. In particular, "the last clause processed grants the entities it mentions a priviledged place in working memory" (Clark & Sengul 1979: 35). As a result, reduced referring expressions (pronouns, zero) are most frequent at short distances from their antecedent (Hobbs 1976; Yule 1981). However, recency only operates on an across-clause basis, as within clauses other constraints such as syntactic and discourse prominence hold sway (cf. Arnold 2010: 190).

Recency effects are generally considered to be a consequence of volume constraints on short-term memory, and hence the mental representations of referents maintained therein. Referents need to be reinstated via the use of more explicit forms if they are no longer maintained in short-term memory, but conversely favour reduced forms if they are. As per Chafe (1976), the activation status of referents gradually decays over time, and reactivation incurs a higher processing cost than maintenance (Section 2.2.1). The relationship between memory scope and sentence comprehension is well documented: Phonological priming effects are reported to disappear within four to seven words (Tanenhaus et al. 1985), a rate that is sometimes modelled as exponential decay (e.g. in Dell 1986 and Lappin & Leass 1994). This follows from long-term memory not storing literal forms, but rather unmarked conceptual "gestalts" (Sachs 1967; Jarvella 1971; Trabasso et al. 1971; Garrod & Trabasso 1973). Similarly, it has been demonstrated that the rate of errors in memory tasks is proportional to the distance between anaphor and antecedent (Myachykov & Posner 2005: 326), as do processing times for pronominal anaphors in reading tasks increase (Garnham 1987). Studies on eye-tracking further show that the time and location of fixations correlate with anaphoric distance, increasing with greater distance to the antecent (Stevenson 1996).

Accessibility theory (Ariel 1990; Section 2.2.4) links anaphoric distance to accessibility in an inverse relationship. All things being equal, more recently mentioned referents should be more readily accessible than more distantly mentioned referents (Ariel 1990: 57). The inverse holds true as well: Closely linked chains of co-referential mentions, especially in subject position, contribute to local discourse coherence (Givón 1983c, 2017; Section 2.4.2.3), and hence strongly tend towards the use of reduced forms, so that that zero and pronominal anaphors are especially common where the antecedent is located in the previous sentence; similar observations also hold for object mentions (Bender 1999; Huang 2000a: 78–89; see also Schnell & Barth 2018 and Schwenter 2016).

Givón (1983c) finds that references to recently mentioned information are more likely to be pronominal than information that is less recent. For a sample of written English, Givón reports the preferred choice of form to transition from reduced to lexical NPs at distances of roughly $d = 7$ clause units. Biber et al. (1998: 106–132), in investigating pronoun use across text types in English, find substantial differences in the average distance (measured in the number of non-co-referential NPs between anaphor and antecedent) between a pronominal anaphor and its antecedent across text types, with news reporting allowing for the longest pronominal references at $d = 11$ intervening NPs, and conversational data the shortest at only $d = 4.5$ NPs; written texts in general allow for longer-distance pronominal anaphors than spoken texts. Similarly, Gipper (2016: 166–167), in a study based on spoken data from Yurakaré (Bolivia, isolate), observes that anaphors in core roles at distances above $d = 5$ clauses from the antecedent show a significantly increased likelihood of being realized overtly. Yurakaré in general has a very strong preference for zero in core argument roles; this preference is elevated further as anaphoric distance increases, and expectedly peaks at distances of $d = 1$.

However, although reduced expressions are chiefly associated with high-accessibility contexts, Ariel (1990: 19) finds that a non-trivial proportion of pronominal expressions occur at distances greater than what would be predicted by accessibility theory, and as such surmises that recency is not the only factor affecting accessibility (see Section 2.2.4). The salience and topicality of the referent interact with recency, Ariel observes, so that when mentions with highly topical antecedents are filtered out, the proportion of high-distance prominal anaphors decreases substantially, while the overall distribution of form types across other distances remains largely the same. Lastly, Schiborr

(2017) finds that outside of clause chains, recency effects in spoken English are most determinative of the broad choice between lexical and reduced expression, and cannot predict most of the finer distinctions among expressions indicated on the accessibility scale (Ariel 1990: 73, ex. 1; cf. Figure 2.1 in Section 2.2.4 above).

All of the examples brought up so far in this section quantify recency in terms of textual distance. Fox (1987b), however, argues that linear distance is an inadequate representation of the mechanisms involved, and that rhetorical distance, measured in terms of the number of intervening nodes along rhetorical structure graphs (Mann et al. 1992; see Section 2.2.5.3), conversely offers a more cogent explanation of referential choices. Kibrik (2000: 74) voices agreement with Fox's assessment, but also contends that linear textual distance should not be dismissed out of hand, though its influence is "modest" by comparison (cf. Kibrik & Krasavina 2005). Kibrik et al. (2016: 9) further qualify this observation, noting that while distance factors are essential for the successful prediction of referential choices, their tests of multiple distance measures in unison reveal that textual and rhetorical distance are highly correlated, so that "using any of them increases accuracy dramatically," but predictions improve even further "if two or three distance factors are included" at the same time.

As mentioned before, anaphoric distance is also of central importance in most approaches to anaphora resolution and referring expression generation (Section 2.3.2). For instance, Krahmer & Theune (2002) note that the informativity of the expression selected for a referent is inversely proportional to the distance since its last mention in the text. Early rule-based models of anaphora resolution (e.g. Lappin & Leass 1994; Ge et al. 1998) likewise assign anaphoric distance the greatest predictive weight, with the relative probability of a candidate antecedent being the correct one decreasing quadratically with every sentence separating it from the anaphor in question. As such, they expect most pronominal anaphors to occur at short distances from the antecedent.

2.4.2.2 | Syntactic prominence

Some of the arguments of a clause tend to be perceived as more salient and hence accessible than others, chief among all the grammatical subject, followed by direct objects (e.g. Brennan et al. 1987; Brennan 1995; Gordon et al. 1993; Arnold 2010; Arnold et al. 2000; Krahmer & Theune 2002). Other syntactic posititions conferring prominence include clefts (e.g. *it's Jane we're looking for*), which make antecedents overall more distinctive in working memory (Foraker & McElree 2007; see also Arnold 1998, Arnold 1999; Almor 1999; Cowles et al. 2007), as well as topicalizations (e.g. *my father, I used to go with him*), such as those used frequently in Japanese (Walker et al. 1994). As such, syntactic prominence is a major contributor to topichood (cf. Givón 1983a), and one of the chief determinants of referent salience in both accessibility theory (Ariel 1990; Section 2.2.4) and centering theory (Grosz et al. 1983; Section 2.2.5.2).

However, it is not only the syntactic prominence of the anaphor itself that influences referential choices, but also that of its immediate antecedent. Halmari (1996: 76), in correlating the accessibility scale (Ariel 1990) with the Keenan-Comrie-scale (Keenan & Comrie 1977), which ranks the grammatical relations of NPs by their accessibility to different syntactic operations, finds that the grammatical role of the antecedent affects the interpretation of anaphors, and concludes that, by and large, subjects make for more salient antecedents than less syntactically prominent roles. In a similar vein, Gundel (2003) notes that non-subjects are more likely candidates for antecedence if referred to by lexical NPs and other expressions indicating low givenness.

The combined relevance of the syntactic positions of antecedent and anaphor hence also play a substantial role in maintaining discourse coherence through same-role clause chains (Givón 2017; see Section 2.2.3 above), so that reduced forms become especially likely in cases of syntactic parallelism (Arnold 1998, 1999; Arnold et al. 2009; see also Kehler 2002 on parallelism superseding syntactic structure as such; see furthermore Chambers & Smyth 1998 for related evidence from psycholinguistics). For example, Arnold (2003: 238) reports a parallelism effect for objects in Mapundungun (isolate, Chile), where object arguments are more likely to be realized as zero if they are co-referential with the object of the preceding clause. Gipper (2016: 166–167) finds a similar effect for subjects in Yurakaré (Bolivia, isolate), but none for objects, suggesting that syntactic parallelism may have a typological dimension.

In addition to syntactic prominence, referential choice has also been found to be affected by the semantic role of an entity (Arnold 2010: 191; Kehler 2002).[9] Experimental data gleaned from sentence completion tasks shows that speakers take the thematic structure of events into account when planning out discourse (Stevenson et al. 1994). For instance, in transitive clauses with a Stimulus and Experiencer, speakers tend to use pronouns more frequently for references to the former than the latter, once syntactic prominence has been accounted for. As Arnold (2010: 191) notes, this preference also reflects the biases of listeners (cf. Stewart et al. 2000).

2.4.2.3 | Discourse prominence

Beyond recency and syntactic position, the prominence of a referent is also influenced by the frequency with which it is mentioned in discourse. High mention frequencies prime addressee's expectations, so that highly frequent referents are more predictable, and hence more readily accessible (Arnold 2010: 191; Yoshida 2011; see also related considerations in Fox 1987b; Poesio & di Eugenio 2001; Stevenson et al. 2000). This is especially the case in the context of anaphoric chains, where co-referential mentions occur in close succession in the same position. Discourse prominence is hence a key building block of the notion of discourse coherence in the sense of Givón (1983a, see Section 2.2.3). Evidence for this association comes from both experimental and corpus-based studies, which consistently find that multiple repeated references to same entity yield progressively shorter, less specific, and more phonologically attenuated forms (e.g. Clark & Wilkes-Gibbs 1986; Ariel 1990; Bard et al. 2000; Aylett & Turk 2004). Ariel (1990: 57), for instance, finds that in written English, 82% of co-referential mentions in adjacent clauses are pronominal or zero. In this view, discourse prominence is chiefly determined by the token frequency (and recency) in a localized stretch of discourse, rather than global token frequency across an entire text (Lichtenberk 1996; Himmelmann 1997), already touched on above in Section 2.4.1.3.

A related observation is that the form of the antecedent influences the form of the anaphor (i.e. form parallelism; Ariel 2001). Specifically, if the

9 An investigation of predicate semantics was originally planned to be part of the present study, but the development and application of the necessary corpus annotations unfortunately did not proceed sufficiently far by the time of its submission.

antecedent is a reduced form, then it is likely to be topical and accessible, which in turn increases the odds of the anaphor also being accessible, assuming a short distance between the two. While this association appears to be relatively consistent for subject anaphors, languages appear to differ in the extent to which other arguments are sensitive to parallelism effects (Arnold 2003 for Mapundungun vs. Gipper 2016 for Yurakaré).

Lastly, there is a semantic dimension of clause chaining related to the spatio-temporal sequencing of sentences in discourse. Schnell & Barth (2020: 274) note that "the most predictable referent is one which is a continuous topic [...] in a sequence of foregrounded, narrative clauses involving the same referent(s)", whereas referents mentioned in clauses involving backgrounded information (i.e. that do not progress the narrative) are less predictable; Myhill (1997), for instance, finds that zero subjects are particularly likely in clauses maintaining a temporal sequence.

2.4.2.4 | Competition between referents

Where there are multiple potential candidates for antecedence with overlapping properties and similar levels of accessibility, speakers supposedly employ more informative and specifically referring expressions to disambiguate, as less informative expressions – whose use might be expected from considerations of referent accessibility alone – would not be enough for addressee to identify intended referent with certainty (Ariel 1990; Arnold & Griffin 2007; Arnold 2010). In essence, the presence of additional potential antecedents reduces the accessibility of all candidates, including the intended one (Gipper 2016: 33). Kibrik (2011: 380) calls this the "referential conflict filter", but understands it to constitute a separate layer from accessibility and activation-based considerations (see Section 2.3.1).

In the strictest sense, this kind of ambiguity can only arise between referents that are equally "semantically compatible" in terms of their "humanity, agentivity or semantic plausibility as object or subject" (Givón 1983a: 14). More generally, one might expect competition to arise more commonly in contexts where a large number of referents are activated concurrently, that is where local information pressure (Durie 2003) or referent pressure (when applied to anaphors, not just new introductions; Du Bois 1987b, 2003a, 2003b) is high; these contexts are also expected to be particularly taxing on attention, increasing cognitive load (Section 2.4.2.5). Of course, since the distinctive-

ness of pronominal expressions in particular differs between languages (e.g. in terms of gender marking, kinship relations, relative social standing, etc.), the effects of high referent competition on referential choices are likely to vary in strength cross-linguistically, as the range of potential candidate antecedents is constrained by the whichever distinctive information a pronoun carries.

Another kind competition between referents – one that has been underrepresented in the literature on reference processing – is effected by similarities in content or context. One particular kind is caused by what Fukumura et al. (2011) label similarity-based interference, which is caused by referents occuring in similar situations (e.g. in terms of spatial location, such as with one or both competitors sitting on a horse). In particular, they find that situational congruence between referents increases the rate of lexical expression (vs. pronouns) in English. However, they conclude that this effect results from speaker-internal production constraints, since it applies even when candidates are sufficiently distinctive, in Givón's (1983a: 14) terms, to receive distinct pronouns, and is as such independent of discourse-based ambiguity avoidance.

2.4.2.5 | Cognitive load

Arnold (2010: 196–197) notes that under circumstances where cognitive tasks compete for attentional resources (cf. Baddeley & Hitch 1974), "processing load will decrease available resources for maintaining activation, and thus increase the use of explicit forms." One such case of high cognitive load is found in contexts in which a large number of referents are active concurrently, such as when establishing new scenes or rapidly switching perspectives in a narrative.

Another dimension relates to the overall complexity of the clause, which Arnold et al. (2009: 142–143) quantify in terms of word length. For their study of English narrative retellings of a short cartoon film, they report that the rate at which grammatical subjects are expressed with reduced forms correlates with the length of the clause, with pronouns and zero being more frequent in shorter than in longer clauses, for a total difference in lexicality rates of about 8% between clauses with fewer or more than seven words. They relate this finding to the cognitive effort expended in planning and producing speech, which is expected to increase proportionally with utterance length (Clark & Wasow 1998; Watson & Gibson 2004). As the planning of longer utterances

results in elevated processing demands, so they argue, fewer resources are available for maintaining referent activation, and hence longer clauses are more likely to have lexical subjects.

2.4.2.6 | Priming from semantic frame relations

Considerations of topicality and discourse prominence are primarily concerned with co-reference, but the conceptual space occupied by a referent can also only partially overlap that of other referents. This the case, for instance, for referents that are inferable from frame semantics (Fillmore 1982), which are called "bridging anaphora" in Huang (2000a) (not to be confused with bridging constructions in the sense of syntactic tail-head linkages, e.g. Guérin 2019). These are referents that are only indirectly evoked by the discourse context, as entailments or mereology Poesio et al. (2004), such as through the relation between a group and its members (e.g. *Lucy + Jane* vs. *they*) or the parts of an entity (e.g. a *dog* has a *tail*, *four legs*, etc.).

As repeated mentions of a referent increase its accessibility, bringing it to the centre of the addressee's attention (cf. the role of mention frequency in Chafe 1976; Ariel 1990; Grosz et al. 1983; etc.), mentions of other referents related to it through bridging relations or semantic entailments presumably do the same (if likely to a lesser degree), in effect priming its subsequent mention, which in turn affects forms choices. For instance, a recent mention of a dog's *tail* indirectly makes the *dog* itself, as well as any of her other body parts and anything else directly associated with her, more accessible. As far as I can tell, this notion has not been systematically investigated in the context of referential choice, though provisions for testing it are provided within the RefLex annotation scheme (Riester & Baumann 2017; see Section 3.3.4.2), which explictly marks bridging relations between referring expressions.

2.4.2.7 | Establishment following introduction

Following the introduction of a new referent into discourse – which, as noted earlier, are predominantly lexical – do speakers shift to reduced forms immediately, or is there a transitional period during which lexical forms are comparatively more common? This idea has, to my knowledge, only seldomly been tested. Kibrik (2000: 78) calculative model of referential choice, for instance, takes note of whether the antecedent of the anaphor in question is the introduction of the referent into discourse, and finds that "when a referent is first introduced into discourse it takes no less than two mentions to fully activate it," until which point it is less likely to be pronominally expressed (cf. also Lichtenberk 1996).

2.4.3 | Miscellaneous properties

2.4.3.1 | Transitivity of the predicate

As per Du Bois's (1987b) non-lexical-A constraint (Section 2.2.5.1), there is an association between the subjects of transitive clauses and non-lexical forms, so that the majority of lexical subjects occur in intransitive clauses. This applies even when factoring out predominantly lexical referent introductions, which according to Du Bois are also associated with intransitive clauses (cf. also Lambrecht 1994: 184–191; see criticism of this view in Section 2.2.5.1).

Gipper (2016: 167) suggests that the dispreference for lexical A arguments might relate to both information structure and processing, as the increased likelihood of the more explicit argument in a transitive constructions being the object facilitates processing in languages with free word order and a lack of case marking; as such, speakers supposedly exploit the association between transitivity and referential choices as a means of ambiguity avoidance. However, what Gipper (2016) finds in actuality is that objects in Yurakaré are only consistently lexical if they are not topical (see Section 2.4.2.3) This aligns with related conclusions in Schnell et al. (2021b), who argue that considerations of transitivity do not affect speakers' choice of referring expressions directly, as the choice of a transitive or intransitive predicate is an independent one, guided chiefly by narrative concerns.

2.4.3.2 | Quantity constraints

In a similar vein, constraints on how much referential information can be reasonably included in a given stretch of discourse or clause have been proposed; the one-lexical-argument constraint in Du Bois (1987b; Section 2.2.5.1), for instance, suggests a universal limit of one lexically expressed argument per clause on average. All else being the same, a lexically realized object should hence be more likely to be paired with a non-lexical subject, and vice versa. Stoll & Bickel (2009: 554) note that while this upper limit on informativity appear to be well established, there is substantial variation among languages in how discourse is structured below the limit, specifically with regards to "the amount of lexical information that speakers are expected to give away when telling a story beyond what is strictly needed for tracking the identities of referents across events."

2.4.3.3 | Clause type

Miltsakaki (2002: 328–329) contends that mentions of referents in subordinate clauses are of lower salience than mentions in the matrix clause, which affects referential choices (though Miltsakaki argues specifically from an anaphora resolution perspective), and that the linear order of subordinated and matrix clause does not matter in this regard (see also Miltsakaki 2003).

Relatedly, Torres Cacoullos & Travis (2019: 659–665; fig. 5 and 6) report that zero anaphors are favoured over pronouns as subjects of coordinated clauses (i.e. via *and*) in samples of spoken Spanish and English. Schnell & Barth (2020: 283–286) find that the opposite is true in Vera'a (Oceanic, Vanuatu), where it is instead pronominal forms that are significantly more frequent; they surmise this to be due to the clause connectives in Vera'a chiefly bearing discourse-structuring function, rather than merely being coordinators. Similar patterns of zero and pronoun use in connected clauses are also found in Kurdish and Turkish (e.g. Matras 1997).

2.4.3.4 | Informativity of pronouns

Another factor with potential influence on the selection of referring expressions is the richness of a language's pronominal system. This criterion has been mostly applied to the selection of zero (vs. pronouns, cf. Stoll & Bickel 2009), but presumably also affects the selection of lexical expressions, as the more informative pronominal expressions are relative to simple lexical expressions, the more mileage speakers get out of their use, making the transition from pronominal to lexical forms happen at comparatively lower degrees of referent accessibility. As such, ceteris paribus, languages with richer pronominal systems – ones that, for instance, distinguish gender (as in English), honorific degree (as in Japanese),[10] or elevation (as in Sanzhi Dargwa) in addition to number – would employ fewer lexical expressions than languages with sparser systems. However, as Stoll & Bickel (2009: 544) note, the number of categories encoded in pronouns often do not matter in practice, as whatever referents need to be distinguished commonly share the same properties, thus rendering pronouns insufficiently informative in practice.

2.4.3.5 | Presence of verbal agreement

Related to the informativity of pronouns is the presence and complexity of verbal agreement markers. This has also been mostly tied to languages' preference for zero anaphora (i.e. in the context of pro-drop, e.g. Huang 1984; Jaeggli & Safir 1989). Richer agreement can – but does not necessarily have to – increase a language's preference for zero; while there are notable cross-linguistic differences in this regard, the general direction of the association seems to be universal, however. For lexical expressions, no connection between rate of zero and rate of lexical expression has been noted in the literature, and neither is there any apparent in the data used for this study; see Chapter 5 further below.

10 But consider confounding factors such as the higher degree of politeness associated with indirect language (especially the use of zero) in Japanese.

2.5 | Lexical expressions

As mentioned above, a number of approches have argued that referential choices are fundamentally about the basic distinction between lexical and non-lexical referential forms (Kibrik 2011; Kibrik et al. 2016; Torres Cacoullos & Travis 2019; Schiborr 2017). In fact, all hierarchies that relate accessibility to referring expressions (e.g. Givón 1983a; Ariel 1990; and others mentioned above) acknowledge this basic distinction, though it tends to be obscured by the finer distinctions made in their taxonomies (Torres Cacoullos & Travis 2019).

Full NPs reside at the higher end of the informativity spectrum, and are ostensibly the ontologically most basic expression type, as speakers strive to reduce the information content of their utterances as much as possible while maintaining the comprensibility of the underlying proposition (cf. Bolinger 1979; Ariel 1990). At the same time, lexical expressions are not the default choice in practice, being outnumbered by reduced expressions in most contexts (Kibrik 2011). In the frameworks of reference discussed in the previous sections, they tend to be used primarily for references to entities with low accessibility that could not be correctly identified via a reduced form, such as those that have not been mentioned for a while or are new to the discourse (Chafe 1976; Givón 1983a; Ariel 1990).

In fact, since new information is generally lexical, lexicality is sometimes conflated with newness (e.g. Francis et al. 1999); but lexicality does not entail newness, and in fact most lexical expressions refer to given information (Schnell et al. 2021b; Du Bois 1987b: 830). Lexicality is also strongly associated with non-subject positions due to the association of subjecthood with topicality, and topicality with reduced forms (Givón 1983a; Gundel 1988; Lambrecht 1994). Lexical subjects are hence quite rare, and so constitute what Francis et al. (1999: 10–13) describe as a "potentially highly anomalous class", but they nevertheless exhibit many of the regularities of non-lexical subjects, in that they are no less agentive than other subjects, and chiefly express topics.

Of course, not all lexical expressions are created equally, and in fact they constitute a rather heterogeneous group, as nouns are an open class, and NPs headed by nouns allow for theoretically infinite modification and elaboration. This variety is represented on Ariel's (1990) accessibility scale, which differ-

entiates a number of types lexical expressions (see Figure 2.1 in Section 2.2.4 above). The remainder of this section briefly touches on the status and treatment of three dimensions specifically (among the many others and their various constellations): proper names (Section 2.5.1), heavy NPs (Section 2.5.2), and demonstrative NPs (Section 2.5.3).

2.5.1 | Common nouns and proper names

Proper names are generally considered to be the type of referring expression whose reference is the most explicit and stable, bar highly elaborated common lexical NPs (Arnold & Griffin 2007; see also Anderson & Hastie 1974; McCoy & Strube 1999; Poesio 2000; and Heller et al. 2012 for various treatments of proper names in discourse). On the accessibility scale (Ariel 1990; see Figure 2.1 in Section 2.2.4), for instance, proper names rank among the forms marking lowest degrees of accessibility.

In most contexts, proper names are the least ambigious means of reference, being both very specific and essentially immutable in their identification of the intended referent. Even so, they do not refer fully independently of context, but are as much subject to considerations of discourse structure as other expression types (Werth 2020). Given names and surnames are limited in number, in some cultural contexts more than in others, and so the use of further identifiers or a combination of both (in cultural contexts in which a person has multiple names) may be needed to make the reference even more specific. This consideration is also reflected on the accessibility scale, where various configurations of personal names are differentiated in terms of accessibility marking.

The personal names of human individuals in particular bear a high degree of salience (Helmbrecht et al. 2018), but of course other entities such as animals, organization, places, events, and so on, may also carry proper names, which alternate with less explicit forms in the same way as personal names do (e.g. *London* vs. *the capital* vs. *it*). While presumably the same constraints apply for their selection, the inherent salience of human referents, as noted earlier, is likely to make a difference; compared to non-human entities, human referents will naturally tend more strongly towards less explicit forms under most circumstances. Indeed, Haghighi & Klein (2010: 386) find that the selection of proper name mentions in general "is governed more by entity frequency than antecedent distance," which suggests that they are selec-

ted more by narrative considerations (i.e. protagonisthood) than by discourse-contextual factors, though they may conceivably also play a role in disambiguating in cases of competition between referents. Additionally, proper names may differ morphosyntactically from common nouns (Schlücker & Ackermann 2017; contributions in Kempf et al. 2020, and may receive special marking (e.g. the proper name article in Vera'a, see Section 3.2.3.10).

2.5.2 | Phrasal complexity

Lexical expressions vary in terms of their informativity, which in most cases is a function of their phrase-structural complexity (e.g. *the book* vs. *the small book with the green cover I read yesterday*; O'Grady 1997; Arnold et al. 2000). An expression's degree of informativity is commonly quantified in terms of phrase weight, for instance through its length in words, morphemes, or phonemes. An alternative approach evaluates information content via the presence or absence of additional modifiers (e.g. attributive adjectives, possessive expressions, adpositional modifiers, relative clauses, etc.; cf. Kibrik et al. 2016).

On the accessibility scale (Ariel 1990), more complex, informative, and specifically referring expressions indicate successively lower levels of accessibility.[11] In the texts analyzed in Givón (1995), speakers chiefly use simple lexical NPs at distances below seven clauses, and more complex expressions at ten clauses and above. The less accessible a referent hence is, the more likely it is to be comparatively more informative and explicit.

2.5.3 | Demonstrative NPs

The third and last type of lexical expressions we will focus on here are those with demonstrative determiners (e.g. *this/that book* vs. *the book*). Demonstratives are generally associated with deictic references, but can also be anaphoric (Diessel 1999), and are frequently used for discourse deixis (Webber 1988). On the accessibility scale, plain lexical expressions mark lower accessibility than those with demonstrative determiners (Ariel 1990: 73); lexical NPs with demonstratives are hence supposedly used with higher frequency in

11 This also applies to the relative complexity of proper names, as discussed in the previous section.

contexts with higher referent accessibility, such as at shorter distances to the antecedent, compared to lexical NPs without them. In English, however, so Ariel (1990: 53) notes, the difference between expressions with and without demonstratives is expected to not be substantial. In a similar vein, Michaelis & Hartwell (2007: 15, tab. 3) observe that in conversational English, lexical NPs with demonstrative determiners are more common for referents that have been mentioned before, that is, are not new to discourse. Zulaica Hernández (2009) instead argues that demonstrative anaphors serve to mark topic shifts in discourse, informing the addressee about said shift by focusing attention on the referent that is the new topic. The use of demonstrative expressions hence does not follow exclusively from considerations of accessibility, they claim, but has an active role in structuring discourse.

Accessibility theory further distinguishes proximal and distal demonstratives, with the latter aligning with lower accessibility and more distant antecedents; though not explicitly stated, medial and other gradations among demonstratives presumably fall somewhere inbetween. This claim has been challenged since (see Botley & McEnery 2001: 226 for written and Schiborr 2017 for spoken English, Zulaica Hernández 2009 for written Spanish, and Jarbou & Migdadi 2012 for Classical Arabic), finding instead that accessibility is not a valid predictor for the selection of proximal versus distal demonstratives, but compare contrary findings in Lichtenberk (1996: 382) for To'aba'ita (Oceanic, Solomon Islands). Botley & McEnery (2001) test the selection of speaker-proximal and speaker-distal demonstratives (pronominal and as NP modifiers) in English news reporting, and find the data do not corroborate the claimed association between lower anaphoric distance and proximal demonstratives, and higher distance and distal demonstratives. In particular, there are few cases of long-distance anaphora involving demonstratives. However, they do find that demonstrative expressions occupy an intermediary position between personal pronouns and full NPs, and that lexical NPs with demonstrative modifiers occur at greater distances than demonstrative pronouns.

2.6 | Research questions

The aim of this study is to investigate how discourse references are encoded in spoken language. It does so from a variationist typological perspective, comparing patterns of usage in spoken discourse data from multiple, typologically diverse languages. Specifically, the research presented here concerns itself with the circumstances under which speakers choose to employ full NPs for anaphoric mentions. The subject matter of this study hence deviates somewhat from the main strand of research on referential choices and discourse typology, which tends to integrate lexical expressions into more granular hierarchies alongside other forms of reference (e.g. Ariel 1990; Givón 1983a; etc.; cf. also recent work in referring expression generation, Section 2.3.2.1). But as noted earlier, it has been argued elsewhere that the alternation between lexical and non-lexical expressions is in fact the most fundamental and universal referential choice speaker have to make (Kibrik 2011; see also Schiborr 2017).

This study addresses three major research questions, which divide it into three parts, with a final chapter providing synthesis: First, can we identify cross-linguistic patterns in the rate of use of lexical expressions for discourse anaphors in various syntactic positions, and how do said patterns differ from the corresponding rates of zero and pronominal anaphors (Chapter 5)? Stoll & Bickel (2009), for instance, contend that languages differ substantially with respect to both the frequency at which lexical expressions are used and the contexts they are used in, but base their conclusions on a highly limited number of languages.

Second, how well can we predict the selection of lexical expressions using multifactorial models built on the basis of discourse-contextual and semantic factors? And just as importantly, what can we learn about the specific mechanisms of referential choice from these models (Chapters 6 and 7)? This echoes similar efforts in Kibrik et al. (2013) and Kibrik et al. (2016), among others, discussed above in Section 2.3.1, which find some success with this approach, but rely exclusively on written data from well-researched languages. Of special interest here are also potential interactions between factors, in particular of contextual properties (i.e. anything related to the structure of the local discourse) and the inherent properties of the referent (such as animacy and protagonisthood): For instance, do human referents behave fundamentally differently from non-human referents, or only in specific contexts, if at

all? Furthermore, psycholinguistic evidence suggests that subjects are given a preferential role in comprehension and cognition relative to other, less privileged roles such as objects (e.g. Kwon et al. 2010; Wang et al. 2009), which supposedly also reflects on the modalities of discourse production (cf. Du Bois 1987b); but how fundamentally, if at all, do the mechanisms of form selection for lexical NPs vary between anaphors in different syntactic positions?

And third and finally, do the patterns identifiable for the broad lexical–non-lexical choice also apply to more nuanced choices between different types of lexical expressions (Chapter 8)? Lexical expression are hardly a homogenous group after all, differing not only in the type of head (e.g. common nouns vs. proper names), but also in the presence and relative complexity of additional modifiers (e.g. demonstrative determiners, possessives, relative clauses, etc.). Accessibility theory (Ariel 1990) posits that proper names, for instance, are selected according to the same considerations as lexical expressions in general, but under more extreme circumstances of referent inaccessibility, that is, that they are simply the most informative expressions at speakers' disposal (cf. Schiborr 2017 for preliminary evidence to the contrary).

Before any of these questions can be addresses, however, the next two chapters deal with the methodological foundations of this study: Data sources and annotations – alongside a short outline of the field of corpus-based typology – in Chapter 3, and the particulars of study design, including sample selection criteria and factor definitions, in Chapter 4.

3 | Corpus-based typology

This chapter begins with a brief outline of the state of research in corpus-based approaches to typology and a short list of corpus projects geared towards investigations of discourse structure (Section 3.1). It then describes one recent corpus project in particular, the Multi-CAST collection (Haig & Schnell 2015), on which the present study is based: its motivations and aims, design philosophy, and contents (Section 3.2), including an outline of the ten corpus languages leveraged for this study (Section 3.2.3) and a working description of the various annotations applied to the data (Section 3.3).

3.1 | Background

Corpus-based typology is the result of the application of the increasingly sophisticated methodologies of corpus linguistics, variationist sociolinguistics, and statistical analysis, to the traditional grammar-based field of typology. It views languages as collections of utterance tokens, rather than as monolithic systems shared by all speakers, and hence treats them as bundles of statistical tendencies rather than individual data points (Schnell & Schiborr 2022: 173).

Research in the framework of corpus-based typology has yielded insights on language use and its functional accounts and the universals of text production; while recent studies have cast doubt on the reality of concrete struc-

tural universals shaping language systems (Bybee 2009), corpus studies have demonstrated the existence of universal tendencies of usage. This includes work on frequency distributions of word tokens and word lengths and Zipf's law of abbreviation (Zipf 1935), which have been shown to hold across a wide range of languages (Bentz & Ferrer-i-Cancho 2016; Piantadosi 2014), on word order variation entropy (Levshina 2019) and dependency lengths (Futrell et al. 2020b), on marking asymmetries as determined by Greenberg (1966) (Mansfield et al. 2020; Haspelmath 2021), on functionalist syntax in the tradition of Chafe and Givón, as well as on prosodic chunking (Seifart et al. 2018; Seifart 2021; see also Himmelmann 2014) and social cognition (Barth & Evans 2017a; Barth et al. 2021). See Levshina (2021), Schnell et al. (2021a), and Schnell & Schiborr (2022) for overviews of the field and its challenges.

Corpus-based typology hence complements the more traditional, system-based approaches to typology in the tradition of Greenberg (1963), whose aim is the distillation of data gleaned from grammars, genealogical factors, and more recently large-scale feature-oriented databases such as WALS (Dryer & Haspelmath 2013) and Grambank (Skirgård et al. 2023). See Wälchli (2009) for a study contrasting the two approaches. Corpus-based typology ultimately rests on the notion of usage-based (or emergent) grammar (Hopper 1987), which holds that discourse and grammar are closely interlinked domains, in that discourse properties are shaped by grammatical properties and their usage in particular contexts, and in turn, grammatical properties emerge from discourse properties (Ariel 2009; MacDonald 2013; Mithun 2015; McDaniel et al. 2015; cf. also Hawkins's 2004 performance-grammar correspondence hypothesis); as Bybee (2006: 730) puts it, "[u]sage feeds into the creation of grammar just as much as grammar determines the shape of usage."

As such, corpus-based approaches to typology focus on individual instances of usage, using corpora as approximations of underlying structures: In the terminology of Levshina (2019), they are in essence the "token-based" counterpart to traditional "type-based" typology, characterizing texts and collection of texts rather than languages.[1] Where traditional approaches to typology compare broad abstractions of entire languages or varieties of languages,

1 This approach is not limited to text corpora; consider, for instance, the experimental work done in neurotypology (Bornkessel-Schlesewsky & Schlesewsky 2013).

corpus-based approaches instead compare concrete instances of actual usage, either as texts and collections of texts (i.e. corpora). Unlike type-based approaches, they also take into account contextual and variationistic conditions, and offer probabilistic as opposed to categorical generalizations.

There are two perspectives on corpus-based typology: One that sees corpus studies as experiments on processing constraints, and one that sees them as incidental observations of verbal behaviour. The former observes that preferences in production lead to constraints on processing, leading in turn to constraints on the diachronic development of grammatical structures, which ultimately manifest as observable structures and their typological distributions (Haspelmath et al. 2014; Mansfield et al. 2020). The latter contends that mutual influences on production behaviour amongst speakers are based on mutual observation, and that consequently, divergent variants are either ignored or adapted by other speakers, and in the latter case are gradually spread throughout a speech community (Croft 1990; Labov 1994). This, in essence, amounts to drawing typological generalizations from observations of speakers' behaviours across texts and corpora (Piantadosi et al. 2011; Futrell et al. 2015; Levshina 2019).

The development of corpus-based typology has been enabled by the increasing availability of (digital) multilingual corpora in recent years, some of which will be outlined in the following section.

3.1.1 | Multilingual corpora for typological research

Corpora come in a variety of difference shapes and sizes, which is important to keep in mind with regards to the representativeness of the data and the comparability of different text types (Levshina 2019: 541–542). Broadly speaking, three types of multilingual corpus resources can be distinguished, within which still further distinctions can be drawn (e.g. in terms of the degree of interactivity, i.e. monologues vs. dialogues/multilogues).[2]

The first type is parallel corpora of texts that are translational equivalents; these are most commonly written fiction, for instance prose literature (Stolz

2 Note that most of the examples listed in the following are either based on written data, or else chiefly focus on well-studied languages from Europe, Asia, and the Americas (cf. Dahl 2015). By comparison, spoken multilingual corpus data is still a rarity.

2007) or religious texts (Cysouw & Wälchli 2007), but may also draw from other sources, such as UN and EU legal documents (Ziemski et al. 2016; Steinberger et al. 2006), proceedings of the European parliament (Koehn 2005), or movie subtitles (Levshina 2016), among others.

The second are parallax corpora, which are built from stimulus-based elicited texts (cf. Barth & Evans 2017a). Perhaps the two best known stimulus texts are the Frog stories (Berman & Slobin 1994), retellings of a picture book for children (Mayer 1969), and the Pear Film (Chafe 1980; used e.g. in Bickel 2003 and Noonan 2003). A more recent example is the SCOPIC corpus (Barth & Evans 2017b; San Roque et al. 2012), which builds on data elicited from a narrative problem-solving picture task.

The third and final type is corpora composed from original texts that result from conventional text production routines, such as storytelling, oral history, and unscripted dialogues, and are as such not controlled for content. Notable corpora of this kind include the CorpAfroAs (Mettouchi et al. 2015) with spoken data from thirteen Afroasiatic languages, the conversational data used in Dingemanse et al. (2013, which is unfortunately not freely available), as well as the various corpora collated by the DoReCo initiative (Seifart et al. 2022; Paschen et al. 2020) from language documentation data. Universal dependency (UD) treebanks (Zeman et al. 2020; Futrell et al. 2015, 2020a) treebanks can also be leveraged in this way (e.g. Levshina 2019), though are at least in part based on pre-existing corpora. Comparable corpora have also been constructed for dialectological studies, such as FRED (Hernández 2006; Anderwald & Wagner 2007) and the various ICE corpora (Greenbaum 1996) for English. This study also draws from original text data, which will be described in Section 3.2 below, following a brief review of some of the corpora designed specifically for studies on anaphora and co-reference in the next section.

3.1.2 | Corpora for research on referential choice

Corpora specifically designed for studies on anaphora, referential choice, and related topics are few and far between, likely owing to the relative complexity (and labour intensity) of the necessary annotations for referential relations. Perhaps unsurprisingly then, corpora based on written data vastly outnumber (and outsize) those based on spoken data in this field, and the latter are almost exclusively monolingual, drawing mostly for English and a few other

languages for which data is readily available. It should also be noted that a number of the corpora listed here are not publicly available, or at least not in their fully annotated form.

A number of notable corpora for research into discourse structure are based on existing treebanks (all from English), such as the RST Discourse Treebank (Carlson et al. 2002), the Discourse Graphbank (Wolf et al. 2005), or the Penn Discourse Treebank (Prasad et al. 2008, 2019), which themselves already note various relations between discourse entities (i.e. not strictly speaking between co-referential expressions). Nicolae et al. (2010) extends the Discourse Graphbank data with custom annotations for co-reference relations; in a similar vein, Loukachevitch et al. (2011) and Kibrik et al. (2013) augment a subset of the RST Discourse Treebank data with their own annotations, which are based on an updated variant of the PoCoS annotation scheme (Krasavina & Chiarcos 2007). In addition to treebanks, Kibrik et al. (2016) also relies on corpus data from newspaper articles.

Among the non-anglophone corpora, the Prague Dependency Treebank (Hajičová et al. 2000) uses Czech treebank data, and the Kyoto Text Corpus (Kawahara et al. 2002) contains data from written Japanese annotated for co-reference and argument structure using a version of the annotation scheme codified in Iida et al. (2007). The Potsdam Commentary Corpus (Stede & Neumann 2014) draws from German newspaper editorials, and offers syntactic parsing as well as annotations for co-reference, rhetorical structure, and entity relations similar to those in the Penn Discourse Treebank mentioned above.

A large proportion of discourse-related corpus project come out of research in computational linguistics and NLP, geared towards the specific aims of these fields. The GNOME corpus (Poesio 2000, 2004) was designed to facilitate study of factors that affect referring expression generation, in particular the salience of referents. It is built on the basis annotated data from three highly specific text types (museum labels, pharmaceutical leaflets, and tutorial dialogues), but is not publicly available. The successor to GNOME, the ARRAU corpus (Poesio & Artstein 2008) marks a complex set of anaphoric features, with special attention paid to abstract anaphors. It contains data from multiple text types: elicitations from task-oriented dialogues, Pear Story retellings, news reporting, as well as subsets of the RST and Penn treebanks.

While this list is far from exhaustive, it is notable that there is essentially no overlap between the corpus resources listed in this and the previous section – I am not aware of any multilingual corpus project specifically geared towards research in discourse structure and co-reference, especially one focused on spoken language, except for the one leveraged for this study. Corpus-based typological research on anaphora and referential choice requires data sets from a large, ideally typologically representative (i.e. genetically and areally diverse) sample of languages, and any annotations applied to them need to be specific enough to capture a language's individual characteristics, but general enough to allow direct comparison across the entire sample. This study is based on data from a corpus project designed specifically to satisfy these needs, which will be outlined in the remainder of this chapter.

3.2 | The Multi-CAST project

3.2.1 | Introduction

The Multilingual Corpus of Annotated Spoken Texts (Multi-CAST, Haig & Schnell 2015) has been designed as research tool for this emerging field of corpus-based typology, intended to address both its needs and the limitations of similar corpus projects. Its primary aims are the expansion of available corpus data into larger and typologically more diverse samples of languages, and the formulation of a methodological framework that facilitates comparison between them. In particular, it was designed to enable cross-linguistic investigations of discourse structure, referential choice, and the interface of grammar and pragmatics, by providing common ground for quantitative analyses, similar to the comparative approach of Mettouchi et al. (2015) mentioned above.

Multi-CAST is a collaborative research project, centered at University of Bamberg, whose ongoing development has in large part run parallel with that of this study. As such, certain aspects of the design of the collection have been informed by the needs of this study, and the extents of this study are in turn defined by the potential inherent in the Multi-CAST annotations. Schnell & Schiborr (2018) contains a discussion of the motivations underpinning the project and annotations, and Schiborr (2019) is an extensive description of

the Multi-CAST collection, documenting its design, content, and technical foundations.

3.2.2 | Corpus design

The Multi-CAST collection encompasses corpora from over a dozen languages, ten of which are leveraged for this study, see Section 3.2.3 and Chapter 4 further below. The corpora consist of natural spoken language data that are non-elicited, unrehearsed, and original in the sense that they are not translations of existing texts (cf. Haig et al. 2011a). The data were recorded in situ in their respective cultural contexts, usually in the presence of an audience of native speakers, often as part of language documentation projects.[3]

As noted earlier, a substantial proportion of work in corpus linguistics is based exclusively on written data; for investigations into discourse structure and referential choices such as this one, the use of natural, unscripted spoken (or signed) data is essential, however, as only here are "considerations of language processing and efficiency-related trade-offs playing out under relevant time constraints of production and comprehension as well as noisy channel and similar considerations of the articulation bottleneck, principles of inference, and so forth" (Schnell et al. 2021a: 10). For instance, Just & Čéplö's (2022) study of bound object indexing in Maltese had to switch from using written Universal Dependency corpora to using spoken texts as the feature under investigation, while widespread in speech, proved too rare in writing for rigorous analysis. In a similar vein, Schnell & Schiborr (2022: 178–179) note that the kind of long, complex NPs that drive the initial claims on dependency length minization and its effects on word order regularities in Futrell et al. (2020b), while reasonably common in written texts, are barely present in spoken data; the greater tolerance for long NPs in written texts likely results from greater ease of written compared to spoken text comprehension, as in the former readers can adjust their reading speed at will, reread passages as needed, and so on, which not possible in spoken language.

3 But note that in typical language documentation settings, most narrative events are in fact "staged communicative events" in the sense of Himmelmann (1998), in that they would not have occurred if not by the request of an outside observer.

The texts in Multi-CAST are furthermore predominantly monologic; monologues avoid the inherent complexities of conversational texts – turn-taking, interruptions, cross-talk, and so on – in favour of a single strand of discourse managed by a single speaker, with only occasional input from secondary speakers. While as such not fully representative of all forms of spoken discourse, the narrow focus allows for a higher degree of comparability.

The narratives in Multi-CAST are original texts in the sense that they are not scripted, translated, or otherwise premediated. They instead have indigenous, spoken, and hence largely natural (or at the very least naturalistic) content. The use of original texts with essentially random narrative content does sacrifice data control, which is often regard as essential for comparative studies, but a major motivation for their use is that they better reflect the routinized verbal behaviour of speakers compared to experimentally elicited ones. In comparison, parallel (e.g. Cysouw & Wälchli 2007) and parallax corpora, such as collections of stimulus-based retellings of the Frog Story (Berman & Slobin 1994) or the Pear Film (Chafe 1980), offer better out-of-the-box comparability by channeling the speaker's choices along predetermined pathways, but in turn force them to react to the stimulus in a highly spontaneous manner, thereby limiting their ability to draw on the kinds of entrenched routines of speech production that of particular interest or referential choice. The Multi-CAST collection contains a number of closely related text narratives types, either narratives with traditional (i.e. indigenous folklore and fairy tales) or personal content (i.e. oral history, biographies, and autobigraphies). The differences between these two text types and their implications are briefly touched upon in Section 4.1.3 below.

Multi-CAST has a deliberate focus on small and lesser-studied languages, in an effort to escape the bias towards better-studied languages of Europe, Asia, and America in much of typological research (cf. Dahl 2015). This focus as well as the exclusive use of spoken data create what is perhaps the greatest limitation of the Multi-CAST data: With only about 10000 words on average per language, it is dwarved by many monolingual corpora, especially those based on written data. This weakness is a direct consequence of its strengths: Spoken data requires highly work-intensive processing before it can be used; the focus on critically understudied languages, many of which are endangered, further limits the availability of data. While the sample is at the point of writing still far from typologically representative, having a noticeable areal focus on Oceanic languages and the languages of Western

Asia, and so lacking data from most of the world's major language families, it nevertheless enables an as yet unprecedented perspective on cross-linguistic distribution of discourse patterns, owing to its careful design and uniform set of annotations. By comparison, even among the data sets prepared specifically for this area of study, such as the small selection discussed in Section 3.1.2 above, most rely on (chiefly written) data from languages overrepresented in linguistic research, the usual suspect being of course English.

As such, the Multi-CAST project leverages its inherent strength by placing its focus on corpus breadth and sample depth, rather than corpus size. In addition to trascriptions, idiomatic English translations, and standard morphological glossing, the data in the collection are richly annotated, with a very high degree of inter-corpus consistency. The Multi-CAST annotations aim to capture certain relevant features of morphosyntax and discourse structure, and do so in a manner that is applicable to a maximally diverse set of languages. Specifically, the annotations target the form, semantics, and syntactic function of referring expressions (Section 3.3.1), explicitly mark co-reference relations (Section 3.3.2), and note relational links between discourse referents, ontological animacy classes, and more (Section 3.3.4). These are layers of analysis that are not immediately recoverable from the standard morpheme-for-morpheme glossing or part-of-speech tagging common in other corpora.

In combination, these specialized annotations allow for a wealth of information to be drawn from a comparatively small sample of texts. That being said, to produce statistically robust results from which generalizations can be drawn, a certain minimum sample size is still inarguably necessary, and we will occasionally come up against the limitations of this quality-over-quantity approach in this study, if only for the most complex interactions between factors that split the data into minutely differentiated subgroups. Nevertheless, the uniform design of Multi-CAST and its annotations together enable a substantial degree of analytical depth across languages.

Lastly, I would like to mention *multicastR* (Schiborr 2018), a companion R package for the Multi-CAST collection, which was developed to provide easy access to the corpus data in R. Its functionality is described in Schiborr (2019), already mentioned above.

3.2.3 | Sampled languages

This section summarizes the affiliation, status, and pertinent typological characteristics of the ten languages from which corpus data is used for this study, as well as relevant properties of the corpus data itself, including a number of language-specific issues regarding analysis and annotations. The data used for this study correspond to version '2101' of Multi-CAST, originally published in January 2021. For more extensive discussions of the properties of each language, see the associated annotation notes for each Multi-CAST corpus (Haig & Schnell 2015). More generally applicable analytical issues are discussed in Section 3.3 as well as in Chapter 4 further below.

Data from the following ten languages are used in this study:

- Cypriot Greek (Indo-European, Greek; Section 3.2.3.1)
- English (Indo-European, Germanic; Section 3.2.3.2)
- Mandarin (Sino-Tibetan, Sinitic; Section 3.2.3.3)
- Nafsan (Austronesian, Oceanic; Section 3.2.3.4)
- Northern Kurdish (Indo-European, Iranian; Section 3.2.3.5)
- Sanzhi Dargwa (Nakh-Daghestanian, Dargin; Section 3.2.3.6)
- Tabasaran (Nakh-Daghestanian, Lezgic; Section 3.2.3.7)
- Teop (Austronesian, Oceanic; Section 3.2.3.8)
- Tulil (Papuan, Taulil-Butam; Section 3.2.3.9)
- Vera'a (Austronesian, Oceanic; Section 3.2.3.10)

Figure 3.1 provides a geographical overview of the area each of these languages is spoken, centered on the specific varieties recorded in the corpus data. The four-letter, four-digit codes given at the start of each of the following sections (e.g. `cypr1239`) identify entries in Glottolog (Hammarström et al. 2020).

The source of glossed examples drawn from the corpus data can be uniquely identified by the combination of the abbreviated name of the corpus, text name, and segment number listed underneath the example. The label "`mc_english_kent01_0240`", for instance, indicates that the example was taken from the utterance with the index "0240" from the "kent01" text in the Multi-CAST English corpus. Comprehensive lists of the texts and speakers included in the sample used for this study along with their associated metadata can be found in Appendices B.1 and B.2.

Figure 3.1 | Language locator map for the corpus data from Multi-CAST.

3.2.3.1 | Cypriot Greek

Cypriot Greek (Indo-European, Greek; cypr1239) is the variety of Greek spoken on Cyprus. It is not mutually intelligible with Standard Greek as spoken in Greece (Arvanti 2006). An overview of the language situation in Cyprus and a short comparative grammar with Standard Greek can be found in Hadjioannou et al. (2011). The texts in this corpus, annotated for MultiCAST by Hadjidas & Vollmer (2015), are folktales in the traditional mode, recorded in the 1960s and later published as part of a collection of Cypriot tales (Giangoullis 2009).

Cypriot Greek makes common use of clitic pronouns in non-subject positions, including as direct objects:

(7) *I arkondissa epernentin kathimera.*

i	*arkondissa*	*epernen*	*=tin*	*kathimera*
DEF.F.NOM	noblewoman	take.IPFV.3SG	=3SG.F.ACC	everyday

'The noblewoman took her everyday.' [mc_cypgreek_minaes_0004]

Occasionally, a co-referential NP may occur in post-verbal position alongside such a clitic pronoun, as in (8).

(8) *Epk'axento jindo psarin.*

epk'axen	*=to*	*jindo*	*psarin*
catch.IPFV.3SG	=3SG.N.ACC	that	fish

'[He] caught it, that fish.' [mc_cypgreek_psarin_0041]

Though both the clitic pronoun and the second NP are instantiations of the object role, only the former is analyzed as an object anaphor for the purpose of this study; however, since the latter occurs later in linear order, it may serve as the immediate antecedent of the next co-referential mention in a subsequent clause (see Section 4.6.2).

3.2.3.2 | English

The English (Indo-European, Germanic; `stan1293`) corpus (Schiborr 2015) consists of autobiographical narratives by speakers of varieties from Southern England (`sout3282`). The texts were originally recorded as part of various oral history projects in 1970s and 1980s, and later collated and transcribed for the Freiburg English Dialect Corpus (FRED, English Dialects Research Group 2005. They feature older working-class speakers recounting their personal experiences with turn-of-the-century and inter-bellum agriculture, animal husbandry, and shipwrighting.

English is notorious for its aversion to anaphoric zeroes (Harvie 1998), preferring pronouns as its default expression type instead, with most instances of pragmatically selected zero are restricted to specific syntactic contexts (Travis & Lindstrom 2016). In the sample used for this study (see sampling criteria in Section 4.1) of the 13% of English subjects that are zero, 69% occur in tighly linked chains of clauses with co-referential subjects as in (9):

(9) *My father'd slip up, get a pony in, go off, and have a look at this pony.*

my	*father*	*='d*	*slip*	*up*
1SG.POSS	father	=would	slip.INF	up

get	*a*	*pony*	*in*	*go*	*off*
get.INF	a	pony	in	go.INF	off

and	*have*	*a*	*look*	*at*	*this*	*pony*
and	have.INF	a	look	at	PROX.SG	pony

'My father'd slip up, get a pony in, go off, and have a look at this pony.'
`[mc_english_kent01_0190]`

Clearly identifiable zero objects in English are even rarer, accounting for only 2% of the objects in the sample. Part of the reason why this number is so low is that many transitive verbs in English also have an intransitive interpretation, for which no zero object is assumed in the annotations.

3.2.3.3 | Mandarin

The Multi-CAST Mandarin corpus (Vollmer 2020) consists of traditional narratives from native speakers of two regional varieties of Modern Standard Mandarin (Sino-Tibetan, Sinitic; mand1415), those of Xī'ān (xian1253) and Northeast China (nort3283). Standard Mandarin is in many ways an artificial construct; an idealized from of the language is taught in schools, but actual usage remains strongly influenced by regional languages. The three texts in this corpus, all traditional narratives, were recorded by Maria Vollmer in 2015 and 2016, and subsequently transcribed, translated, and annotated for Multi-CAST between 2016 and 2019.

Left-dislocated topic constructions as in (10) and (11) are relatively frequent in Mandarin (cf. Li & Thompson 1981); 7% of the clauses in the texts start with one. Their treatment and the issues they and similar constructions in other languages pose for the analysis are discussed in Section 4.6.2.1 below.

(10) *Dànshì zhù yuánwài tā yǒu yígè nǚér.*

dànshì zhù yuánwài tā yǒu yígè nǚér
but Zhu landlord 3SG have one CL

'But Landlord Zhu, he had a daughter.' [mc_mandarin_lzh_0010]

(11) *Dànshì mùlán ne mùlán zhè gè shíhòu zuò le yī gè hěn yǒnggǎn de juédìng.*

dànshì mùlán ne mùlán zhè gè shíhòu zuò le yī gè hěn
but Mulan MP Mulan this CL moment make ASP one CL very

yǒnggǎn de juédìng
courageous MOD decision

'But Mulan, Mulan at this time made a very brave decision.'
[mc_mandarin_hml_0047]

Lasty, it should be noted that Mandarin has relatively flexible word class assignment; this is particularly relevant for verbs doing double duty as adpositions (Sun 2006: 26), which in the context of serial verb constructions has implications for the identification of post-predicate argument as either objects or oblique arguments; for instance, in (12), *gěi* 'give' and *tì* 'take place of' functions as an adposition in *gěi nǐmén* 'to you' and *tì fù* 'instead of (her)

father'. Refer to the Multi-CAST annotation notes for this corpus (Vollmer 2020) for more details on their treatment in the annotations.

(12) *Wǒ jiù gěi nǐmén jiǎng yī gè huāmùlán tì fù cóng jūn zhè gè gùshì*
wǒ jiù gěi nǐmén jiǎng yī gè huāmùlán tì fù
1SG ADV give 2PL tell one CL Hua Mulan take.place.of father
cóng jūn zhè gè gùshì
enlist army this CL story
'I will tell you the story of how Hua Mulan joined the army instead of (her) father.' [mc_mandarin_hml_0011]

3.2.3.4 | Nafsan

Nafsan (sout2856), also known as South Efate, is a Southern Oceanic language spoken in the island of Efate in central Vanuatu. As of 2005, there are approximately 6000 speakers of Nafsan living in coastal villages on the island. A description of the language can be found in Thieberger (2006).

The Multi-CAST Nafsan corpus (Thieberger & Brickell 2019) is a subset of a larger data set collected by Nick Thieberger between 1995 and 2000. The entirety of the data has been archived in PARADISEC (Thieberger 1995); see Thieberger (2004) for a description of the documentation practices applied to the larger Nafsan corpus.

3.2.3.5 | Northern Kurdish

Northern Kurdish (Indo-European, Iranian; nort2641), also known as Kurmanjî, is an Iranian language spoken in eastern Turkey, northeastern Syria, northern Iraq, and western Iran, with a large diaspora into larger cities in Turkey and into Europe (Haig 2018a: 106). Reliable figures for the number of speakers are hard to come by, but in Turkey range somewhere between 8 and 15 million (Haig 2018a: 106). For a description of the language and its areal context, see Öpengin & Haig (2014), Haig & Öpengin (2018), and Haig (2018a). The texts in the Multi-CAST Northern Kurdish corpus (Haig et al. 2019) are from speakers of the Northern Kurmanji of Erzurum and Muş in eastern Turkey (nort3328). Northern Kurdish is notable for its tense-based alignment split, being ergative in the past tense, but accusative otherwise.

3.2.3.6 | Sanzhi Dargwa

Sanzhi Dargwa (Nakh-Daghestanian, Dargin; sanz1248) is a variety of Southwestern Dargwa, a Nakh-Daghestanian language spoken in the Caucasus Mountains, Republic of Daghestan, Russia. Starting in 1968, all speakers of Sanzhi resettled from their eponymous village in the mountains to multilingual and multiethnic communities in the lowlands. Today Sanzhi Dargwa is spoken by approximately 250 speakers and critically endangered, as the younger generation of speakers has shifted towards using Standard Dargwa and other varieties of Dargwa for daily use (Forker 2020: 1–4).

A large corpus of Sanzhi texts has been collected by Diana Forker and other researchers as part of a DOBES language documentation project that ran from 2012 to 2019, and has been archived at *The Language Archive* (Forker et al. 2019). Forker (2020) is a comprehensive grammar of the language, compiled on the basis of these materials. The eight texts in the Multi-CAST Sanzhi Dargwa corpus (Forker & Schiborr 2019), a mixture of autobiographical and traditional narratives, constitute only a small subset of these data.

Sanzhi is one of two languages with consistently ergative-absolutive morphosyntactic alignment in the sample, the other being Tabasaran (Section 3.2.3.7), in addition to the split-alignment system in Northern Kurdish. In Sanzhi, the Patient-like argument of a transitive verb is coded the same way as the single argument of an intransitive verb, distinct from the marking of the Agent-like argument of a transitive verb (Forker 2020; Dixon 1994). The latter is marked with ergative case, while the former two are in the unmarked absolutive case:

(13) *Il k:alk:ira kabičaqib cab hel ażdahal.*

il	*k:alk:i*	*=ra*	*ka-b-ič-aq-ib*	*ca-b*	*hel*	*ażdaha-l*
that	tree	=and	down-N-occur.PFV-CAUS-PRET	be-N	that	monster-ERG

‘And the monster made the tree fall.’ [mc_sanzhi_dragon_0030]

(14) *Il admi sarituqun caw.*

il	*admi*	*sa-r-ituq-un*	*ca-w*
that	person	ANTE-ABL-CROSS.PFV-PRET	be-M

‘The man went by.’ [mc_sanzhi_mill_0011]

There are certain circumstances under which case assignment deviates from this basic pattern, such as experiencer predicates. The definitions of grammatical roles employed in this study rely on cross-linguistically applicable generalizations; the ergative alignment of Sanzhi and Tabasaran constitutes a special case, which will be discussed alongside the general case in Section 4.3 further below.

Note also that Sanzhi Dargwa does not have a separate paradigm of third person pronouns, instead using an extensive set of demonstrative pronouns that express proximity, elevation, and cardinal direction in addition to case and number (see Forker 2020: 89–103). The following is an example of a simple demonstrative used for third-person reference:

(15) *C'il ilt:ira razi biχub cab.*

c'il	*ilt:i*	*=ra*	*razi*	*b-iχ-ub*	*ca-b*
then	those	=and	happy	HPL-be.PFV-PRET	be-HPL

'And then they were happy.' [mc_sanzhi_patima_0009]

3.2.3.7 | Tabasaran

Tabasaran (Nakh-Daghestanian, Lezgic, Samur; taba1259) is spoken in the Caucasus, in the Republic of Daghestan, Russia. Recent census data puts the number of speakers at about 120 000 speakers; Campbell et al. (2010) classify the language as vulnerable.

The texts in the Multi-CAST Tabasaran corpus (Bogomolova & Ganenkov & Schiborr 2021) were collected by Natalia Bogomolova with the assistance of Dmitry Ganenkov in 2010, and subsequently transcribed, glossed, and translated by Bogomolova. The annotations with GRAID and RefIND as well as preparations for inclusion of the data into Multi-CAST were undertaken between 2019 and 2020. The corpus is composed of four traditional and one biographical narrative.

Tabasaran is the second ergative language in the sample (Bogomolova 2021), alongside related Sanzhi Dargwa (Section 3.2.3.6), and the same considerations concerning the identification of grammatical roles apply to both; refer to Section 4.3 for a general outline of the issues involved. Although

there are notable differences between the two language, from a methodological angle, they are largely inconsequential for the present study.

(16) *Hamu haˤjvni qana raˤbʁuru hamu nuq'.*

ha-mu	*haˤjvni*	*qana*	*raˤ<b>ʁ-uru*
EMPH-PROX(ATTR)	horse(ERG)	again	<NSG>grind-FUT

ha-mu	*nuq'*
EMPH-PROX(ATTR)	grave(ABS)

'The horse again destroyed the grave.' [mc_tabasaran_horse_0027]

(17) *Hamcːi aldabʁu nuq'ǯin ʁʷanǯilan χab.*

ha-m-cːi	*alda<b>ʁ-u*	*nuq'.ǯi-n*	*ʁʷan.ǯi-l-an*	*χab*
EMPH-PROX-ADV	<NSG>rise-FUT	grave-GEN	stone-SUPER-ELAT	edge(ABS)

'Like this the edge of the gravestone rose up.'
[mc_tabasaran_horse_0015]

Like Sanzhi, Tabasaran does not have personal pronouns distinct from demonstratives in the third person, as exemplified by (18).

(18) Tabasaran
Ap'uru muvu čaz sab uǯub χalla.

ap'-uru	*muvu*	*ča-z*	*sa-b*	*uǯu-b*
do-FUT	PROX(ERG)	REFL(SG)-DAT	one-NSG	good-NSG

χal	*=la*
house(ABS)	=ADD

'He built a good house for himself.' [mc_tabasaran_nuradin_0015]

3.2.3.8 | Teop

Teop (teop1238) is an Oceanic language from the Nehan-North Bougainville network of the North-West Solomonic group of the Meso-Melanesian cluster (Ross 1988) spoken on Bougainville island, Papua New Guinea. Accurate figures for the number of speakers are difficult to ascertain; figures from the early 2000s range from 5000 to 10000 active speakers (Mosel & Thiesen 2007: 1). From 1989 onward, the island of Bougainville was torn by a decade-long civil war that resulted in an estimated 18000 to 20000 casualties, with a devastating effect on the speech population. Compounding factors such as marriages outside of the Teop speech community, the pressure exerted by neighbouring languages, and the growing influence of Tok Pisin and English all contribute to making Teop highly endangered.

A significant body of language data was collected under the direction of Ulrike Mosel between 2000 and 2007 as part of a DOBES language documentation project;[4] Mosel & Thiesen (2007) is a sketch grammar of the language, and Mosel (2019) an online dictionary. The four texts in the Multi-CAST Teop corpus (Mosel & Schnell 2015) are all traditional narratives.

3.2.3.9 | Tulil

Tulil (Papuan, Taulil-Butam; taul1251), also known by the exonym Taulil, is a Papuan language spoken in the East New Britain Province of Papua New Guinea. In 2000, the last year for which census data is available, Tulil was spoken by approximately 2000 people (Meng 2018: 1).

The six texts in the Multi-CAST Tulil corpus (Meng 2019) were recorded and transcribed by Chenxi Meng in 2012 and 2015. They constitute a small subset of a larger collection of material, all of which has been deposited into PARADISEC (Meng 2014). A comprehensive grammar of Tulil (Meng 2018) has been written on the basis of these data. The sample used for this study comprises a mix of traditional and autobiographical narratives.

4 https://dobes.mpi.nl/projects/teop/

3.2.3.10 | Vera'a

Vera'a (Austronesian, Oceanic; vera1241) is an Oceanic language spoken by about 450 speakers on Vanua Lava, Vanuatu, of which about half live in the eponymous village of Vera'a and the rest spread out along the coastline (Schnell 2011: 1).

Most of the data was collected by Stefan Schnell starting in 2007, with several additional narratives recorded by Makson Vores in 2012 and 2013. A sketch grammar is available in Schnell (2011). The Multi-CAST Vera'a corpus (Schnell 2015) is one of the largest in the sample, and consists entirely of traditional narratives.

Most zero subjects in Vera'a occur in clause-chaining contexts (i.e. where the antecedent was the subject of the immediately preceding clause; Schnell 2011; Schnell & Barth 2020), not unlike English (cf. also Schnell & Barth 2018). Furthermore, among the corpora used for this study, the Vera'a data are notable for having a relatively high proportion of intransitive clauses ($P = 0.70$ in Vera'a vs. mean $P = 0.61$ in other corpora, $\sigma = 0.061$), and like Mandarin, having a relatively low proportion of subordinated clauses ($P = 0.11$ vs. mean $P = 0.33$, $\sigma = 0.162$).

3.3 | Corpus annotations

Corpus-based typology is based on the statistical analysis of multiple usage samples from different languages. This necessitates the identification of whatever quantifiable units are the focus of research in a uniform manner across different languages, and a definition of these items in a way that is both conceptually and practically comparable (see discussion in Haig et al. 2021). Finding and defining the right categories to compare is a difficult task – see the comments on the "comparability problem" in typology in Evans (2020). One way of approaching this task is at the annotational stage, by finding appropriate generalizations across marking and constructional properties.

The analyses in this study rely primarily on annotations of the formal and functional properties of referring expressions (Section 3.3.1) and of coreference relations between referring expressions (Section 3.3.2). Additionally, specific analyses make use of annotations of the ontological animacy class membership of discourse referents (Section 3.3.4.1), mereological rela-

tions between referents (Section 3.3.4.2), as well as the information status of newly introduced referents (Section 3.3.4.3). Each of these layers of annotation targets fairly low-level aspects of linguistic and semantic structure; in combination, they enable the identification of highly complex relationships in the data, which in turn make a wide range of quantitative inquiries into the structure of discourse and referential choices possible. How exactly the annotations are leveraged for the various factors tested in this study is explained in the respective sections that discuss them in Chapter 6. From this point on, examples drawn from the corpus data also include the relevant annotation values in addition to the standard morphological glossing, to aid in the clarification of some of the analytical decisions made in this study. Most importantly, morpheme-by-morpheme glossing is by itself not cross-linguistically comparable, a shortcoming that the annotations were specifically designed to alleviate. In addition, the annotation also capture certain categories (such as zero anaphora and the distinction between subjects of intransitive and transitive clauses) that are not inferable from standard morphological glossing (or part-of-speech tagging etc.).

It should be noted that these annotations have (necessarily) all been done by hand. Manual annotation is particularly labour-intensive, which puts a practical limit on corpus size, as already mentioned earlier. Furthermore, unlike dependency-based treebank approaches (e.g. Levshina 2019), for instance, which are parsed automatically, Multi-CAST contains an indelible interpretive component in its annotations that introduces a low certain degree of random variability into the data. But in turn, the Multi-CAST collection offers a rare perspective on spoken data and underrepresented languages, for which no parsing algorithms can be feasibly trained.

There are a number of approaches to corpus annotation that share similarities with that of Multi-CAST, most of which seem to centre around part-of-speech tagging and treebanks, and have chiefly been designed for application to monolingual and written data sets. Some of these were already mentioned in Section 3.1.2 above in the context of the various corpora they have been applied to: Iida et al. (2007) describe an annotation scheme for predicate-argument relations and co-reference relations, designed for a corpus of written Japanese; Nicolae et al. (2010) extend data from the Discourse Graphbank (Wolf et al. 2005) with custom annotations for co-reference, and also note ontological classes of referents, referential status, and form of expression, but not grammatical relations; Loukachevitch et al. (2011), Kibrik et al. (2013),

and Kibrik et al. (2016) likewise utilize data from existing treebanks, in their case the RST Discourse Treebank (Carlson et al. 2002). Their annotations mark co-reference, including split antecedence, and a nuanced sets of expression types and discourse factors (see Section 2.3.1 above). More superficially related systems of co-reference annotation that have the same aim, but different motivations, include, among others, the Entity Detection and Tracking system (EDT, Doddington et al. 2004), the Potsdam Coreference Scheme (PoCoS, Krasavina & Chiarcos 2007), and the scheme described in McEnery et al. (1997).

The following sections outline the various annotation schemes applied to the corpus data used for this study. Each section starts with the motivations and general operational principles of the annotations, followed by a brief rundown of their formal and functional properties. The primary purpose of these sections is to illustrate one approach to building annotation systems for corpus-based typological research. In doing so, they aim to provide a foundational understanding of how these annotations work on a basic level, and so make it easier to follow along with the analyses presented in this study as well as allow for better replication of the results. Beyond that, these sections are focused mainly on those aspects of the annotations that are directly relevant to the analyses; exhaustive descriptions are available in the respective manuals and guidelines cited below.

3.3.1 | Referring expressions and grammatical relations

The GRAID annotation scheme (Grammatical Relations and Animacy in Discourse, Haig & Schnell 2014) was designed for the corpus-based investigation of the intersection between discourse and grammar, a field pioneered by the work of Chafe (1976, 1980, 1987, 1994, etc.) and Givón (1978, 1979, 1983a, etc.), among others. GRAID was designed with typological comparability in mind, attempting to generalize across universally shared morphosyntactic categories while leaving open avenues for language-specific adaptations. The scheme targets the expression of participants in states of affairs as they occur in connected discourse features, marking

- the basic form of referring expressions,
- person and humanness distinctions,

- the syntactic relations of major clause constituents, and
- the boundaries of phrase-level and clause-level syntactic units.

Given a long enough stretch of discourse, the GRAID system enables the analysis of the associations between these layers of annotation, in addition to information gleaned from their relative positioning. Do note, however, that GRAID annotations are focused on the arguments of verbs, meaning that not all form classes across attested across languages are captured with the same degree of detail.

In terms of formal categories, GRAID notes whether clause constituents are realized as full noun phrases or pronouns, or are left unexpressed. Consider the following example from the English corpus:

(19)

she	*used*	*to*	*stand*	*there*	*and*		*sell*	*the*	*whelk-s*
pro						0		ln	np

[mc_english_kent02_0159]

There are three constituents of interest in this example: a pronominal NP *she* (⟨pro⟩), a zero expression in a same-subject construction (⟨0⟩),[5] and a full NP *the whelks* (⟨np⟩). The annotation symbols align with individual word units (or nothing, in the case of zero), but in fact target entire phrases. In multiword noun phrases, the symbols ⟨ln⟩ and its counterpart ⟨rn⟩ are assigned to all subconstituent elements, respectively to those to the left and right of the phrasal head, as with the definite determine *the* in (19). These glosses identify the phrase membership of all pertinent elements in a clause and so delineate the extent of phrases, which allows checking for the presence of specific modifiers (e.g. demonstrative determiners or possessors), compare the length and complexity of NPs, and so on.

GRAID defines formal categories on a fairly general level, relying on distinctions that can be assumed to be identifiable across a diverse range of languages. For instance, the form gloss ⟨pro⟩ is applied to various types of free and bound person indices including personal pronouns, but also to demonstratives where they are used pronominally as anaphoric expressions, for instance in languages that do not possess a distinct paradigm of third-person pronouns and instead use demonstratives or other kinds of pro-forms, as in

5 Which in English also omits any associated auxiliaries.

Sanzhi Dargwa and Tabasaran (Sections 3.2.3.6 and 3.2.3.7; see also Bhat 2004). Pro-forms that are used in lieu of definite pronouns (Lyons 1968) receive an additional specifier denoting their status, for instance ⟨dem_pro⟩ for a demonstrative. This procedure allows for cross-linguistic comparison with regards to a comparative category of pronouns ⟨pro⟩ – and analogously for other categories – as well as for consideration of different form types encompassed by this broader category.

Crucially, GRAID also captures zero anaphors. This practice is hardly uncontentious; it relies on a conception of zero as a "form" type, even though zero anaphors are in essence the absence of form and referential only by inference. Note that GRAID only captures unexpressed arguments on the clause level, which are fundamentally different from zero in morphological paradigms. The annotations further distinguish pragmatically selected zero forms such as the one in (19) from structural and other syntactially conditioned cases of argument deletion, such as in non-finite clauses. As defined here, pragmatic zero anaphors are those that are licensed by the predicate in question, generally imply a specific referent retrievable from the discourse context, and are not systematically suppressed by the argument function they are placed in. The last criterion is particularly crucial in the context of referential choice, as there is no alternation between zero and other forms in these positions since speakers do not exercise a choice of expression.

In addition to the formal properties of referring expressions, GRAID also marks certain person and animacy classes. (20) shows how the various levels of symbols combine linearly into GRAID "words":

(20)

she	*used*	*to*	*stand*	*there*	*and*		*sell*	*the*	*whelk-s*
pro.h						0.h		ln	np

GRAID distinguishes first (⟨.1⟩) and second (⟨.2⟩) person from third person forms, which further distinguish human ⟨.h⟩ from non-human referents; the latter are not explicitly marked. Humanness is assumed to be entailed in the first and second person, and hence the annotations conflate both categories. A fourth animacy category (⟨.d⟩) is used for anthropomorphized third-person referents such as talking animals and spirits, whose appearence is not uncommon in folktales.

On the functional side, GRAID is focused primarily on syntactic arguments: subjects, objects, and various oblique arguments such as goals, instru-

mentals, and benefactives, with the first two being the most relevant for the present study. Their definition adapts the concept of a semantic transitive prototype (Andrews 2007: 135–140), in essence combining cross-linguistically identifiable semantic prototype features such as proto-agent and proto-patient with the identity of argument encoding and the number of arguments (see Haig & Schnell 2014: 12–14). In this sense, verbs such as *hit* and *make* describe transitive events with a prototypical agent and patient, which in English are respectively encoded as a pre-verbal NP (triggering agreement on the verb if third person and present tense), and a post-verbal NP. In a transitive clause, any syntactic arguments *encoded* in the same manner as proto-agents or proto-patients are annotated as carrying "A" or "P" function, though this does not yield an interpretation of them *being* agents or patients. Any clause lacking either an A or P argument or both is classified as intransitive, and may contain a single core argument function "S", in addition to any oblique functions; S is also the function of the subjects of non-verbal clauses. For the purposes of this study, the A and S functions are then collapsed into a combined category of "subject". While this is not an uncontentious decision, it should not be understood as claiming that the two roles pattern the same way or that the various languages in the sample (some of which have ergative morphology, as noted earlier) posses a comparable notion of subjecthood. Rather, it follows from the methodological principle that such distinctions should be made as part of the modelling process instead of as an a priori decision. The use of transitivity as a factor in this study is discussed in Chapter 4.

In the extended example (21), then, *she* is the subject and sole argument of an intransitive clause headed by the verb *stand*, and is hence glossed with the symbol ⟨:s⟩ for the S function. The second clause, headed by *sell*, is transitive, as it contains two arguments that map unto the A (⟨:a⟩) and P (⟨:p⟩) functions, the former of which is unexpressed.

(21)

she	*used to stand there and*		*sell the whelk-s*
pro.h:s		0.h:a	ln np:p

Altough *sell* is usually ditransitive, here the recipient is contextually unavailable and hence not coded. In languages where the recipient of a ditransitive clause receives the marking a the patient of a prototypical monotransitive clause, it is also coded as P. In order to distinguish the recipient from the theme of a ditransitive construction, the latter is glossed ⟨:p2⟩ in GRAID. S ar-

guments whose marking deviates from that of a regular intransitive subject in a language (e.g. dative subjects in Tabasaran) are classified as non-canonical subjects (⟨:ncs⟩). Crucially, the functional classification is always based on the surface-level coding of arguments; the semantic roles and prototyplicality of specific instances is irrelevant for this classification. Lastly, the symbol ⟨:obl⟩ is used for oblique arguments, that is any required argument that does not meet the criteria for identfication as subject or object, with two specific subtypes of oblique arguments receiving distinct glosses: ⟨:g⟩ for goals and addressees, and ⟨:l⟩ for static locations.

As with form categories above, the use of cross-linguistically stable features of semantic prototypes (albeit collapsed into a broad "subject" category) in favour of language-specific definitions of subjecthood is motivated by the needs of cross-linguistic comparability. Compared to similar approaches that employ, for instance, semantic macro-roles as in role-and-reference grammar (Van Valin & LaPolla 1997) or in Bickel (2011), the advantage of the framework defined in Andrews (2007) is the relative ease with which core argument functions can be identified and delimited across diverse languages.[6]

The properties of other clause constituents such as predicates and adjuncts are registered with comparatively less analytic depth, since they are not the focus of the annotation system:

(22)

she	*used*	*to*	*stand*	*there*
pro.h:s	lv	lv	v:pred	other:l

and		*sell*	*the*	*whelk-s*
other	0.h:a	v:pred	ln	np:p

The symbol ⟨v⟩ marks the primary verb of the clause, which here receives the function symbol ⟨:pred⟩ since it is the head of the verbal complex. The symbols ⟨lv⟩ and ⟨rv⟩ are subconstituents of the verbal complex, here used

6 Do note, however, that this system requires modification to work with languages that exhibit multiple transitive patterns that vary systematically in the way core arguments are encoded, as for instance with symmetrical voice systems of the so-called Philippine type (Foley 2008; Riesberg 2014). While none of the languages examined in this study belong to this type, the Multi-CAST collection to date contains data from two languages that do, Tondano (tond1251; Brickell 2016) and Arta (arta1239; Kimoto 2019).

for auxiliary *used to*; they works analogously to the similar symbols used for NP subconstituents described above. Lastly, the symbol ⟨other⟩ is assigned to any element for which no other form or function symbol can be applied. In this case, *there* also receives the function gloss ⟨:l⟩ for the expression of a static location. A full list of GRAID symbols can be found in Appendix A.3 and in the documentation for the Multi-CAST collection.

The final component of the GRAID annotations involves the delineation of phrasal and clause boundaries, and concomitantly with the latter, the identification of certain dependent clause types (complement, relative, adverbial, etc.), illucutionary force, and direct speech segments.

(23)

	she	*used*	*to*	*stand*	*there*
##	pro.h:s	lv	lv	v:pred	other:l

	and		*sell*	*the*	*whelk-s*
##	other	0.h:a	v:pred	ln	np:p

The ⟨##⟩ symbols indicates the left-edge boundary of a clause that is not syntactically dependent on another clause. These are for the most part fully independent clauses capable of expressing an utterance by itself, but also include constructions such as those in the example above, which meet most but not all criteria for full independence. Every other type of clause instead receives the ⟨#⟩ symbol, and the right-edge boundary of embedded clauses is additionally marked with the ⟨%⟩ symbol to better represent syntactic hierarchization. Where an embedding splits its matrix clause in half, the latter continues after the ⟨%⟩ symbol. The left-edge boundary markers may combine with additional tags, such as ⟨##neg⟩ for a negated independent clause or ⟨#cc⟩ for a complement clause; see the aforementioned Appendix A.3 for a list of all clause boundary symbols and rules for their combination.

The core GRAID symbol inventory outlined here can be extended via the use of additional specifiers, which attach to the basic symbol with an underscore ⟨_⟩; the example of ⟨dem_pro⟩ for demonstrative pronouns was already brought up above. Other specifiers relevant here include ⟨pn_np⟩ for proper names (vs. NPs headed by common nouns) and ⟨ln_dem⟩ for demonstrative determiners (as in e.g. *this book*), among others. This is also the primary way in which GRAID captures language-specific properties while maintaining cross-linguistic comparability on a more general level: Symbol specifiers

narrow down (but do not extend) the definition of the more general category they attach to, and so can readily be subsumed under them if needed, as is done for those analyses in this study that do not rely on the distinctions drawn by them.

The GRAID annotations constitute the foundation of this study, but their true potential is realized only in combination with the annotations for co-reference with the RefIND scheme, described in the next section.

3.3.2 | Referent identification

The RefIND annotation scheme (Referent Indexing in Natural-language Discourse, Schiborr et al. 2018) is a system for identifying co-reference relations and referent tracking. It does so by assigning a unique numerical identifier to each referent in a text, which is repeated for every mention of that referent:

(24)

	she	*used*	*to*	*stand*	*there*
	3SG.F	used	to	stand.INF	there
##	pro.h:s	lv_aux	lv	v:pred	other:l
	0016				0137

	and		*sell*	*the*	*whelk-s*
	and		sell.INF	the	whelk-PL
##	other	0.h:a	v:pred	ln_det	np:p
		0016			0133

[mc_english_kent02_0159]

As seen here, the RefIND indices align with the heads of referring expressions, similar to the GRAID annotations. This procedure allows the various realizations of individual referents to be followed throughout a stretch of connected discourse, which in combination with the GRAID annotations makes it possible to determine whether a specific referent or a type of referents behaves differently from others, for instance with regard to animacy, protagonisthood, or contextual properties such as anaphoric distance. This information can then be leveraged to answer questions related to referent accessibility (Ariel 1990) or activation (Chafe 1976; 1994) as conditioned by the local discourse context; consider also the related notion of lookback in Givón (1983a). Beyond

the description of individual referents, RefIND allows capture of the global properties of texts as well as the local properties of subsegments of texts, for instance in terms of referential density (Noonan 2003; Bickel 2003; Stoll & Bickel 2009) and referent pressure (Du Bois 2003a, 2003b; Durie 2003), or analogously in terms of a local competition between concurrently active referents (see Section 6.6), all of which have been claimed to have influence on certain distributional patterns of referring expressions.

The notion of referent identification rests on the assumption that speakers assign a continuous identity to a referent in their mental representation of it (cf. Du Bois 1980: 208), meaning that a connection between this singular abstracted identity and the various linguistic forms occuring in a text can be drawn; see also the discussion of co-referentiality and co-reference annotations in van Deemter & Kibble (1999).

While the annotation of co-reference relations is at first glance a comparatively simple process, there are a number of conceptual and analytical issues that make it less straightforward in practice. The first and most contentious step is to determine which expressions are referential in the first place, which is discussed in detail in Section 4.1.2 further below, and then to decide which previously mentioned referential entity they represent and should share an index with, or else if they should constitute a new referent to be assigned its own identifier. In example (24) the subjects of both constituent clauses share the same index ⟨0016⟩ since they point to the same referent (the speaker's mother), while the location expressed by *there* (⟨0137⟩) and *the whelks* (⟨133⟩) each receive their own, distinct index. Further details on annotational issues are discussed in the RefIND annotation guidelines (Schiborr et al. 2018), and briefly touched on where relevant in the following chapters.

In addition to the referent indexing, the RefIND scheme also track a number of further referent properties which are outlined in the next sections: These include animacy class (Section 3.3.4.1), relationships between related referents (Section 3.3.4.2), and the inferability of newly introduced referents (Section 3.3.4.3). A number of longer, fully annotated examples can be found in Section 3.3.3.

3.3.3 | Longer annotated examples

The following are a number of longer examples with complete GRAID and RefIND annotations, showcasing the application of the two systems in unison. Further examples can be found spread throughout this study.

The first example (25) is from the English corpus. Notable here in particular is the continuity of reference in the subject role (annotated ⟨pro.h:a⟩ and ⟨pro.h:s⟩), which is easily identified via the referent indices (here specifically index ⟨0090⟩).

(25) English

a. *He put on his jacket.*

	he	*put*	*on*	*his*	*jacket*
	3SG.M	put.PST	on	3SG.M.POSS	jacket
##	pro.h:a	v:pred	rv	ln_pro.h:poss	np:p
	0090			0090	0108

'He put on his jacket.'

b. *Where's my jacket, Mother, he said.*

	where	*='s*	*my*	*jacket*	*Mother*
	where	=be.PRS.3SG	1SG.POSS	jacket	Mother
##ds	other:pred	=cop	ln_pro.1:poss	np:s	pn_np.h:voc
			0090	0108	0091

	he	*said*
	3SG.M	say.PST
##	pro.h:s_ds	v:pred
	0090	

'Where's my jacket, Mother, he said.'

c. *And he put on his jacket.*

	and	*he*	*put*	*on*	*his*	*jacket*
	and	3SG.M	put.PRS	on	3SG.M.POSS	jacket
##	other	pro.h:a	v:pred	rv	ln_pro.h:poss	np:p
		0090			0090	0108

'And he put on his jacket,'

d. *He went down there.*

	he	*went*	*down*	*there*
	3SG.M	go.PST	down	there
##	pro.h:s	v:pred	adp	other:g
	0090			0102

'He went down there.'

e. *He come back with them boots.*

	he	*come*	*back*	*with*	*them*	*boot-s*
	3SG.M	come.PST	back	with	DIST.PL	boot-PL
##	pro.h:s	v:pred	rv	adp	ln_dem	np:obl
	0090					0107

'He come back with them boots.'

[mc_english_kent03_0102-0103]

This example also contains a number of adnominal expressions of possession (e.g. ⟨ln_pro.h:poss⟩ 'his') in conjunction with the 'jacket' (index ⟨0108⟩). These and other adnominal modifies can be readily associated with the head of a phrase (which carries the symbol for grammatical role as well as referent indices), and in this way leveraged for assessing the relative complexity of nominal expressions. (25b) furthermore demonstrates the treatment of direct speech and the constructions that frame it (i.e. *he said*) as distinct structures rather than syntactic embeddings.

In this example from the Vera'a corpus, the subject shifts back and forth between the ten brothers (⟨0002⟩) and their eleventh brother *Qo'* (⟨0001⟩). References to the latter demonstrate the use of GRAID symbols with additional specifiers for subcategorization, here ⟨pn_np⟩ for proper names. Note here in particular the goal argument ⟨np:g⟩ to 'go' in (26a), the use of the negation tag on the clause boundary marker ⟨##neg⟩ in (26b), and in (26c) the way Vera'a employs possessive suffixes (⟨-rn_pro.h:poss⟩), which, despite being structurally quite different from the possessive expressions in the English example above, are immediately identifiable as such due to sharing the same annotation categories.

(26) Vera'a

a. *Dirm van kal sar lēn wōm̃ōm̃ō'.*

	dir	*=m*	*van*	*kal*	*sar*	*lē*	*=n*	*wōm̃ōm̃ō'*
	3PL	=TAM1	go	upwards	inland	LOC	=ART	bush
##	pro.h:s	=lv	v:pred	rv	rv	adp	=ln	np:g
	0002							0008

'They went up into the bush.'

b. *Ba e Qo' ē van ros.*

	ba	*e*	*Qo'*	*e*	*van*	*rōs*
	but	ART	Qo'	GEN.NEG1	go	GEN.NEG2
##neg	other	ln	pn_np.h:s	lv	v:pred	rv
			0001			

'But Qo' didn't go.'

c. *'Eraga 'i'isigi wuva dirm van.*

	e	*raga*	*'i-'isi*	*-gi*	*wuva*
	ART	people	RED-brother	-3SG	only
##	ln	ln	np.h:dt_s	-rn_pro.h:poss	rn
			0002	0001	

dir	*=m*	*van*
3PL	=TAM1	go
pro.h:s	=lv	v:pred
0002		

'Only his brothers, they went.'

d. *Dir sañwul.*

	dir	*sañwul*
	3PL	ten
##	pro.h:s	np:pred
	0002	

'They were ten (brothers).'

e. *Dirk kal ’ar ēn naka.*

	dir	*=k*	*kal*	*’ar*	*ēn*	*naka*
	3PL	=TAM2	go.upwards	cut	ART	canoe
##	pro.h:a	=lv	v:pred	rv	ln	np:p
	0002					

‘They went upwards to cut canoes.’

f. *Qo’ ga ’og’og lēn lōlō vunuō, ne ’og ’i.*

	ba	*Qo’*	*ga*	*’ōg’ōg*	*lē*	*=n*	*lōlō*	*vunuō*
	but	Qo’	STAT	RED:stay	LOC	=ART	inside	village
##	other	pn_np.h:s	lv	v:pred	adp	=ln	ln	np:l
		0001						0004

	ne	*’ōg*	*’i*
	TAM2:3SG	stay	DEL
##	0.h:s	lv-pro_h_s	v:pred	rv
	0001			

‘But Qo’ stayed in the village, (he) stayed behind.’

[mc_veraa_jjq_0007-0012]

The third and final example is taken from the Tabasaran corpus. Here, the complement clause (⟨#cc⟩) in the first part (27a) does not allow for overt subject realization, and so its omission is marked with the symbol ⟨f0⟩ instead of ⟨0⟩. The distinction between pragmatically selected (⟨0⟩) and structurally enforced (⟨f0⟩) subject ellipsis is a crucial one, and one we will discuss in more detail in Section 4.1.3 below.

(27) Tabasaran

a. *Abguz kːun šulu murariz.*

```
                        a<b>gu-z
                        <NSG>search-INF
## #cc f0.h:a 0:p       vother:pred    %
       0004   0083
```

```
kːun        šul-u          mu-rari-z
want        become-FUT     PROX-PL-DAT
other:lvc   v:pred         dem_pro.h:ncs
                           0004
```

‘They wanted to search (for a gazelle).’

b. *Aˁʁuˁra sab jarχʷla, hotmu mučʼu jarkʷrariz.*

```
           aˁʁ-uˁra    sa-b      jarχʷ-l-a
           (PL)go-PRS  one-NSG   long-SUPER-ELAT
## 0.h:s   v:pred      other     other
   0004
```

```
ho-tmu              mučʼu   jarkʷr-ari-z
EMPH-DIST(ATTR)     dark    forest-PL-DAT
ln_dem              ln      np:g
                            0084
```

‘(They) went far away, to those dark forests.’

c. *Hacːib jišariz aˁʁuˁru murar, šubred čjirra.*

```
   ha-cːi-b                 jiš-ari-z     aˁʁ-uˁru     mu-rar
   EMPH-(DIST)ADV-NSG       place-PL-DAT  (PL)go-FUT   PROX-PL(ABS)
## ln                       np:g          v:pred       dem_pro.h:s
                            0084                       0004
```

```
šub-r-ed        čʼ-jir              =ra
three-NSG-DEF   brother-PL(ABS)     =ADD
ln_num          np.h:appos          =other
                0004
```

‘They went to that place, these three brothers.’

d. *Qa aˤχuˤ čuč:u jivnu dišlaji dubk'u, sab žejran χuru.*

qa	*aˤχuˤ*	*čuč:u*		*jiv-nu*	*dišlaji*
then	big	brother(ERG)		hit-PCVB	immediately
## other #cv	ln	np.h:a	0:p	lv_v	other
		0011	0085		

du-b-k'-u	*sa-b*	*žejran*	*χ-uru*
PFV-NSG-kill-PCVB	one-NSG	gazelle(ABS)	carry-FUT
v:pred %	0.h:a ln_num	np:p	v:pred
	0011	0085	

'The oldest brother killed a gazelle right away and brought (it) back.'

e. *ʁabχiri muʁu čan χp:iriz k'ur, ap'in ip'rub k'ur.*

	ʁ-a<b>χ-iri
	PFV-<NSG>bring-PCVB
## #cv 0.h:a 0:p	v:pred %
0011 0085	

muʁu	*ča-n*	*χp:iri-z*	*k'ur*
PROX(ERG)	REFL.SG-GEN	wife-DAT	CIT
dem_pro.h:a_ds	ln_refl.h:poss	np.h:g	other
0011		0053	

	ap'-in	*ip'-ru-b*	*k'ur*
	do-IMP	<NSG>eat-FUT.PTCP-NSG	CIT
##ds 0.2:a	v:pred	np:p	other
0053		0086	

'Having brought back (the gazelle), he said to his wife, Prepare food [from it].'

3.3.4 | Additional annotations and data

In addition to the annotations with the GRAID and RefIND schemes, this study makes use of a number of further levels of annotation applied to the corpus data as well as external data: annotations of ontological animacy classes, which extend the animacy labels in GRAID, annotations of mereological links between related referents, and annotations of the information status of newly introduced referents. With the exception of the links between related referents, these data are not used for the primary investigation of referential choice, but are leveraged only on the side, for excursions into specific interactions between other factors.

3.3.4.1 | Ontological animacy classes

In addition to assigning each referent a unique identifier, the RefIND annotations also note its membership in a number of broadly defined animacy classes (Schiborr et al. 2018: 15). These classes and their associated codes in the annotations are, in particular:

- human animates including anthropomorphized entities (⟨hum⟩), and
- non-human animates (⟨anm⟩);
- body parts (⟨bdp⟩),
- non-individuable masses (⟨mss⟩), and
- all other inanimate objects (⟨inm⟩);
- locations (⟨loc⟩),
- points in or intervals of time (⟨tme⟩), and
- all other abstract entities and notions (⟨abs⟩).

The human class overlaps with the GRAID symbols ⟨.h⟩ and ⟨.d⟩ for third-person references. Human and inanimate referents constitute the by far most common classes. The other classes individually are comparatively rare, and the occurrence of some, especially non-human animates, is highly dependent on the specific content of a text; this issue discussed further in Section 6.1 below.

3.3.4.2 | Mereological relations between referents

As noted above, the definition of referents in terms of the RefIND scheme is based on their individuation as uniquely identifiable entities, and as such does not inherently recognize where these entities overlap in conceptual space. To remedy this, the RefIND annotations further note certain mereological relationships between referents (Schiborr et al. 2018: 16):

- split antecedence (⟨>⟩) as in (29),
- partial co-reference (⟨<⟩) as in (28), and
- meronymic part-whole relations (⟨M⟩) as in (30).

These relations are noted for all relevant referent indices on a text-by-text basis using the symbols given above. A referent can have multiple relations of different types with other referents.

(28) English

a. *And my father and him couldn't get on at all.*

	and	*my*	*father*	*and*	*him*	*could-n't*
	and	1SG.POSS	father	and	3SG.M.OBL	could-NEG
##neg	other	ln_pro.1:poss	np.h:s	rn	rn_pro.h	lv_aux
		0000	0027		0108	

get	*on*	*at*	*all*
get.INF	on	at	all
v:pred	rv	other	other

'And my father and him couldn't get on at all.'

b. *They was always flying at one another.*

	they	*was*	*always*	*fly-ing*	*at*	*one*	*another*
	3PL	be.PST.3SG	always	fly-PTCP.PRS	at	one	another
##	pro.h:s	lv_aux	other	v:pred	adp	ln	refl.h:g
	0111						0111

'They was always flying at one another.'

[mc_english_kent02_0134]

(29) Vera'a

a. *Vus seqeg diē lēn mōgmōglēge dōl: …*

	vus	*seqeg*	*diē*	*lē*	*=n*	*mōgmōglēge*	*dōl*
	hit	bless	3SG	LOC	=ART	things	all
##	0.h:a	v:pred	rv	pro.h:p	adp	=ln	np:obl	rn
	0010			0001			0011	

'[He] blessed him with all (the customary) things: …'

b. *Dim gis qal gōr ēn mañra, …*

	di	*=m*	*gis*	*qal*	*gōr*	*ēn*	*mañra*
	3SG	=TAM1	hold	hit	secure	ART	money
##	pro.h:a	=lv	v:pred	rv	rv	ln	np:p
	0010						0012

'He bore money, …'

c. *dim gis qal gōr ēn qō.*

	di	*=m*	*gis*	*qal*	*gōr*	*ēn*	*qō.*
	3SG	=TAM1	hold	hit	secure	ART	pig
##	pro.h:a	=lv	v:pred	rv	rv	ln	np:p
	0010						0013

'he bore pigs.' [mc_veraa_iswm_0024-0026]

(30) Northern Kurdish

a. *Bal-a xwe didê xortekî.*

	bal-a	*xwe*	*di-d*
	attention-EZ	REFL	IND-give.PRS.3SG
##	0.h:s_cp	np:lvc	rn_refl.h:poss	v:pred
	0033		0033	

=ê	*xort-ek-î*
=3SG.OBL	youth-INDF-EZ
=pro:other	np.h:obl
	0046

'[He] sees a young man…'

b. *Digre zêrekî dike kefa destê wî.*

	di-gir-e	*zêr-ek-î*	*di-k-e*	*kef-a*
	IND-put.PRS-3SG	gold-INDF-OBL	IND-do.PRS-3SG	palm-EZ
##	0.h:a	lv	np:p	v:pred	np:g
	0033		0050		0126

dest-ê	*wî*
hand-EZ	3SG.OBL.M
rn_np:poss	rn_pro.h:poss
0048	0046

'[He] places a piece of gold in the palm of his hand.'

[mc_nkurd_muserz03_0084-0085]

Bridging relations are annotated only in chronological order of referent introduction, that is, no relations are noted with referents introduced later in the discourse, as they arguably do not yet "exist" in the universe of the discourse prior to their introduction. However, since partial co-reference and split antecedence are converse relations, one is necessarily entailed by the other, and hence any implicit relations can be established programmatically. The same is true for meronymic relations and their inverse holonyms, but the latter are only implied by the former in the annotations, not explicitly marked.

3.3.4.3 | Information status of new referents

Although this is primarily a study of anaphoric mentions and not of referent introductions, the contextual inferability of new referents at point of introduction might affect form of second-and-subsequent mentions; this idea is expounded on further in Section 6.9; see also Prince (1981b) (and Section 2.2.2 above) on the interrelation of new and given information.

The ISNRef scheme (Information Status of New Referents, Schiborr et al. 2018: 15) notes the information status of newly introduced referents (i.e. whenever a distinct referent index is assigned) with a simply taxonomy distinguishing

- brand new referents (⟨new⟩) from
- referents inferable from frame semantics (⟨bridging⟩) and
- globally known entities (⟨unused⟩).

ISNRef is in essence a greatly simplified version of the RefLex scheme (Riester & Baumann 2017).

The operative distinction is between referents that are in some way contextually inferable from frame semantics (Fillmore 1982) and those that are not. Referents may be inferable from the linguistic context via mereological relations (see Section 3.3.4.2) or from frame semantics (e.g. mention of a room implies a floor, ceiling, walls, etc.; cf. Hawkins 1978: Ch. 3 and Lambrecht 1994: 91), or else from the deictic context or shared world knowledge. The latter captures any reference to known but not previously mentioned (i.e. "unused") entities – discourse-new but hearer-old in Prince's (1992) terms – such as those that are present in the immediate physical surroundings (i.e. external deictic references) or universally recognizable from shared encyclopaedic knowledge (e.g. *the sun*, *the stars*, but also *the spirits*, *the capital*), including the interlocutors themselves. Which entities are drawn from shared world knowledge is of course culturally dependent, and necessitates a certain familiarity with the speech communities from which the data are drawn. References to globally known entities are quite rare overall, accounting for only 3% of the unique referents in the sample. For the purpose of this study, all referents that are not brand new are grouped together into a single class, reducing the three-way categorization given above to a simpler, binary one.

The following examples exemplify usage of the annotation scheme:

(31) Cypriot Greek
Mian foran ishen enan vasilean.

	mian	*foran*	*ishen*	*enan*	*vasilean*
	one.F	time	have.IPFV.3SG	INDF.M.ACC	king
##	ln	np:other	other:predex	ln	np.h:s
					0001
					new

'Once there was a king.' [mc_cypgreek_jitros_0001]

Example (31) was taken from the very beginning of a narrative, and as such there are no prior mentions of *vasilean* 'king' (⟨0001⟩) in the local discourse record, making this perhaps the least ambigious kind of brand new introduction (⟨new⟩). As seen here, the information status labels align with the referent indices in the annotated texts.

In (32), *nok* ‘hand’ (⟨0010⟩) and *kiou* ‘mouth’ (⟨0011⟩) are both new to the discourse, but inferable (⟨bridging⟩) from being body parts of two previously introduced referents, the lizard (⟨0004⟩) and the conch shell (⟨0001⟩).

(32) Tulil

mənəbət, məlanga vuna vənok da vətauk da ikiou.

	mən=	*nə-bət*	*məlang*	*=a*	*v-un*	*=a*
	from=	LOC-MED	lizard	=SG.CL:MASC	3SG.M.PST-throw	=PAT
##	=adp	dem_other:dt	np.d:a	=rn	v:pred	=rv
			0004			

və=	*nok*			*da*	*və-tauk*	*da*
3SG.F.POSS.INAL=	hand			PURP	3SG.M.PST-grab	on
=ln_pro.d:poss	np:p	#	0.d:s	other	v:pred	adp
0004	0010		0004			
	bridging					

i=	*kiou*
3SG.F.POSS.INAL=	mouth
=ln_pro.d:poss	np:g
0001	0011
	bridging

‘From there, the lizard stretched out his hand to hold on to her [= the conch shell’s] mouth.’ [mc_tulil_lnsl_0026]

4 | Study design and methodology

This chapter describes how the various categories counted and factors tested in this study are defined. The first half concerns itself with the general selection criteria applied to the corpus data (Section 4.1) and the definitions of form and function categories (Sections 4.2 and 4.3), which are capped off with a brief numerical overview of the resulting sample (Section 4.4). The second half lists the various properties of the discourse and referent that feed into the multifactorial analysis of lexical choices in Chapters 6 and 7 (Section 4.5) and discusses a number of specific methodological issues related to these properties and their definitions (Section 4.6).

4.1 | Sample selection criteria

4.1.1 | General selection criteria

The following sections outline the general selection criteria used for this study, that is those guided by the specific research motivating this study rather than by methodological considerations; the latter will be discussed in the subsequent sections.

To be considered for sampling, an expression must meet four basic criteria: First, since this is a study of reference and anaphora, only expressions that

are fully referential are considered, and second, since it is further a study of referential choice, it is limited to expressions that are pragmatically selected, that is those that are part of a paradigm of choice that allows free alternation with other expressions. These two criteria are discussed in more detail in Sections 4.1.2 and 4.1.3.

Third, as the target of this investigation is specifically anaphoric references, only the second and subsequent mentions of a discourse referent are included. The first mentions of discourse referents, that is their introduction into discourse (as per Chafe 1976 and Prince 1981b; see Section 2.2.2), is subject to different constraints than those that apply to subsequent anaphoric references (cf. discussion in Section 2.2.5.1). This is reflected in the selection of expressions for referent introductions, which, as will be seen later in Section 5.3 in Chapter 5, are overwhelmingly realized as full, lexical NPs in all ten corpora in the sample. Since first mentions are excluded, the sample consequently (and obviously) encompasses only discourse referents which are mentioned at least twice in a given text. This excludes a significant proportion of the discourse referents identified by the annotations that only ever mentioned once and never picked up again in the following discourse (cross-corpus mean $P = 0.56$, $\sigma = 0.084$).

And fourth, since this study focuses its attention specifically on anaphors in argument positions – subjects, objects, and obliques – the sample includes only mentions in those positions. Precisely which arguments in which constructions are encompassed by the definition of these roles follows from the definitions outlined in Section 4.3 below. Do keep in mind that while the anaphors in question are limited to these positions, any factors that consider the preceding discourse, such as anaphoric distance or competition between candidate antecedents, involve all relevant mentions in any position, not just those that are subject, objects, or oblique arguments. This also applies to introductions and discourse referents with only a single occurrence, as mentioned above.

Lastly, there are a number of minor but nevertheless relevant considerations that need to be taken into account in addition to those listed here; these are also part of Section 4.1.4 below.

4.1.2 | Referentiality

As discussed in Section 3.3.2 above, a crucial and not uncontroversial challenge revolves round determining which referring expressions instantiate a new discourse referent, and which refer to the same referent (cf. van Deemter & Kibble 1999). The definition employed here in essence is that given in Du Bois (1980: 208), this being that an expression is deemed referential "when it is used to speak about an object as an object, with continuous identity over time". In other words, in the ideal case a referring expression evokes a specific, individuable discourse entity that can be commented on (i.e. as a topic, see Gundel 1985: 90) and can be identified as referring to the same entity across separate instances (i.e. it can be "tracked" in the terminology of RefIND, Schiborr et al. 2018).

As such, this definition excludes certain expressions that do not refer to a specific entity, of which the clearest cases are indefinite pronouns, classifiers (e.g. in Mandarin), and other forms with non-specific reference (e.g. English *you*, German *man*). Other cases of non-specific entities are less clearly delineated in terms of meeting the criterion of referentiality. Category classifiers (e.g. *elephants live in Africa, they have long trunks*) and mentions in irrealis contexts (e.g. *if I were to adopt a dog, it would be a collie*) are assumed to instantiate a trackable discourse referent only where they are taken up again in the following discourse, as is the case in these two examples (Schiborr et al. 2018: 5–6).

Where this is not the case, or where a classifier merely serves to classify another referent (e.g. with non-identificational nominal predicates, e.g. *she is an artist*), or where a reference to a class of entities does not evoke an individuated discourse referent (e.g. conflated objects, e.g. *she is wearing glasses*), the referentiality criterion is deemed to not be met. Similarly, expressions under scope of negation (e.g. *no true Scotsman*) are excluded, as are those that are part of idiomatic expressions (e.g. *break the ice*), the non-verbal elements in complex predicates (e.g. *take prisoner*), and incorporated nouns.

A related issue concerns uniqueness of reference: Where an entity is composed of multiple constituent parts (and vice versa, see Section 3.3.4.2 above), part and whole are considered separate discourse referents. However, where one entity is transformed into another, either via a progression of small changes (e.g. a material being worked into an object, such a tree being turned into a canoe) or instantly (e.g. a human being magically transfigured into some

non-human shape), it maintains co-referentiality throughout,[1] as in the following example taken from Huddleston & Pullum (2002):

(33) *Wash a bunch of fresh spinach well and then shred it finely. Sauté it in a little butter until it is wilted, drain Ø, then put a little into each ramekin.* *(Huddleston & Pullum 2002: 1457, ex. 17)*

For a more exhaustive description of these and other issues, see the relevant sections of the RefIND annotation guidelines (Schiborr et al. 2018: 5–14). Lastly, it should be mentioned that there is ongoing controversy about the referentiality of bound person markers in some languages, which will be briefly touched on in the discussion in Section 4.2.2 below.

4.1.3 | Pragmatic choice

The second selection criterion concerns free alternation of referring expression, which is arguably at the core of the notion of referential choice. Here, only positions where speakers may exercise a pragmatic choice of form are considered, and specifically only those positions where lexical expressions are one of the available options, which excludes most, but not all, first and second person expressions.[2] See Section 4.6.3 for concomitant methodological issues.

Crucially, the potential for variation only considers alternations in the form of NPs, not in the syntactic structures they are embedded in. This definition is based solely on the potential for alternation within the confines of grammaticality, not on acceptability; it hence captures both of what Grüning & Kibrik (2005: 172–173) call "categorical" and "potentially alternating" referential choices, where among the former observed tendencies are absolute, but variation is still theoretically possible. The criterion of pragmatic choice

1 In essence, this means the physical constitution of the ship of Theseus is irrelevant, so long as it is still considered to be the ship of Theseus in the speaker's mental representation of it.

2 Use of lexical NPs for second person reference (e.g. *would Ma'am like to see the menu?*) is of marginal frequency and has been excluded. Consider also first and second person pronouns that have grammaticalized from nouns (e.g. Indonesian *saya* 'I' < *sahaya* 'servant') (Cysouw 2003: cf.), which form an intermediary case. For the purposes of this study, these are treated as first/second person pronominal expressions.

conversely specifically excludes structurally conditioned forms such as gaps in relative and suppressed arguments of non-finite clauses, as well as reflexives.[3] The exact grammatical constraints on these positions differ between languages; compare participial clauses in English, which categorically forbid subject realization, with the corresponding constructions in Tabasaran, where no such constraint is in effect, even if overt subjects remain extremely uncommon in non-finite clauses. Similarly, the requirement that positions must be able to alternate with lexical expressions specifically rules out, among others, relative pronouns in English, which only alternate with zero.

Any of the expressions excluded by this selection criterion may nevertheless serve as antecedents for later co-referential mentions, as in (34), where the antecedent of *Father* (index *0027*) in (34b) is the second-person pronoun *you* in (34a):

(34) English

a. *He says, Give us a fiver for it, Edward, and you can have it.*

	he	*says*		*give*	*us*
	3SG.M	say.PRS.3SG		give.IMP	1PL.OBL
##	pro.h:s_ds	v:pred	##ds 0.2:a	v:pred	pro.1:p
	0026		0027		0026

a	*fiver*	*for*	*it*	*Edward*	*and*	*you*
a	fiver	for	3SG.N	Edward	and	2SG
ln_deti	np:p2	adp	pro:obl	pn_np.h:voc ##ds	other	pro.2:a
			0029	0027		0027

can	*have*	*it*
can	have.INF	3SG.N
lv_aux	v:pred	pro:p
		0029

'He says, Give us a fiver for it, Edward, and you can have it.'

3 As noted earlier, structurally conditioned gaps are included in Ariel (1990)'s accessibility marking scale despite not being part of speakers' discourse planning due to considerations of intra-sentential constraints on referring expressions. As per the the criterion of pragmatic choice, these forms fall outside of the scope of accessibility-influenced choice.

b. And so Father gave him a fiver for this horse.

	and	*so*	*Father*	*gave*	*him*	*a*	*fiver*
	and	so	Father	give.PST	3SG.M.OBL	a	fiver
##	other	other	pn_np.h:a 0027	v:pred	pro.h:p 0026	ln_deti	np:p2 0030

for	*this*	*horse*
for	PROX.SG	horse
adp	ln_dem	np:obl 0029

'And so Father gave him a fiver for this horse.'

[mc_english_kent02_0027-0028]

4.1.4 | Further considerations

4.1.4.1 | Noun phrase subconstituents

Where a referring expression is composed of multiple other referring expressions, only the primary reference instantiated by the entire phrase is counted. This means that any anaphoric mentions in subordinated NP modifiers, such as phrasal modifiers in (36) or possessives in (35) are not captured except as potential antecedents for later expressions.

(35) English

She clears that colt's mouth so that it can breathe.

	she	*clear-s*	*that*	*colt='s*	*mouth*
	3SG.F	clear-PRS.3SG	DIST.SG	colt=POSS	mouth
##	pro:a 0324	v:pred	ln_dem	ln_np:poss 0320	np:p 0333

	so	*that*	*it*	*can*	*breathe*
	so	that	3SG.N	can	breathe.INF
#ac	adp	adp	pro:s 0320	lv_aux	v:pred

'She clears that colt's mouth so that it can breathe.'

[mc_english_kent03_0308]

(36) English

A farmer down Churston Court, that's the farmer beside the church, he bought one of they cows.

	a	*farmer*	*down*	*Churston Court*		*that*
	a	farmer	down	Churston_Court		DIST.SG
##	ln_deti	np.h:dt_a	rn_adp	rn_pn_np	#	dem_pro.h:s
		0104		0105		0104

='s	*the*	*farmer*	*beside*	*the*	*church*	
=be.PST.3SG	the	farmer	beside	the	church	
=cop	ln_det	np.h:pred	rn_adp	rn_det	rn_np	%
		0104			0106	

he	*bought*	*one*	*of*	*they*	*cow-s*
3SG.M	buy.PST	one	of	DIST.PL	COW-PL
pro.h:a	v:pred	np:p	rn	rn_dem	rn_np
0104		0107			0067

'A farmer down Churston Court, that's the farmer beside the church, he bought one of those cows.' [mc_english_devon01_0047]

Due to the limitations of the corpus annotations, the above-mentioned also applies to coordinated NPs as in (37).

(37) Vera'a

Dirēk vesir ēn 'amagi won vēvēgi so, …

	dir	*=ēk*	*vesir*	*ēn*	*'ama*	*-gi*
	3PL	=TAM2	ask	ART	father	-3SG
##	pro.h:a	=lv	v:pred	ln	np.h:p	-rn_pro.h:poss
	0024				0026	0024

wo	*=n*	*vēvē*	*-gi*	*so*
and	=ART	mother	-3SG	QUOT
rn	=rn	rn_np.h	-rn_pro.h:poss	other
		0025	0024	

'They asked their father and mother, …' [mc_veraa_iswm_0059]

The second and subsequent elements in the list are annotated the in same way as subordinated elements, so that only the first element is counted as the head of the phrase. However, since coordinated NPs of this kind appear quite seldomly in the data (a rough estimate puts them at a frequency of about 1 every 40 clauses, less when only considering subject and object NPs), the loss of information incurred by their exclusion is minimal.

4.1.4.2 | Clausal and VP references

For the sake of simplicity, we focus exclusively on nominal expressions (and zero anaphors) in this study. Clause-level anaphors such as headless relative clauses as in (38) are excluded, although they technically meet the criterion of pragmatic choice and other criteria outlined above.

(38) Vera'a
Qo' sas van kelkel 'i 'anē!

	Qo'	*sa*				
	Qo'	EMPH				
##ds	pn_np.h:pred	rn				
	0001					
		=s	*van*	*kēlkēl*	*'i*	*anē*
		=SIM	go	RED:back	DEL	DEM1.A
#ds_rc:s	0.h:s	=lv	v:pred	rv	other	other
0001	0001					

'[They said,] It must be Qo' (who) is walking there.'

[mc_veraa_jjq_0250]

Constructions of this kind constitute a class of super-heavy expressions that is distinct from other forms of reference (see Section 8.2), and so merit further examination, but are unfortunately beyond the scope of the present study.

4.1.4.3 | Incomplete and interrupted segments

In unrehearsed spoken language, false starts and incomplete or interrupted segments of the kind in (39) are not uncommon:

(39) English
He said, They— I said, They won't be doing you any good.

	he	*said*	*they*		*I*	*said*
	3SG.M	say.PST	3PL		1SG	say.PST
#nc	nc_pro.h 0199	nc	nc_pro 0210	##	pro.1:s_ds 0000	v:pred

	they	*won't*	*be*	*doing*	*you*	*any*
	3PL	will.NEG	be.INF	do-PTCP.PST	2SG	any
##ds.neg	pro:a 0210	lv_aux	lv_aux	v:pred	pro.2:p 0199	ln_deti

good
good
np:p2

'He said, They— I said, They won't be doing you any good.'
[mc_english_devon01_0097]

As mentioned above in Section 3.3.1, only the repaired expressions are fully annotated, as they are what the speaker deems to be the intended form (cf. Gipper 2016: 150). As such, anaphors occuring in such segments are excluded from the sample, and only complete mentions in fully-formed clauses are included. However, the content of these segments does affect activation states of referents, as even if speakers does not commit to them, they are still written to the discourse record (Lowder et al. 2018), provided they are intelligible. As such, anaphoric mentions in false and starts and interrupted clauses may serve as antecedents to later co-referential mentions for the purposes of, for instance, calculating anaphoric distances. Likewise, they are counted for any factors that consider the properties of the surrounding discourse, such as competition.

Do note however that the intended grammatical role of these types of antecedents is often difficult to ascertain, and so as a rule is left unannotated in the Multi-CAST texts. For the purposes of our analysis, their role is set to default to the unspecified 'other' category; refer to Section 4.3 later in this chapter for an outline of the definitions of the role categories used in this study.

4.1.4.4 | Secondary speakers

The texts in the Multi-CAST collection are mostly monological, but do get interrupted by other speakers on occasion, usually by audience members commenting on the narrative or asking for clarification as in (40):

(40) Nafsan

a. *Me natap ita pi natopu?*

	me	*natap*	*i=*	*ta*	*pi*	*natopu*
	but	idol	3SG.RS=	not	be	spirit
#nc	nc	nc_np.h	nc	nc	nc	nc_np.d
		0021				0027

'[An audience member asks,] But the idol is not a spirit?'

b. *E, natopu teƥtae.*

	E	*natopu*	*te-ƥtae*
	e	spirit	DET-different
##	other	np.d:s	np:pred
		0027	

'[The speaker answers,] The spirit is different.'

[mc_nafsan_lelep_0020-0021]

For this study, only anaphors in the speech of the primary speaker are sampled, but as with incomplete and interrupted segments, mentions in the speech of other speakers are still considered for the identification of antecedents and the properties of the surrounding discourse. Due to limitations in the annotations of these segments, the role of mentions in them also defaults to the unspecified 'other' category, even if these clauses are in most cases well-formed.

4.2 | Definition of expression types

For the purposes of this study, "lexicality" is defined fairly broadly, as a "comparative concept" (Haspelmath 2010), to facilitate cross-linguistic comparison. For most analysis, the significant split is between expressions with nominal heads (*the woman*, *Jane*; Section 4.2.1) on the one hand, and various other, largely pronominal expressions (*she*, *her*; *that*) and zero (Section 4.2.2) on the other. Each is briefly outlined in the following sections, along with their concomitant analytical issues. For further details on language-specific issues in coding and classification, refer to the annotation notes for each Multi-CAST corpus (Haig & Schnell 2015).

4.2.1 | Lexical expressions

For the purposes of this study, any noun phrase that meets is selection criteria in Section 4.1 and is not one of the reduced forms defined in Section 4.2.2 is counted as a lexical expression. The deliberately broad and under specified definition of lexicality used here is meant to allow for easier comparison between languages. Which expressions are considered eligible is ultimately based on the formal morphological and syntactic criteria for lexical NP'ness in each language, but the core criterion is the ability to bear anaphoric reference and function as an argument of a predicate while not being headed by a pronominal form (see Section 4.2.2.1 below). This largely aligns with the definition of the ⟨np⟩ form symbol and its variants in the GRAID annotation scheme (Section 3.3.1) on which the analyses in this study are based, that is, interlocutors' use of conceptual or name-giving knowledge to identify or establish a referent. In addition to NPs with common-noun heads (e.g. *tree*, *girl*, *idea*, etc.), which make up the majority of lexical expressions in the sample, the sample also includes expressions such as proper names, kin and address terms (e.g. *Mother*), de-numeral nouns (e.g. *the one*, *the three*; but not anaphoric *one*, Payne et al. 2013) and other nouns with little semantic content (e.g. *the other*), whose inclusion could be considered contentious. The cross-linguistic nature of the corpus data inevitably raises theoretical questions regarding the inclusion and exclusion of certain types of nominal expression. These (and other issues) are beyond what can reasonably be explained within the scope of this study, and I instead direct readers to the extensive annotation notes for each of the Multi-CAST corpora (see citations in Section 3.2.3).

For certain parts of this study, lexical NPs are further subcategorized along the following three axes, each defined in the following sections:

- nominal expressions with common and proper heads (Section 4.2.1.1),
- light and heavy expressions (Section 4.2.1.2), and
- nominal expressions with and without demonstrative modifiers (Section 4.2.1.3).

These expression types are not distinguished in the sample used for the main analyses in Chapters 6 and 7, which focuses on the more basic lexicality distinction, only in Chapter 8 further below.

4.2.1.1 | Proper names

Referring expressions with proper noun heads mostly fall into one of a few broad categories: personal names and the names of products and organizations (e.g. *Jane*, *Ms Rigby*; *Samwell's Shipyard*), placenames (e.g. *London*, *the Isle of Wight*), events (e.g. *World War Two*), and titles and other codified descriptors (*Mother and Father* vs. *my mother and father*). Of these, the first and last are the ones to occur the most frequently in the sample, as locations and events seldomly occupy core argument positions.

The last of the categories listed above deserves further comment. Codified descriptors, as I call them here, are common noun labels used in lieu of proper names, usually titles or role labels used to refer to very prominent referents, such as *the queen* or *the vizier* in folkloristic narratives, which may attain a status of rigidity that resembles that of proper names through frequent enough use. In essence, these are descriptors that are used to refer to specific characters so routinely that they become unique identifiers rather than class labels. The line between the two is blurred especially in languages in which common and proper nouns are not formally distinguished, such as in Northern Kurdish. Conversely, in other languages such as English, where most codified descriptors lack determiners (*Father* vs. *my father*), and Vera'a, where proper names receive specialized marking via personal articles that are distinct from the articles used with other types of nouns, the distinction is easier to make based on formal criteria.[4]

4 Compare in this context the tendency for definite articles in Maltese English to be omitted

Lastly, it bears mentioning that the usage rates of proper names differs between the two text types in the sample. Proper names are overall more common in the biographical texts than in the traditional narratives, as their occurrence is highly content-dependent. As such, there are texts in the corpora in which no proper names appear, which makes them ineffective for studies of proper name selection. We will return to this issue and how it limits the scope and validity of parts of the present study in Section 8.1 below.

4.2.1.2 | Light and heavy NPs

Phrasal complexity is a measure of the informativity of a referring expression, with more complex, heavier phrases being associated with lower accessibility and greater identificational ambiguity (see Section 2.5.2). For the sake of simplicity, the analyses presented in this study adopt a relatively undifferentiated classification system with only two levels, in essence distinguishing maximally simple NPs – those with no additional modifiers or markers beyond the obligatory – and those with any kind of additional modifiers. The definition of phrase weight used here is hence not so much based on syntactic complexity as on information content and specificness of reference. The presence of one or multiple of the following modifiers marks out a given noun phrase as heavy:

- attributive adjectives: *the old school* (41); 'this little girl' (42),
- possessive determiners: 'her father' (42); 'this king's daughter' (43),
- attributive numerals: 'his own two daughters' (44)
- adpositional modifiers: *the books on the table*,
- coordinated NPs: *the books and magazines*, as well as
- relative clauses: *the book she bought*.

Of these, possessive modifiers are by far the most common, accounting for almost half of the instances in the sample ($P = 0.47$), followed by adpositional and adjectival modifiers ($P = 0.29$) and relative clauses ($P = 0.11$). These expressions are annotated variously in GRAID, as exemplified by the following:

"when the uniqueness or identifiability of a referent is salient in context" (Krug & Lucas 2018: 261), as in *But government yesterday distanced itself from the censorship* (2018: 289, ex. 68). See also (Rupp & Tagliamonte 2019) for related findings in York English.

(41) English

They've turned the old school into dwellings.

	they	*='ve*	*turn-ed*	*the*	*old*	*school*
	3PL	=have.PRS	turn-PTCP.PST	the	old	school
##	pro.h:a	=lv_aux	v:pred	ln_det	ln	np:p

into	*dwelling-s*
into	dwelling-PL
adp	np:obl

'They've turned the old school into dwellings.'

[mc_english_devon01_0006]

(42) Vera'a

En 'amagi ne 'ēn gōr ēn ni'i reñe 'anē.

	e	*=n*	*'ama*	*-gi*	*ne*
	DISC	=ART	father	-3SG	TAM2:3SG
##	other	=ln	np.h:a	-rn_pro.h:poss	lv-pro_h_a

'ēn	*gōr*	*ēn*	*ni'i*	*reñe*	*anē*
see	secure	ART	small	woman	DEM1
v:pred	rv	ln	ln	np.h:p	rn_dem

'And her [= the girl's] father looked after this little girl.'

[mc_veraa_anv_0006]

(43) Tabasaran

Qa mu pač:ihǯin riš šad šulu.

	qa	*mu*	*pač:ihǯi-n*	*riš*	*šad*	*šul-u*
	then	PROX(ATTR)	king-GEN	girl(ABS)	glad	become-FUT
##	other	ln_dem	ln_np.h:poss	np.h:s	other	v:pred

'Then this daughter of the king was delighted.'

[mc_tabasaran_horse_0197]

(44) Mandarin
Ā jiào gěi le zìjǐ de liǎng gè nǚér.

	ā	*jiào*	*gěi*	*le*	*zìjǐ*		*de*
	MP	teach	give	ASP	REFL		MOD
##	0.h:a	other	v:pred	rv_svc	rv	ln_refl.h:poss	ln

liǎng	*gè*	*nǚér*
two	CL	daughter
ln_num	ln	np.h:p

'He taught his own two daughters.' [mc_mandarin_hml_0018]

4.2.1.3 | Demonstrative modifiers

The third and final type of lexical expression examined in this study concerns those with demonstratives as used as determiners. In the GRAID annotations, demonstrative determiners are annotated as ⟨ln_dem⟩ or ⟨rn_dem⟩, depending on their position relative to the head noun of the NP:

(45) English
And 'course that old mare knew her job.

	and	*'course*	*that*	*old*	*mare*	*knew*	*her*	*job*
	and	of_course	DIST.SG	old	mare	KNOW.PST	3SG.F.POSS	job
##	other	other	ln_dem	ln	np:a	v:pred	ln_pro:poss	np:p

'And 'course that old mare knew her job.'
[mc_english_kent03_0039]

(46) Sanzhi Dargwa
Ilil aždahal il k:alk:ira buk:unne, …

	il-i-l	*aždaha-l*	*il*	*k:alk:i*	*=ra*
	that-OBL-ERG	monster-ERG	that	tree	=and
#cv	ln_dem	np.h:a	ln_dem	np:p	=other

b-uk:-un-ne
N-eat.IPFV-PRET-CVB
v:pred

'The monster was also eating that tree, …' [mc_sanzhi_dragon_0031]

As noted earlier in Section 2.2, demonstratives are being given special consideration in a number of models of referential choice, most notably accessibility theory (Ariel 1990), where they occupy higher levels on the accessibility scale compared to plain NPs.

Note that in many languages in the sample, demonstrative determiners and demonstrative pronouns are formally equivalent (bar the presence of case markers, etc.), meaning it is the co-occurrence with a head noun that is distinctive. Demonstrative pronouns (⟨dem_pro⟩ in GRAID) are counted as non-lexical expressions instead; see Section 4.2.2.1 below.

Ariel (1990) further holds (for English) that speaker-proximal demonstratives (*this*, *these*) indicate higher referent accessibility than speaker-distal demonstratives (*that*, *those*), both as NP modifiers and as pronouns, in effect reflecting the semantics of these expressions. We here do not distinguish demonstrative determiners further by type, given that not all languages in the sample used for this study draw the same proximal–distal (or proximal–medial–distal) distinctions, and not along the same lines – consider, for instance, the highly complex demonstrative systems found in Sanzhi Dargwa and Tabasaran, a characteristic of Nakh-Daghestanian languages – which poses considerable obstacles for cross-linguistic comparability, and that there is preliminary evidence contrary to the claims in Ariel (1990), with Schiborr (2017) finding that anaphoric distance, a key determinant of accessibility, has no bearing on the selection of proximal and distal demonstratives in spoken English discourse.

4.2.2 | Non-lexical expressions

The other category of expressions in the sample are those that are reduced, that is do not carry lexical information in the manner that noun phrases headed by common or proper nouns do. As such, the definition of what constitutes a reduced expression is inherently less specific than the corresponding one given for lexical expressions above. In essence, any noun phrase that meets is selection criteria in Section 4.1 and does not have a nominal head is counted as a non-lexical expression. Additionally, we are only interested in expressions occuring in positions that alternate with lexical expressions, corollary to the criterion of pragmatic choice.

There are two broad types of non-lexical expressions examined here: pronominal expressions of various kinds (Section 4.2.2.1), including demonstratives and certain bound pronominal forms, and zero anaphors (Section 4.2.2.2). While the choice between these two types of expression is central to work on referential choice (cf. e.g. Torres Cacoullos & Travis 2019 for a recent example), they are here subsumed into a single category, which follows naturally from the focus on lexical choice in particular in this study.

4.2.2.1 | Pronominal noun phrases

As discussed above in Section 3.3.1, the GRAID annotations by design do not register many of the minute language-specific distinctions among forms, but rather something akin to broader, comparative categories (Haig & Schnell 2014: 8–9). This generalized definition is leveraged for the analyses in this study, which are deliberately designed to be widely inclusive so as to capture the full range of options at speakers' disposals – provided these options freely alternate with lexical expressions. As such, we only exclude form types that do not meet the other selection criteria outlined in Sections 4.1.2 and 4.1.3 above, such as indefinite pronouns, relative pronouns, as well as reflexive and reciprocal expressions. The overwhelming majority of the pronominal expressions included in the analysis (i.e. third person only) are free definite pronouns, especially personal pronouns ($P = 0.80$) and demonstratives ($P = 0.11$).[5]

However, there is one contentious type of reduced referring expression that merits a few additional comments: pronominal affixes/bound person indices (see Lehmann 1988; Mithun 2003; Kibrik 2011; Haig & Forker 2018; among many others, see also Haig & Schnell 2014: 31–45). Corbett (2006) observes that anaphora are related to the phenomenon of agreement, which replicates specific features of a controller, just as anaphors replicate certain features of an antecedent, though agreement usually involves much tighter relations with the controller compared to those between anaphor and antecedent. Co-occurrence with a controller has been typically taken as a key diagnostic for distinguishing (non-referential) agreement from (anaphoric) pronouns; Bresnan & Mchombo (1987), for instance, assume a categorical

5 As noted above in Section 3.2.3, two of the languages examined here (Sanzhi Dargwa and Tabasaran) lack a paradigm of third-person pronouns that is distinct from demonstratives.

distinction between agreement and pronouns based on the presence of co-nomination, that is whether a further, usually more informative, expression co-occurs with the index in the same clause.[6] In a similar vein, Falk (2006: 55) considers "agreement on the verb is itself [to be] a realization of the pronominal argument, a kind of incorporated pronoun." However, Haig & Forker (2018: 719) note that this approach is "conceptually problematic, because it conflates […] obligatoriness of the person marker […] and a language-specific tolerance [for referential zero] subjects." They argue that these two dimensions are logically independent and need to be kept apart. Instead, Haig & Forker (2018: 718) note that cross-linguistically, "there are pervasive parallels and overlaps between anaphora and agreement," which lead "to a multiplicity of hybrid forms and mixed systems." As such, the division between anaphora and agreement can be difficult to maintain in all cases even within a single language, much less typologically. They instead suggest that both anaphoric pronouns and bound agreement markers can be considered potentially referential devices, which are deployed according to conditions of pragmatic appropriateness (Haig & Forker 2018: 718; cf. Lehmann 1982; de Groot & Hengeveld 2005; Croft 2013).

For the present analysis, we adopt the selection criteria that already guide the annotation of person indices in the GRAID scheme (see Section 3.3.1 above), which operates on two criteria: obligatoriness of the bound marker (cf. Corbett 2003) and the potential for (rather than the actual presence of) co-nomination. The English third-person index (e.g. *she speaks*), for instance, is classified as agreement, as its presence is required irrespective of what the remainder of the clause looks like. If the index in question is not obligatory, co-nomination becomes decisive: GRAID adapts Haspelmath's (2013) three-way typology of "gram-indices", "cross-indices", and "pro-indices" – where the first two types, gram-indices and cross-indices, respectively require or optionally permit co-nomination, the third type, pro-indices, prohibits it, and hence never enter into a co-reference relation with a co-nominal in the same clause that could be considered "agreement". As such, apart from their phonological dependence on a host, these forms resemble free pronouns in most regards, including in terms of their information content. Pro-indices and

6 Compare the various criteria of uniqueness of reference formulated in, e.g., government and binding theory (Chomsky 1981, 1982) and lexical-functional grammar (Falk 2006: 55–56).

cross-indices show similar referential characteristics; in some languages, pro-indices occur in free variation with free pronominal or lexical NP expressions on the clause level, and the latter type of expression would be used, for instance, in contrastive or otherwise pragmatically marked contexts. For the purposes of this study, only those bound forms that can be categorized as pro-indices according to Haspelmath (2013) are counted. Where verbal indices allow for co-nominals (i.e. cross-indices), but none is present, we instead assume a zero anaphor, and accordingly do not analyze the index as referential.

4.2.2.2 | Zero anaphors

Zero anaphors are fairly easy to define in terms of formal criteria, as they are simply the absence of any overt form. The crucial question is deciding which instances of argument omission constitute cases of pragmatically selected (rather than structurally conditioned) zero that evoke a specific discourse referent; the relevant criteria have already been discussed in the context of defining the ⟨0⟩ annotation symbol in the GRAID scheme and the concomitant discussion in Section 3.3.1 above. Beyond these criteria, the sample is limited to those zero anaphors that further meet the criterion of pragmatic choice discussed above in Section 4.1.3. For instance, in this example from Sanzhi Dargwa, the subjects of both clauses are omitted (as is very common in this language), but could freely be replaced with overt forms. They also point to the same, clearly identifiable discourse referent:

(47) Sanzhi Dargwa
k'ult'a qaˁškaʁible t:ura haq:ibcab... cinna rucbe.

	k'ult'a	*qaˁš*	*k-aʁ-ib-le*	
	belly	cut	down-do.PFV-PRET-CVB	
## #cv	0.h:a	np:p	other:lvc v:pred	%

	t:ura-h-aq:-ib	*ca-b*
	outside-upwards-take_out.PFV-PRET	be-HPL
0.h:a	v:pred	rv_aux

cin-na	*ruc-be*
REFL.SG-GEN	sister-PL
ln_refl.h:poss	np.h:p

'Cutting open the (wolf's) belly, (they) took her sisters out.'
[mc_sanzhi_patima_0036]

4.3 | Role definitions

The present study concerns itself with referential choices in subject and object position, largely to the exclusion of other syntactic positions. Similar to form categories, the definitions of grammatical roles used appeals to relatively broad generalizations in an effort to capture the highly variable patterns found in the languages of the world, and so to enable comparison across diverse sets of languages. Since at least Keenan (1976), this has been a topic fraught with many complications and laden with unresolved controversy (cf. e.g. Falk 2006; Andrews 2007; among many others). We will only tangentially touch on these issues here, as chiefly we aim for definitions that are suitable to the task at hand, even if they are not quite perfect.

The definition of grammatical roles follow from the definitions of syntactic functions in Andrews (2007: 137–140), have already been outlined in Section 3.3.1 on the corpus annotations above. To recapitulate, Andrews's account rests on the notion of a semantic transitive prototype. Any construction with two arguments that share the marking properties of the agent and patient roles of a prototypical transitive event (such as *smash* or *kick*) in a language is identified as transitive, and its arguments classified as carrying "A" and "P" function respectively. Any clause lacking either an A or P argument or both is classified as intransitive, and may contain a single core argument function "S" and other oblique arguments. Crucially, only those clauses that have arguments coded identically to an A and a P argument are deemed transitive, though neither needs be overtly realized.[7]

Andrews (2007: 139) distinguishes the syntactic functions S, A, and P from grammatical relations such as subject and object. The relation between syntactic functions and grammatical roles remains a topic of controversy in the literature, but a substantial body of research suggests that they do represent a valid level of syntactic description, and, more importantly, provide a framework within which significant cross-linguistic generalizations on the possible

7 This interpretation contrasts with recent views such as Dixon (2009: 151), who extends these functions to arguments not marked the same as the A and P arguments of a primary transitive verb. Similarly, most approaches simply take S to be the "single argument of a one-place predicate"; consider for instance Donohue (2008), which classifies the single argument of any monovalent verb as S, regardless of marking.

shapes of grammars can be formulated (Comrie 1989; Farrell 2005; Andrews 2007; Haspelmath 2011; among many others).

As noted earlier, we here group S and A arguments together into a combined notion of "subject" on purely methodological – rather than theoretical – grounds, and align P arguments with the notion of "object".[8] Clauses with A arguments are counted as transitive, clauses with S arguments as intransitive. Since S and A have been found to have different discourse profiles (see Section 2.2.5.1), we include transitivity as a factor in the multifactorial analysis part of this study; more on that in Sections 4.5 and 6.12. There is a subset of languages that treats certain S arguments differently, which we will get to below in Section 4.3.1. The next sections also discuss the treatment of the ergative languages in the sample (Section 4.3.2) and briefly touch on the issue of ditransitive constructions (Section 4.3.3).

4.3.1 | Non-canonical subjects and objects

A problematic case for our definition of grammatical roles are those arguments that are marked differently from S, A, and P arguments as defined above, which are usually labelled "non-canonical" subjects and objects. Non-canonical subjects in particular generally do not align with prototypical transitivity, as they tend to not be agentive. As a result, it is possible that languages treat them differently from canonical subjects, though they are generally understood to nevertheless be, for all intents and purposes, subjects (Helasvuo & Huumo 2015). Crucially, canonical and non-canonical subjects tend to pattern similarly in discourse.

In the present sample of languages, non-canonical subjects chiefly show up in Sanzhi Dargwa, Tabasaran, and occasionally in Northern Kurdish, usually as Experiencers and with certain marginal predicates as in (48), and are annotated as ⟨:ncs⟩ in GRAID:

8 In a general sense, we consider the most workable definition of subjecthood to align with the notion of the "privileged syntactic argument" of a clause as used in role and reference grammar (Van Valin 2005), according to which subjects are the convergence point of a number of properties that are relevant to most languages.

(48) Tabasaran
Ja muvaz mal gundar.

	ja	*muva-z*	*mal*	*gun-dar*
	or	PROX-DAT	cattle(ABS)	want-PRS.NEG
##neg	other	dem_pro.h:ncs 0005	np:p	v:pred

'Neither did he want cattle.' [mc_tabasaran_naz_0030]

As they are not particularly frequent ($P = 0.03$ of subjects in Sanzhi, $P = 0.06$ in Tabasaran, and one case in Northern Kurdish) they are here grouped together with intransitive predicates for considerations of transitivity (Section 6.12).[9]

4.3.2 | Ergativity

Three of the languages in the sample, Sanzhi Dargwa, Tabasaran, and Northern Kurdish, have ergative alignment in morphology, that is the subject of an intransitive clause (i.e. S) is marked the same as a direct object (P) (cf. Dixon 1994), though in the case of Northern Kurdish it is limited to the past tense (split alignment). A arguments receive ergative case marking, the absolutive case on S and P arguments is unmarked:

(49) Sanzhi Dargwa
Pat'imal has:ible q:apra, agurcar wac'ac:e.

	pat'ima-l	*h-as:-ib-le*	*q:ap*	*=ra*
	Patima-ERG	upwards-take.PFV-PRET-CVB	sack	=and
## #cv	pn_np.h:a_cv	v:pred	np:p	=other %

	ag-ur	*ca-r*	*wac'a-c:e*
	go.PFV-PRET	be-F	forest-INE
0.h:s	v:pred	rv_aux	np:g

'Patima took a sack and went into the forest.'
[mc_sanzhi_patima_0010]

9 Notably, virtually all instances of non-canonical subjects in the corpus data involve human referents ($P = 0.96$).

This way of classifying A and S arguments leads to a small number of fringe cases where an ergative-marked subject is analyzed as A even if the predicate does not license an absolutive-marked P argument, contrary to the criteria given above (i.e. number of core arguments superseding marking properties). This mostly occurs with complex predicates as in (50), where the non-verbal complement of the vector verb is not analyzed as an argument (though in some cases the status of nominal complements can be fuzzy; cf. Butt 2003; 2010):

(50) Sanzhi Dargwa
wallah, ʁubza hel xunul admil c'aq'le ħaʕħaʕ darq'ib.

	wallah	*ʁubza*	*hel*	*xunul*	*admi-l*	*c'aq'-le*
	by_god	EMPH	that	woman	person-ERG	very-ADVZ
##	other	other	ln_dem	ln	np.h:a_cps	other

	ħaʕħaʕ	*d-arq'-ib*
	laughter	NPL-do.PFV-PRET
	other:lvc	v:pred

'By god, that woman laughed very much.' [mc_sanzhi_tape_0023]

Here, the vector verb (or 'light verb') 'do' assigns ergative case to the subject, but entire predicate 'laugh' (lit. 'laughter do', with mimetic *ħaʕħaʕ* 'laughter') lacks an absolutive-marked object. In these (quite rare) cases, the subject is counted as an A argument and the clause is treated as transitive.

4.3.3 | Ditransitive constructions

In ditransitive primary-object constructions where both the Recipient and Theme are coded as P arguments (Malchukov et al. 2010; Margetts 2007), both are counted as objects for the purposes of this study. That is, since the Recipient *mi mother* in example (51) is coded the same way as the Patient of a primary transitive verb in English, it is analyzed as a P argument. The secondary object expressing the Theme, *a letter*, is also counted as a P argument for the same reason.

In other languages, the corresponding arguments can be coded differently: For instance, the Recipients of transitive verbs of transfer (e.g. *give*) may coded as goals; this is also a possibility in English (i.e. dative alternation). As

is generally the case in GRAID, formal morphosyntactic coding properties take precedence over semantics. Primary and secondary objects are distinguished by their function symbols, these being ⟨:p⟩ for objects expressing the recipient and ⟨:p2⟩ for those expressing the theme.

(51) English
The priest there gave mi mother a letter.

	the	*priest*	*there*	*gave*	*mi*	*mother*
	the	priest	there	give.PST	1SG.POSS	mother
##	ln_det	np.h:a	rn	v:pred	ln_pro.1:poss	np.h:p

a	*letter*
a	letter
ln_deti	np:p2

'The priest there gave my mother a letter.'

[mc_english_devon01_0003]

4.4 | Summary of the sample data

A tabular overview of the corpus data is provided in Table 4.1: Listed there are the number of unique speakers, texts, and individual clauses, as well as the total number of identifiable discourse referents and number of those with mentions that meet the selection criteria outlined in this chapter, in addition to the final number of subject and object mentions that will be analyzed in the following; mentions in oblique positions are not shown here, as they will not be relevant beyond Chapter 5, where their numbers are listed separately.

Notably, only a little over a third of all identifiable referents in the corpora have mentions that meet the sampling criteria (cross-corpus mean $P = 0.38$, with a standard deviation of $\sigma = 0.066$). This means that the majority of referents are either never mentioned in subject or object position, or are only ever mentioned once in a text. The latter case is particularly common, with single-mention referents making up almost half ($P = 0.56$, $\sigma = 0.084$) of all referents in the ten corpora. Among referents with mentions that do meet the selection critieria, each is mentioned 7.32 time on average (cross-corpus mean, $\sigma = 2.057$).

	total number of			referents		mentions	
corpus	speakers	texts	clauses	all	sampled	subjects	objects
C. Greek	1	3	1070	296	119	441	235
English	3	4	4184	1933	616	1355	723
Mandarin	3	3	1194	466	140	719	203
Nafsan	4	9	1012	262	126	698	228
N. Kurdish	1	2	1359	297	117	643	257
S. Dargwa	4	8	1066	359	121	524	125
Tabasaran	2	5	1383	398	134	786	244
Teop	4	4	1303	267	93	751	255
Tulil	5	6	1264	383	149	620	237
Vera'a	10	10	3608	656	314	2430	614
totals	37	54	17443	5317	1929	8967	3121

Table 4.1 | Overview of the corpus data.
The 'all referents' column contains the total number of discourse entities referred to in the corpus data, while 'sampled referents' are those matching the sampling criteria outlined in this section (i.e. types); 'mentions' are the anaphoric instantiations (i.e. tokens) of the sampled referents in subject or object position.

Lists of the texts and speakers included in the sample used for this study along with their associated metadata can be found in Appendix B.

4.5 | Tested properties

As outlined in Section 2.6, one of the aims of this study is to construct multifactorial models to help explain and predict the choice of lexical expressions for anaphoric mentions cross-linguistically (Chapters 7 and 8). As such, we expect certain factor levels to align with either non-lexical or lexical expressions (or neither), and do so either independently or in combination, and in the same way for subject and object mentions, or differently for each role. Every factor is tested for every observation in the sample, to yield a full picture of variation across all observed contexts.

This section lists and briefly outlines the twelve factors tested in the main part of this study and the motivations for their selection. How exactly each

of them is defined and calculated is described in detail in the respective sections in Chapter 6. The tested factors can be split into four broad categories: those that derive from the inherent semantic properties of the referent in question (such as humanness), those that capture the properties of the surrounding discourse (such as the number of recent co-referential mentions) and the properties of the immediate antecedent (such as syntactic position and form), and those that relate to the structural properties of the clause and the predicate (such as clause dependency and transitivity).

1. *Animacy and humanness* (Section 6.1). Animacy distinctions span a nuanced continuum of values (Dixon 1994; Yamamoto 1999), but due to the relative scarcity of non-human referents in subject position especially, we will restrict our analysis here to a broad humanness distinction, and examine finer animacy distinctions only in passing.
2. *Total mention frequency* (Section 6.2). The total token frequency of a referent in a text – relative to the frequency of all other referents – is used here as a quantitative estimate of protagonisthood (Dooley & Levinsohn 2001; Section 2.4.1.3), as coding protagonist status manually for every referent in the corpus data (as is done, e.g., in Kibrik et al. 2016; Kibrik 2000 for much smaller samples) is not practical.
3. *Anaphoric distance* (Section 6.3). Recency effects are deemed the single most influential discourse-based contributor to referent accessibility in most theories of discourse structure (e.g. in accessibility theory, Ariel 1990 and centering theory, Grosz et al. 1995) as well as in many computational approaches to anaphora (see references in Section 2.3. Recency is most commonly quantified in terms of anaphoric distance, which is the length of the interval between an anaphor and its antecedent. However, how exactly this interval is to be measured, and between precisely which pair of mentions, differs from approach to approach, and hence needs to be carefully defined, which we will do in Sections 4.6.1 and 4.6.2 below. A number of studies (e.g. Kibrik et al. 2016) augment the use of measures of textual distance with evaluations of rhetorical structure (in terms of elementary discourse units from rhetorical structure theory, Matthiessen & Thompson 1988), others with syntactic distance (e.g. in terms of nodes along parse trees), but these are avenues not explored for the present study.

4. *Number of recent co-referential mentions* (Section 6.4). This factor quantifies of one of the many contributors to the somewhat nebulous notion of discourse prominence (Section 2.4.2.3), which has significant overlap with the related concept of topicality (cf. Givón 1983a). In particular, this factor captures the idea that repeated mentions of the same referent in the preceding discourse serves to maintain its activation in speakers' memory. Other contributors to discourse prominence are humanness and protagonisthood, already mentioned above. This factor is expected to correlate noticeably with anaphoric distance, as a higher frequency of mentions entails lower distances between each individual mention.
5. *Number of recent mentions of related referents* (Section 6.5). Due to the way discourse referents are defined and tracked by the corpus annotations (see Section 3.3.2), measures that evaluate co-reference relations (such as anaphoric distance) are agnostic of overlaps between referents in conceptual and referential space (see Section 2.4.2.6). However, annotations that capture a number of mereological relations between referents are in place (Section 3.3.4.2), and are leveraged here as a (limited) measure of priming from semantic inferability.
6. *Number of recent mentions of competing referents* (Section 6.6). Counts of the number of mentions non-co-referential (and unrelated) referents in recent discourse here serves both as a measure of local cognitive load (cf. information pressure/referent pressure, Du Bois 2003b) and of potential competition between candidate referents. For the latter, this measure is at best a crude approximation, as we here do not (and cannot) check for the specific properties of referents that would make reduced references to them ambigious (as per Givón 1983a: 14), though it should be noted that this measure has been suggested as a valid quantification of competition in Ariel (1990).
7. *Role of the antecedent* (Section 6.7). The role of antecedence is linked to the syntactic prominence of the referent, due to a general tendency towards role persistence and repeated mentions in the same syntactic position. This factor also captures the effect of shifts in position, such as promotion to and demotion from subject.

8. *Form of the antecedent* (Section 6.8). As with the role of the antecedent, form parallelism is expected to affect form choices. Since the key distinction is likely one of lexicality, we here categorize by type of expression; an alternative approach might also capture the phrase length of antecedent, as is done in Kibrik et al. (2016).
9. *Sequence of mention* (Section 6.9). This factor relates to the notion of gradual establishment of newly introduced referents described in Section 2.4.2.7 above. As mentioned there, the only study that deals with this, to my knowledge, is Kibrik (2000), which only notes whether the antecedent is the first mention of a referent in a given text. Here, we attempt to capture the gradual transition from almost exclusively lexical introductions to fully established referent in a slightly more nuanced way, and so distinguish second, third, and later mentions of a referent.
10. *Clause type* (Section 6.10). This factor tests whether the syntactic embeddedness of clauses might affect the lexicality of mentions within them. It should be noted, however, that clause dependency is not equally well distinguishable in all languages in the sample, and that at times it can be difficult to establish exactly which matrix clause – if any – a given dependent clause belongs to. We will address these methodological issues and how they limit the interpretability and cross-linguistic comparability of this factor, in Section 6.10.
11. *Clause length* (Section 6.11). Clause length serves as measure of cognitive load in Arnold et al. (2009), where longer clauses correlate with greater processing demands, which in turn lead to a higher rate of lexical subject expression, as more informative forms tax speakers' attentional budget less strongly (Arnold et al. 2009: 11–12) Here, it is the second factor capturing the influence of cognitive load alongside the frequency of competing referents.
12. *Transitivity* (Section 6.12). The syntactic function of subjects (i.e. A vs. S) has been linked to the rate at which they are realized lexically (cf. the non-lexical-A constraint in Du Bois 1987b, see Section 2.2.5.1). For the sake of simplicity, we will refer to this factor broadly as a question of "transitivity", though it is important to keep in mind the specific definitions of syntactic functions given in Section 4.3 above. Naturally, this factor is not applicable to object anaphors.

13. *Corpus and speaker.* As the key aim of this study is to test for cross-linguistic variation in the drivers of referential choice, it is crucial to include the corpus language as facetting variable in any model, both as a random effect in regression models and as a general check on predictive accuracy in other kinds of models. However, we first need to account for potential structural differences between corpora, and since the texts are not controlled for content, we also need to account for inter-speaker variation, to ensure that variation between speakers does not exceed variation between corpora, or else we cannot properly assess cross-linguistic regularities. This is addressed in Section 5.4 below.

Lastly, it should be mentioned that there are a number of factors that are likely to offer valuable insights into the mechanisms of referential choice, but happen fall outside the scope of the present study for one reason or another. Of these, perhaps the two most notable are semantic number and semantic role (Section 2.4.2.2), for which no corpus annotations are available. These factors undoubtably offer avenues for future research, but unfortunately cannot be included in the present study.

Before moving on the analysis of the corpus data in the next chapters, we first need to discuss a number of additional methodological issues that are tied specifically to our selection of tested factors.

4.6 | Additional methodological issues

As mentioned before, a lack of clarity in methodology limits comparability between studies, and as such it is essential to provide precise definitions of all applied measures (within reason). This section deals with some of the more esoteric, but no less important, methodological considerations relevant to this study: How should discourse be segmented into quantifiable units (Section 4.6.1)? And how should antecedence be operationalized (Section 4.6.2)? Lastly, this section also comments on the distribution of first and second person mentions (vs. third person) in the corpus data (Section 4.6.3).

4.6.1 | Segmentation of discourse

One of the central methodological questions in investigations of discourse data is how to best partition discourse into segments, as well as how to best quantify these segments. The first part of this issue involves deciding on an appropriate unit of segmentation (Section 4.6.1.1), the second deciding which segments (given a particular unit of measurement) should be included and counted, and which should not (Section 4.6.1.2).

4.6.1.1 | Units of measurement

Finding appropriate units for the segmentation of discourse is a central question in this field of research; see, among others, the discussions in Taboada & Hadic Zabala (2008) and Moosegard Hanse (1998).[10] Which segmentation units are the most appropriate is of course dependent on their ultimate application, and like all aspects of methodology their selection needs to be guided by the aims of the study in question. This section briefly reviews some of the many options used in the literature, evaluating their advantages and disadvantages in the light of the needs of the present investigation.

A non-exhaustive lists of possible measurements, either used or proposed in the literature, might include the following units of discourse segmentation, among many others:

- elapsed time,
- phonemes and phonological words,
- intonation units;
- paragraphs and narrative episodes,
- sentences and clause units,
- noun phrases,
- orthographic or grammatical words,
- morphemes;
- intervening NPs or referring expressions,
- nodes of syntactic, rhetorical, etc. trees, or
- propositions, speech acts, etc.

10 Approaches such as Ford (2004), which argue for a situational rather than a general definition of discourse units – that is one that adapts to the particular characteristics of the examined context on a case-by-case basis – possess limited utility for quantitative analysis.

All of these segmentation units[11] come with their own drawbacks and limitations. Some require specific types of data: The notion of "paragraph", for instance, is only properly applicable to written texts, and there only to certain genres; divisions by elapsed time and measures dependent on phonology require the presence of adequately analyzable audio recordings. Some units are constrained by practical feasibility – given the corpus data and annotations we are working with – as they require specialized tools: Units based on hierarchical syntactic structures (i.e. as gleaned from a treebank, e.g. in Haghighi & Klein 2010), for instance, require automatic parsing that is only possible with languages for which sufficient training data is available, which rules out investigations of most less-studied languages; similar concerns hold for measures of rhetorical distance (Mann et al. 1992; see Section 2.2.5.3). Finally, other units are conceptually unsound or lack explanatory utility, either because they are too coarse-grained or else fail to adequately capture relevant conceptual units that we are interested in (e.g. content units and propositions, Moosegard Hanse 1998). As such, many of the options listed here can be dismissed out of hand for the present study, leaving only a number viable options: intonation units, connected textual units, individual word units, syntactic units, and referring expressions.

The first option are intonation units, this being segments of speech "uttered under a single, coherent intonation contour" (Du Bois et al. 1993: 47). As such, they are delineated more by considerations of prosody than of grammar; see the related notion of prosodic sentences in Chafe (1994: 139–140). As units of discourse segmentation, intonation units find use in Torres Cacoullos & Travis (2019: 658), among others, as well as in the RefLex annotation scheme Riester & Baumann (2014, 2017). While they are a technically viable option, they are not applicable to the corpus data in a straightforward manner that ensures cross-corpus comparability.

Second in the list are the more high-level structural divisions of discourse. Ariel (1990: 20) argues that the segmentation of discourse into larger, connected "textual units" serves as a suitable reflection of the scope of speakers' and listeners' working memory. As mentioned above, the division into 'proper'

11 For conversational data with multiple speakers, we might further add instances of turn-taking, such as turn-constructional units (TCUs; Sacks et al. 1974; Ford et al. 1996), and similar measures to the list.

paragraphs is only feasible with written data, as spoken texts lack partitioning on an equivalent level, or at least one that would be readily identifiable. For narrative texts, the closest approximation might be narrative episodes or scenes (cf. Anderson et al. 1983; Vonk et al. 1992), whose onsets in particular have been identified as influencing referential choices (Marslen-Wilson et al. 1982; Tomlin 1987a; Fox 1987a).[12] As a plain measure of anaphoric distance, however, they are less suitable, as they are generally of substantial length – certainly longer than most paragraphs in written narratives – and so would yield measurements of rather low resolution. As such, smaller, more self-contained segments are preferable for measurements of anaphoric distances and related factors.

Towards the other end of the granularity spectrum are word units, which notoriously lack anything so much as approaching a widely agreed-upon definition. Even so, words are inarguably the most commonly selected choice of unit in the literature on referential choice and anaphora resolution both (e.g. in Kibrik et al. 2013; Kibrik et al. 2016; Jarbou & Migdadi 2012; and many others), owing largely to the comparative ease of counting orthographic words in tokenized written texts – in English, this is as simple as counting any token between two spaces or punctuation marks. Simple word counts require no additional layers of annotation or analysis, are enhanced by combination with (automatic) part-of-speech taggers, and, as mentioned above, offer relatively high granularity. However, their use is problematic for a number of reasons, even putting aside the arbitrariness of orthographic word counts and their limited applicability to non-Latin scripts. Other problems arise from difficulties in defining word boundaries in a cross-linguistically applicable fashion; see, among others, relevant discussions of wordhood in Dixon & Aikhenvald (2002) as well as Schiering et al. (2010) on the typology of wordhood.

Measures with somewhat better comparability target divisions based on syntactic structures. Sentences are a popular unit of division, defined either

12 Divisions into larger stretches of discourse also find use in computational linguistics, where they typically revolve around textual content or, more generally, locutionary acts (Hearst 1994; Passonneau & Litman 1997). These units are then decomposed into smaller segments of varying length, as in Grosz & Sidner's (1986) discourse model. These approaches ultimately rest on the notion of speech acts as used in the philosophy of language (cf. e.g. Austin 1962; Searle 1969).

orthographically (i.e. delimited by punctuation marks) or syntactically; in the latter case, sentences are the smallest syntactically (and propositionally) independent unit of discourse, and can minimally be composed of one clause, or multiple independent and dependent clauses. However, Torres Cacoullos & Travis (2019: 658) caution that while "the sentence has featured prominently in syntactic analysis, it is not a straightforward unit of spoken language" (cf. also Harvie 1998: 24; Izre'el 2005: 3; Miller 1988: 132). A more rigorously defined reinterpretation of sentences are T-units ('minimum terminable units'), which are composed of a matrix clause plus modifiers including subordinated clauses. T-units originate from studies on text analysis (Hunt 1977), and have found widespread use in studies on language acquisition (e.g. Fox 1987b; Fries 1994).

An alternative to whole sentences are individual clause units, which might be defined in terms of predicates plus their accompanying phrases (cf. Thompson & Couper-Kuhlen 2005). Givón (1983a) notes that distance measurements in terms of clauses represent cognitive patterns reasonably well, as language is processed in clause-sized chunks rather than word-for-word (or sentence-by-sentence). In a similar vein, Thompson & Couper-Kuhlen (2005) identify the clause as the basic unit of interaction and the locus of interaction, as their predictable internal structure allows listeners to anticipate where the predicate occur and where the clause is going to end. The inherent predictability of clausal structure differs from language to language, they note, but regardless of how a language constructs clauses, the clause will always be the locus of interaction. In addition to the numerous studies on anaphora that utilize clauses as their unit of disourse segmentation (e.g. Arnold 2010; Arnold et al. 2009; Kibrik et al. 2016; etc.), clause units (specifically finite clauses) have also been proposed as the preferred unit of analysis in centering theory-based approaches Taboada & Hadic Zabala (2008).

Like intonation units and connected textual units, measures based on clauses or sentences also have the drawback of providing a lower resolution than other more fine-grained measures such as word units. Clauses can be longer or shorter, but are in all cases counted as one unit, which is of course even more of an issue when measuring in terms of entire sentences. For measuring anaphoric distances, this can result in inaccuracies at very low distances to the antecedent in particular. That being said, extremely-short distance anaphora (i.e. intra-clause) is likely to be subject to different constraints than cross-clause anaphora (Arnold 2010: 190); Falk 2006: 60–66, and hence best

excluded from studies that focus primarily on discourse-contextual factors; see Section 4.6.2.1 below. Although clause (and sentence) length is fairly variable, most clauses are in fact of roughly similar length (in the present sample, the cross-corpus mean length is $l = 4.95$ words, with a standard deviation of $\sigma = 0.851$ words), and as such, the lower granularity of clause-based measurements becomes less of an issue as distance increases.

Another viable measure for anaphoric distances in particular is counting the number of referring expressions (or more generally, any NP) that intervene between anaphor and antecedent. This unit of measurement is used in Biber et al. (1998), among others, and suggested in Ariel (1990) as a measure of competition between candidate antecedents (Section 2.4.2.4), which we also adopt for the present study, as noted above. However, it should be noted that since referring expressions tend to be more or less evenly distributed throughout discourse (cf. Levy & Jaeger 2007; Tily & Piantadosi 2009; Jaeger 2010), the number of intervening referential mentions is strongly correlated with anaphoric distance (Schiborr 2017: 33, fig. 2), mostly making the two measures mutually redundant.

Which units to choose for which factors, then? For our measures of anaphoric distances as well as all other factors that capture the structure of the surrounding discourse (i.e. those that involve limited lookback, e.g. recent coreferential mention frequency), we will here employ clauses as our unit of discourse segmentation, using the predicate-centered definition given above.[13] Given our focus on cross-clause anaphors (see Section 4.6.2.1 below), the minimal meaningful distance is a single clause, meaning higher resolutions (as offered, e.g. by elapsed time or words) are unnecessary. Of course, not all clauses are created equally, as they may differ in terms of their syntactic dependency and the finiteness of their predicates; we will discuss this point further below in Section 4.6.1.2. We here use (grammatical) word counts only for measuring the length of clauses (Section 6.11). In doing so, we will be largely glossing over the aforementioned issues with comparability, as the discrepancies between mean clause lengths in the ten corpus languages are fortunately fairly small overall ($P = 4.95$ words, $\sigma = 0.851$).[14]

13 A strong argument for going this route, of course, is that the corpus annotations already explicitly mark clause boundaries (see Section 3.3.1).

14 A potential way of addressing the comparability issue would be by normalizing clause

A last comment on a key methodological decision that was only implied in the previous paragraph, to wit, the decision to select only a single unit of segmentation for a given factor. In their models of referential choice, Kibrik et al. (2016) employ multiple distance measures in unison – words, intervening anaphors, sentences, paragraphs, as well as rhetorical distance – all of which are highly correlated, which is reflected in their results. Nevertheless, they find that the accuracy of their predictive model improves if two or more distance factors are combined (2016: 9). Given the focus on maximizing predictive accuracy in Kibrik et al. (2016), the inclusion of multiple, largely redundant distance measures is likely justified, as evidenced by their combined effect. However, for the more explanatory focus of the present (otherwise methodologically quite similar) study, I would argue that in the inevitable trade-off between accuracy and simplicity, we should favour the latter, since methodological parsimony and transparency should aid our ultimate goal of providing a readily interpretable account of the mechanisms involved.

In the next section, we will discuss the aforementioned question of which clauses to count (and how much), and in Section 4.6.2 we will address the related issue of defining antecedence in operationalizable terms.

4.6.1.2 | Clause types

With the selection of clause units as the basic unit of discourse segmentation comes an additional question – exactly which clauses should be counted? Specifically, in question are dependent and/or non-finite clauses, which have been excluded in a number of studies on referential choice, such as Toole (1996) and Arnold (2003) (see also the discussion in Thompson 1987) Relatedly, Taboada & Hadic Zabala (2008) conclude that for Centering-theory based approaches, the recommended unit of discourse segmentation is finite clauses specifically. However, finiteness distinctions are very difficult to generalize across languages, and can be tenuous to maintain even within a single language (e.g. in Sanzhi Dargwa, Forker 2020; see also Forker 2011, 2013 for a treatments of finiteness in Hinuq, a related Nakh-Daghestanian language).

lengths by the mean length of all clauses in each corpus or text. However, this would prevent the identification of specific length thresholds (e.g. one or two-word clauses) that could be potentially interesting.

For the purposes of this study, no distinctions are made between independent and dependent or finite and non-finite clauses at the sample selection stage, and so all anaphors in any context are captured, provided they meet the other selection criteria; in non-finite clauses, certain argument positions are commonly suppressed, in particular subjects, and are hence excluded due to not meeting the criterion of pragmatic choice. Clause dependency is instead included as a distinct variable in the analysis (Section 6.10). Since all clauses are treated equally, this also means that calculations of anaphoric distance and related measures do not take syntactic structure into account, so that matrix clauses do not count for more than their subordinations in terms of distance. This all-in approach also encompasses fragmentary and abandoned clauses and speech from secondary speakers: All material that is intelligible, even if it is not syntactically complete or well-formed, is assumed to be written to the discourse record, even if speakers ultimately do not comit to it; see the related sections on sample selection above.

4.6.2 | Selecting antecedents

Antecedence is generally defined in terms of co-reference, that is reference to the same discourse referent across multiple referring expressions in a text (but see discussion in McEnery 1995). This, however, leaves open the question of which co-referential mention in the preceding discourse should be considered the antecedent of a given anaphor for the purposes of calculating anaphoric distances and related measures. There are three main approaches to this issue in the literature, the first two of which having found use primarily in computational linguistics-based studies of anaphor resolution, and as such actually attempt to *identify* antecedents, not just establish anaphoric relationships. The first considers antecedence relationships to reach back to the first mention of the referent in discourse, ignoring all co-referential mentions inbetween (e.g. in Botley & McEnery 2001). This kind of measure of distance-to-introduction is not useful for our purposes, as it fails to capture any information about the structure of discourse at the point speakers make their referential choices. It also poses issues of comparability between texts, since the range of distance-to-introduction values in a given text varies greatly depending on its overall length, and would as such require some sort of normalization procedure; this

is especially relevant with the spoken corpus data used for this study, where the shorter texts are only a few minutes long, while the longest last for well over an hour. For the written news reports on which Botley & McEnery (2001) is based, which tend to have a more or less uniform length, this is less of an issue.

In the second approach, the antecedent is taken to be the most recent full NP mention of the referent, again ignoring all reduced mentions of that referent falling inbetween. According to this approach, only forms expressing more explicit (i.e. lexical) information about a referent (compared to a reduced form) allow speakers to identify the intended referent of a less explicit (i.e. pronominal) form; consider, for instance, the discussion in Botley (1999), and from a different perspective, Huddleston & Pullum (2002: 1457). While this line of reasoning has not been without its proponents, in particular in earlier studies on anaphora resolution, it is also not unproblematic. For one, pronouns are hardly devoid of information, even if they are not as informative as most lexical expressions (at least not in English, on which all of these studies are invariably based), and for another, as Givón (1983a) (cf. also Kehler 2002; 2004) argues, it is the local discourse cohesion gained from sequences of co-referential mentions that maintains the identifiability of the intended referent in the absence of more informative forms. If the intermediary segments in the sequence are disregarded, as they are with this approach, the effects of close anaphoric linkage cannot be properly captured. In fact, Botley & McEnery (2000b: 33) recognize this issue in evaluating their own methodology, asking that "where pronouns occur in long anaphoric chains, does this constitute a long distance between anaphor and antecedent [= the most recent full NP mention], or merely a short distance between one pronoun and its most recent coreferent token?"

The third and final approach, which we will employ for the present analysis, instead selects the most recent co-referential mentions as the antecedent, regardless of its form, including zero anaphors (but excluding agreement markers, forced zero, etc.). As mentioned above, this approach allows chains of mentions to be mapped more accurately, and the exact properties of each link in the chain to be determined. Antecedents are subject to same general selection criteria as anaphors themselves (Section 4.1), but can be any role (i.e. not just subjects and objects), any person (not just third), or clausal. The form of the antecedent as a factor influencing the choice of form of anaphors is examined in Section 6.8.

The next sections deal with more specific issues related to antecedent selection: the question of how to treat same-clause anaphors (Section 4.6.2.1) and how to deal with abstract and clausal antecedents (Section 4.6.2.2), as well as whether to make special provisions for anaphors in direct and reported speech (Section 4.6.2.3).

4.6.2.1 | Same-clause anaphors

Anaphoric references to antecedents in the same clause are excluded from the sample, as they are subject to different constraints on form than cross-clause anaphora (Arnold 2010: 190; Falk 2006: 60–66). These constraints are either syntactic or simply a matter of what Kibrik (2000: 78) calls "supercontiguity" (i.e. close adjacency) as in (52) and (53).

The reasoning behind the exclusion of cases of this kind is chiefly a matter of comparability with cross-clause anaphors, and as such, analytical simplicity. That is not so say these cases of close-proximity anaphora are not valid instances of discourse anaphora; however, we will here leave them for future study.

(52) English
My father, then he used to buy a lot of ferret.

	my	*father*	*then*	*he*	*used*	*to*	*buy*
	1SG.POSS	father	then	3SG.M	used	to	buy.INF
##	ln_pro.1:poss	np.h:dt_a	other	pro.h:a	lv_aux	lv	v:pred
	0000	0003		0003			

a	*lot*	*of*	*ferret*
a	lot	of	ferret
ln	ln	ln	np:p
			0042

'My father, back then he used to buy a lot of ferrets.'

[mc_english_kent01_0021]

(53) English

In her young days, somebody had took her out and…

	in	*her*	*young*	*day-s*	*somebody*
	in	3SG.F.OBL	young	day-PL	somebody
##	adp	ln_pro:poss 0245	ln	np:other	indef_other.h:a 0255

had	*took*	*her*	*out*	*and*	*...*
have.PST	take.PTCP.PST	3SG.F.OBL	out	and	
lv_aux	v:pred	pro:p 0245	rv	other	

'In her young days, somebody had taken her out and…'

[mc_english_kent02_0355]

Aside from adnominal possessives, the most common type of construction affected are dislocated topic constructions as in (52) and (54) below. This type of construction is especially common in the Tulil ($P = 0.19$), Mandarin ($P = 0.07$), and Vera'a ($P = 0.07$) corpora. In Mandarin, for instance, dislocated topics receive special marking with a topic particle such as *bā* or *ne* (termed "pause particles" in Li & Thompson 1981). In other languages, they are only distinguished via their position and optional co-occurrence with a co-referential element in the clause (as well as intonation, etc.), as in the example from English in (53). In languages where these constructions are not explicitly marked, they are only analyzed as a dislocated element if an overt co-referential phrase in the expected position is present; the default interpretation is instead as an argument of the predicate (where applicable). In languages where they are marked, as in Mandarin, a zero anaphor is assumed if no co-referential phrase is present in the clause. In these cases, the topic construction and any co-referential subject or object argument are treated as same-clause anaphors, and the latter is excluded from the analysis for the reasons stated above.

(54) Tulil

Me tərevənik o bidəmmang idə.

	me	*tərə*	*=vənik*	*o*	*b=*	*idə-m~mang*	*idə*
	and	thing	=PL.CL:DIM	TOP	ASP=	3N-RED~burn	3N
##	other	np:dt_s	=rn	other	=other	v:pred	pro:s
		0019					0019

'And the things, they were dying from heat.' [mc_tulil_alrm_0011]

4.6.2.2 | Abstract anaphora and clausal antecedents

Anaphoric references to clause-level expressions and the larger linguistic context, or more generally where the antecedent is not a well-delineated noun phrase, but a connected stretch of discourse as in (55) are difficult to chart in terms of a number of the factors tested in this study, chief among them anaphoric distance. Their inclusion would require special provisions to achieve results given the employed annotation schemes, which is unfortunately not feasible at this point, and so they have been excluded.

Navarretta (2011: 106) reports that 90% of the abstract anaphora in a corpus of spoken and written Danish occur within one clause distance from their clausal antecedent. They also find that abstract anaphora are predominantly demonstratives, and only rarely lexical NPs or personal pronouns. A similar pattern has been found for English, as personal pronouns often cannot refer to abstract entities when the antecedent is a clause (Webber 1988). See also Gundel et al. (2003), Hedberg et al. (2007), and Dipper et al. (2011) for further discussion of this issue.

(55) English

a. *And I paid them seven and sixpence for my lodgings.*

	and	*I*	*paid*	*them*	*seven*	*and*	*sixpence*
	and	1SG	pay.PST	3PL.OBL	seven	and	sixpence
##	other	pro.1:a	v:pred	pro.h:p2	np:p	rn	rn_np
		0000		0092			

for	*my*	*lodging-s*
for	1SG.POSS	lodging-PL
adp	ln_pro.1:poss	np:obl
	0000	0093

‘And I paid them seven and sixpence for my lodgings.’

b. *That left me half a crown.*

	that	*left*	*me*	*half*	*a*	*crown*
	DIST.SG	leave.PST	1SG.OBL	half	a	crown
##	dem_pro:a	v:pred	pro.1:p	ln	ln	np:p2
	0094		0000			

‘That left me half a crown.’ [mc_english_kent03_0087-0088]

4.6.2.3 | Direct and reported speech

Direct speech of the kind in (56) is quite common in the corpus data, accounting for over a quarter of all clauses (cross-corpus mean $P = 0.29$, $\sigma = 0.184$).

(56) *a.* *Now, he said, […]*

	now	*he*	*said*
	now	3SG.M	say.PST
##	other	pro.h:s_ds	v:pred
		0128	

b. when you put that plug in, […] tie it in with that bit of thong.

	when	*you*	*put*	*that*	*plug*	*in*
	when	2SG	put.PRS	DIST.SG	plug	in
##ds #ds_ac	adp	pro.2:a 0000	v:pred	ln_dem	np:p 0143	rv %

	tie	*it*	*in*	*with*	*that*	*bit*	*of*	*thong*
	tie.IMP	3SG.N	in	with	DIST.SG	bit	of	thong
0.2:a 0000	v:pred	pro:p 0143	rv	adp	ln_dem	np:obl 0145	rn	rn_np

[mc_english_kent03_0140-0141]

Brown & Yule (1983) contend that direct speech constitues a separate "verbal record" from the surrounding discourse; in this vein, Botley & McEnery (2001: 228–229) argue that the assessment of anaphors with antecedents in a different discourse frame such as direct speech leads to flawed results (vis-à-vis the default configuration), as accessibility states are liable to differ between frames. These anaphors are hence not part of a continuous textual reference, they argue, but "rather a situational one".

However, it is important to observe that these arguments are likely specific to the newspaper articles employed by Botley & McEnery (2001), where direct speech mostly consists of verbatim quotations from third-party speakers who are entirely separate from the journalist writing the article in question. In the narrative texts used for this study, speakers merely assume the role of a character in the narrative, in effect speaking "through" them. Since the entirety of the discourse is planned and produced by the same speaker, switching between the roles of narrator (and commentator) and the various characters in the narrative (in autobiographical texts, this can even include the speaker themselves at earlier points in time), direct speech could instead be argued to occur largely within same verbal record (or at least a closely linked one, within the scope of the narrative; but see Lichtenberk 1996 for further discussion), making the caveats in Botley & McEnery (2001) less relevant here. As such, no special exceptions for references to referents occurring within a direct speech context are made in the present investigation.

4.6.3 | Distribution of first and second person mentions

Next, a few comments on potential issues for cross-corpus comparability that result from the exclusion of first and second person mentions. The sample data include two closely related text types, as introduced in Section 3.2.2 above: traditional narratives (i.e. folktales, often with fantastical and supernatural elements) and personal narratives, with the majority of the latter being autobiographical.

One difference between the various text types in the corpus data, apart from content, is the relative frequency at which first and second person forms occur. Especially in autobiographical texts, autological references to the speaker are quite common; in the traditional narratives, they chiefly involve depictive reference to an imitated, imagined speaker. First and second person forms make up a somewhat larger share of the mentions in the personal narratives compared to the traditional folktales in the sample (cross-corpus mean $P = 0.35$ with a standard deviation of $\sigma = 0.120$ in autobiographical texts, vs. $P = 0.20$ with $\sigma = 0.094$ in traditional narratives).[15]

As a consequence of the qualitative differences between first and second person references in personal and traditional narratives, their distribution differs as well: In traditional narratives, first and second person mentions are generally restricted to direct speech contexts (see Section 4.6.2.3), while in autobiographical narratives they may occur anywhere and in any context. As only third-person mentions are included in this study, it is important to keep these differences in mind, in particular for those discourse factors that are sensitive to local mention densities, for instance as regards (the lengths of) anaphoric chains (see Section 4.5 for an overview).

15 Notably, content-dependent variation in this regard appears to be mostly restricted to subjects ($P = 0.31$, $\sigma = 0.132$ with a range of $P = 0.16$ to $P = 0.52$), as objects are predominantly in the third person ($P = 0.09$, $\sigma = 0.054$ with a range of $P = 0.02$ to $P = 0.20$). This echoes similar observations in Haig (2018b: 810–811), which finds that regardless of content, the proportion of first and second person P does not rise above about 20%.

5 | Rates of lexical expression

The first part of this study addresses the question of how frequent lexical anaphora is in natural discourse from a typological angle; the question of under which circumstances lexical expressions are selected by speakers is dealt with in Chapters 6–8.

Most studies on the relative proportions of various referring expressions focus on the distinction between overt and covert forms, as in the traditional classification of languages into pro-drop and non-pro-drop (Perlmutter 1971). More recent approaches consider languages in terms of a continuum of overtness: Bickel (2013: 710), adapting terminology from media theory (McLuhan 1964), refers to "hot" and "cold" languages, with the latter being characterized by comparatively sparse information density.

One such measure that has been proposed within typological research is the notion of "referential density" (RD; Bickel 2003, 2005; Noonan 2003; Stoll & Bickel 2009), which represents ratio of overtly expressed arguments to potentially available argument positions, that is the balance between the prominence of the structure of an event and the participants involved in it.[1]

1 Noonan calculates a number of additional variant ratios, one of which also considers agreement, and others that more directly capture the proportion of predicates to overt arguments. Bickel's RD is equivalent to Noonan's "RD1".

Through referential density, languages can be characterized as being either relatively noun-prominent (i.e. with focus on the participants of events, with high RD) or verb-prominent (i.e. with focus on the events themselves, with low RD). Referential density is calculated as the proportion of overtly expressed arguments to all possible argument slots. In other words, the lower the referential density in a given text, the higher the rate of zero anaphors. RD is a measure calculated on the basis of individual texts, but taken to characterize the discourse of entire languages, at least within a given mode and text type. The optionality of overt expressions is judged "from a strictly syntactic point of view" (Stoll & Bickel 2009: 543; a definition that is largely compatible to the one used here, see Section 4.2.2.2). A density value of $RD = 1$ indicates that every possible argument slot is filled with a pronominal or lexical expression, and a value of $RD = 0$ that there are no overt NP references at all. Naturally, languages are expected to fall somewhere inbetween these extreme values, and there are indeed substantial differences between languages in terms of referential density (Noonan 2003; Bickel 2003; cf. the null subject parameter, Holmberg 2009; Nicolis 2008).

Conversely, the overall rates of lexical expressions have only seldom been investigated cross-linguistically. One such approach is Stoll & Bickel (2009), who complement considerations referential density with a variant measure that targets the explicitness of discourse (i.e. rate of lexical anaphors) rather than capturing its overtness (i.e. rate of zero anaphors). This "lexical referential density" (RD_{lex}) is calculated analogously to RD, but as the ratio of lexical NPs to non-lexical NPs in a given text. Stoll & Bickel (2009: 553–554) report substantial cross-linguistic differences both in levels of RD_{lex} as well as the use of lexical expressions between two corpora of spoken Russian and Belhare (Sino-Tibetan, Nepal). They conclude that while all languages comply with limitations on the maximum number of lexical expressions permissible in a clause (cf. Section 2.2.5.1), there are "significant differences" in how languages operate below this limit, "specifically in the amount of lexical information that speakers are expected to give away when telling a story beyond what is strictly needed for tracking the identities of referents across events" (Stoll & Bickel 2009: 554), including stylistic habits of individuals and cultural narrative traditions.

That being said, the figures drawn from the corpora used in this study are principle not directly comparable to those reported in Stoll & Bickel (2009) and similar studies, as a fundamental aspect of these studies is the use of data

derived from retellings of the Pear film (Chafe 1980), which contains a highly limited number of referential entities that speakers may mention in their discourse. As discussed above, the Multi-CAST corpora are composed of free narratives that are virtually uncontrolled for content, and hence differ substantially from this parallax approach to corpus design. While Pear story retellings may themselves vary in structure and length (cf. Bickel 2003; Kumagai 2006), they and similar elicitation experiments crucially offer speakers the same number of opportunities to verbalize referents, which makes them eminently comparable across speakers of different languages – a trait that the corpora used here lack (see Schnell & Schiborr 2018 and Haig et al. 2021 for further discussion of these and related issues). Besides this, there are other, more general methodological caveats as well, as in the absence of suitable standards of comparison and precise descriptions of methodologies (i.e. precisely what is counted in which contexts and when?) it can often be difficult to judge what constitutes a substantial difference in lexicality profile of language, or whether variation falls well within range of random variation between data sets. While unqualified, direct comparison is thus not viable, we can nevertheless gauge general cross-linguistic tendencies by placing the values reported by earlier, Pear-story based studies and others side-by-side with those taken from Multi-CAST.

In this and the next chapters, we examine further evidence regarding the use of lexical NPs, starting in the following sections with the overall proportions of lexical and other referring expressions. But rather than quantify matters in terms of "densities" – that is, as a variable of information pressure – we here instead frame observations in more neutral terms, as a simple "lexicality rate" among all discourse anaphors. In the following, we first look at overall proportions of different form types – zero anaphors, pronominal and lexical NPs – in the ten Multi-CAST corpora (Section 5.1), yielding values that are mathematically (if perhaps not conceptually, see below) comparable to RD_{lex}. We then delve a little deeper, looking at the corresponding rates in various argument positions (Section 5.2) and put these figures in the context of referent introductions, which are known for being predominantly lexical (Section 5.3). Lastly, we check whether the cross-linguistic regularities identified in these sections are also stable between individual speakers and texts (Section 5.4).

5.1 | Overall lexicality rates

Figure 5.1 shows the relative proportions of three basic forms types (zero anaphors, pronominal NPs, and lexical NPs) across the ten corpora from the MultiCAST collection. These figures only capture mentions in argument positions (subjects, objects, and various obliques such as goals, instrumentals, etc.) that meet the selection criteria outlined in the previous chapter (third-person anaphoric mentions only, i.e. excluding introductions, etc.). As with (lexical) RD (Bickel 2003; Stoll & Bickel 2009), mentions in other, non-argument positions fall outside the purview of this study, but are presumably predominantly lexical.

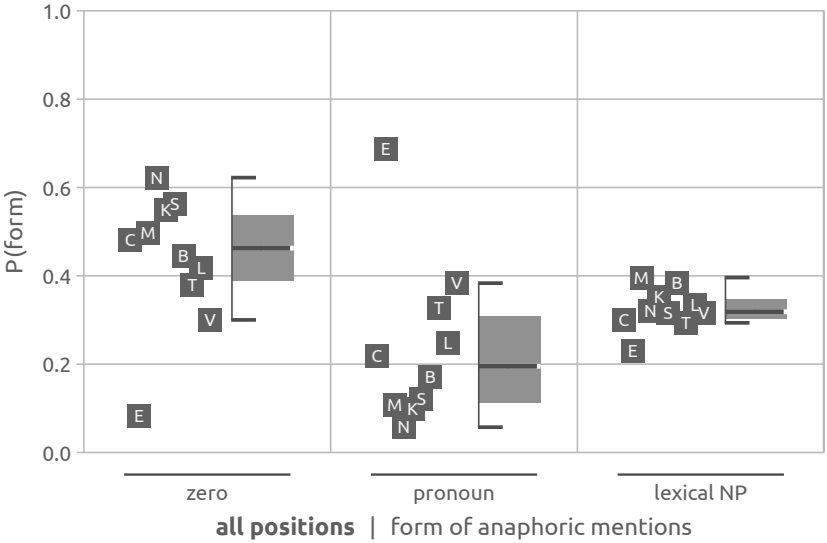
As Figure 5.1 shows, there is a quite noticeable degree of similarity between the lexicality rates in the ten corpora, especially compared to the corresponding distributions of zero anaphors and pronominal forms. Lexicality rates range from 23% lexical in English to 40% lexical in Mandarin, but mostly fall between 30% and 35% lexical:

(57) cross-corpus mean form rates of anaphors in argument positions

a.	zero	$m = 0.43;\ \sigma = 0.155$
b.	pronominal	$m = 0.24;\ \sigma = 0.188$
c.	lexical	$m = 0.32;\ \sigma = 0.047$

We here hence find a remarkable level of uniformity: All corpora in the sample tend towards a single baseline rate of lexicality. In other words, the information content carried by explicit referring expressions is roughly the same across essentially random stretches of natural, narrative discourse from different languages. This runs counter to Stoll & Bickel (2009), who find substantial differences in terms of lexical referential density between Russian and Belhare; since their analysis includes referent introductions, however, it is possible that a significant locus of cross-linguistic variation is found in how languages introduce new referents into discourse; see Section 5.3 for further discussion.

While the figures in Figure 5.1 do not capture the precise circumstances under which these expressions are deployed, it does suggest that the strategies for form selections are to a degree similar across languages (cf. claims to universality in Ariel 1990). Although properly comparable data are hard to



		zero		pronoun		lexical NP		
	corpus	N(form)	P(form)	N(form)	P(form)	N(form)	P(form)	N(all)
C	C. Greek	388	0.48	177	0.22	242	0.30	807
E	English	194	0.08	1600	0.69	534	0.23	2328
M	Mandarin	501	0.50	109	0.11	400	0.40	1010
N	Nafsan	608	0.62	56	0.06	313	0.32	977
K	N. Kurdish	590	0.55	105	0.10	378	0.35	1073
S	S. Dargwa	414	0.56	89	0.12	232	0.32	735
B	Tabasaran	524	0.44	202	0.17	452	0.38	1178
T	Teop	394	0.38	340	0.33	305	0.29	1039
L	Tulil	409	0.42	244	0.25	328	0.33	981
V	Vera'a	1017	0.30	1299	0.38	1072	0.32	3388
	totals	5039	—	4221	—	4256	—	13516

Figure 5.1 | Form of third-person anaphoric mentions in all argument positions across corpora from ten languages.

find as noted above – especially with regards to which argument positions and whether and how zero anaphors are counted – lexicality rates in corpora from other text types do seem to span a somewhat wider spread of values. Clancy (1980: 133) finds comparatively lower values, with lexical expressions making up 16% and 27% of all anaphoric references in all positions in Pear story retellings from English and Japanese. For their Pear story retellings, Stoll & Bickel (2009) report mean lexical referential densities of $RD_{lex} = 0.47$ ($\sigma = 0.07$) in Russian and $RD_{lex} = 0.32$ ($\sigma = 0.08$) in Belhare (Sino-Tibetan, Nepal), a difference of about 15 percentage points, which falls well within the range of values found here, though their figures do include referent introductions and would likely be lower if they were excluded.

The cross-linguistic stability of lexicality rates in the Multi-CAST corpora also highlights the symmetry of the correspoding zero–pronoun choice across languages. Since lexicality rates are practically stable across languages, the rates of zero and pronouns are in essence mirror images of one another; this means that a language's preference for either zero or pronominal subjects does not directly affect its rate of lexical subject expression (cf. Stoll & Bickel 2009: 544). Zero anaphors and pronouns are hence also where most variation between languages in terms of referential choices is found (cf. pro-drop, null subject parameter, etc.) – consider the characteristically low rate of zero anaphors and correspondingly high rate of pronouns in English ($P = 0.08$ vs. $P = 0.69$), and the inverse pattern in languages such as Nafsan, Sanzhi Dargwa, and Northern Kurdish, the last of which has highest rate of zero in the sample ($P = 0.55$ vs. $P = 0.10$). Other corpora, like Vera'a and Teop, fall in the middle, with roughly equal proportions of the three form types (e.g. $P = 0.30$ zero vs. $P = 0.38$ pronouns vs. $P = 0.32$ lexical in Vera'a).

The stability of the lexical baseline is especially remarkable in light of the essentially uncontrolled nature of the corpus in terms of content. We would expect overall rates of lexical expression to be highly content-dependent, and hence differ substantially between texts and corpora; hence the focus on parallax corpora such as Pear story retellings in the literature. But as we will see later in Section 5.4, lexicality rates are stable not just across corpora, but also across individual speakers and texts.

The figures in Figure 5.1 group anaphors in all positions together, but as we will see in next section, different positions have quite different lexicality profiles, not all of which are as cross-linguistically stable as the overall picture suggests.

5.2 | Role-based lexicality rates

5.2.1 | Lexicality of subjects

Figure 5.2 shows same perspective as Figure 5.1 above, but limited to anaphoric, third-person subjects. Here, the variation in lexicality rates between corpora is even smaller, ranging from 12% lexical in English to 29% lexical in Mandarin, with most other corpora clustering very tightly around the 20% lexical mark:

(58) cross-corpus mean rates of form of subject anaphors

a. zero $m = 0.56;\ \sigma = 0.207$

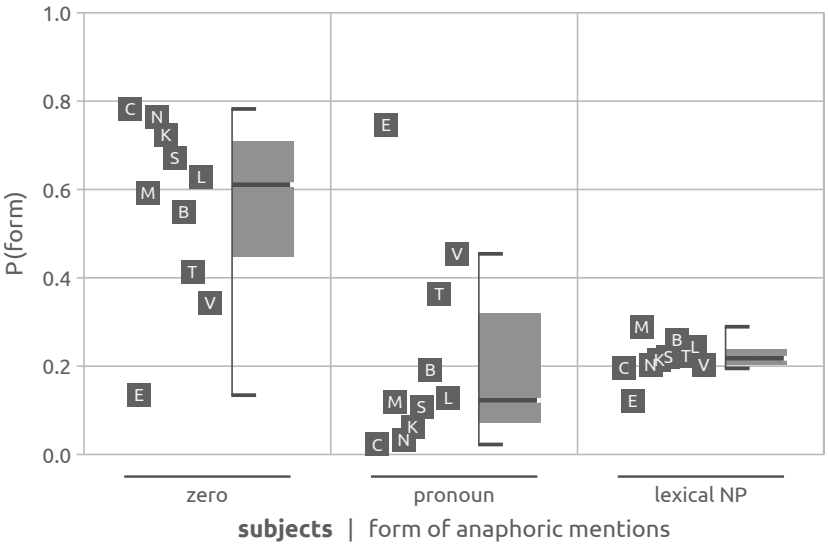
b. pronominal $m = 0.22;\ \sigma = 0.232$

c. lexical $m = 0.22;\ \sigma = 0.045$

The rates for zero and pronouns are consequently even more diverse for subjects specifically than across positions. The subject role is generally associated with non-lexical forms, as it is most topical argument position, and topicality is in turn associated with reduced forms (Chafe 1994; Givón 1983a; Mithun 1991; etc.).

But Figure 5.2 shows that there is in fact a cross-linguistically robust baseline of lexical subjects in natural discourse. Subject mentions are very frequent, making up well over half of the full sample (mean $P = 0.66$, $\sigma = 0.066$), and so are a major contributor to the overall lexicality rates discussed in the previous section. Many of the observations made there hence also apply here.

As in the overall rates, the English corpus has a somewhat lower rate of lexical subjects, the Mandarin corpus a slightly higher, but the spread of values inbetween is noticeably narrower. This makes English the language with the lowest rate of highly explicit subjects, and Mandarin the one with the most; English in particular is an outlier, possibly as a result of a combination of differences in text type and relatively informative pronouns. Notably, Mandarin is generally considered to be a language that is heavily reliant on inference from contextual cues (Huang 2000a: 261–277) rather than explicit marking, but patterns in this way seemingly not due to a high rate of zero subjects, as other corpora with higher zero rates come closer to the central tendency in terms of lexicality.



		zero		pronoun		lexical NP		
	corpus	N(form)	P(form)	N(form)	P(form)	N(form)	P(form)	N(all)
C	C. Greek	345	0.78	10	0.02	86	0.20	441
E	English	182	0.13	1010	0.75	163	0.12	1355
M	Mandarin	426	0.59	85	0.12	208	0.29	719
N	Nafsan	533	0.76	23	0.03	142	0.20	698
K	N. Kurdish	465	0.72	40	0.06	138	0.21	643
S	S. Dargwa	352	0.67	56	0.11	116	0.22	524
B	Tabasaran	432	0.55	150	0.19	204	0.26	786
T	Teop	311	0.41	273	0.36	167	0.22	751
L	Tulil	390	0.63	79	0.13	151	0.24	620
V	Vera'a	834	0.34	1104	0.45	492	0.20	2430
	totals	4270	—	2830	—	1867	—	8967

Figure 5.2 | Form of third-person subject anaphors across corpora from ten languages.

By and large, however, subject lexicality (in the third person) appears to be largely impervious to cross-linguistic differences in subject overtness, and also indifferent towards narrative content. As comparison with figures from the literature indicates, however, it is highly sensitive to differences in text type. While compatible figures for subjects are a little easier to find than for overall rates including objects and oblique arguments, as defined above, it is nevertheless important to keep in mind that direct comparability with the values in Figure 5.2 is not possible, as it is not clear in all cases whether reported figures include zero anaphors, and first and second person forms and others for which speakers do not exercise a pragmatic choice in the totals.

Payne (1993: tab. 26) reports that 19% of subjects in traditional narratives from Yagua (Peba-Yaguan, Peru) are lexical. Haig & Schnell (2016) calculate corresponding figures for two narratives from Gorani (Iranian, Iraq and Iran) published in Mahmoudveysi et al. (2012: 89–103), and find a subject lexicality rate of about 20%. Givón (1990) reports 26% lexical subjects in spoken English narratives, Arnold (2003: fig. 2a–c) 31% in Mapundungun (Araucanian, Chile) narratives, and Lichtenberk (1996: tab. 8) 34% in To'aba'ita (Oceanic, Solomon Islands). Subjects in the personal narratives from Venezuelan Spanish and French analyzed in Ashby & Bentivoglio (1993: tab. 3) are 16% and 21% lexical, respectively. Various studies of Pear story retellings report values between 27% lexical in German (Himmelmann 1997) and 31% in English (Kumagai 2006: tab. 8).

Corresponding figures for conversational data are sparser; Kärkkäinen (1996: Tab. 2 and 5) reports a low 9% lexical subjects in informal English conversation from the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2005), and Francis et al. (1999: tab. 1) similarly find a rate of 9% lexical in a subset of the Switchboard corpus of English telephone conversations (Godfrey et al. 1992). Notably, written texts appear to have substantially higher rates of subject lexicality, with about 60% in the ZPG fund-raising letter (English, Prince 1992). Genre differences are likely to have a significant effect here; this is a likely avenue for future investigations, given the scarcity of currently available data. Overall, subject lexicality rates in monologic narratives texts such as the ones analyzed here seem to fall between the much lower rates in conversational data and the (presumably) much higher rates in written texts.

Lastly, lexicality rates also differ between the subjects of transitive and intransitive clauses (cf. the non-lexical-A-constraint in Du Bois 1987b, see Sec-

tion 2.2.5.1), in the Multi-CAST data by about 12% (mean difference across corpora). Notably, even with this difference, lexicality rates for subjects are stable across corpora, as the rate of transitive (vs. intransitive clauses) is similarly stable; these patterns are examined further in Section 6.12 in the next chapter.

5.2.2 | Lexicality of objects and oblique arguments

Compared to the remarkably stable pattern for subjects anaphors, the corresponding pictures for direct objects and oblique arguments (i.e. goals, addressees, instrumentals, etc.) in Figures 5.3 and 5.4 show substantially greater degrees of cross-linguistic variation as regards lexicality rates. For objects, lexicality rates fall between a minimum of 32% in English and a maximum of 61% in Tabasaran, closely followed by Northern Kurdish and Mandarin.

(59) cross-corpus mean rates of form of object anaphors

a.	zero	$m = 0.27;\ \sigma = 0.140$
b.	pronominal	$m = 0.24;\ \sigma = 0.221$
c.	lexical	$m = 0.49;\ \sigma = 0.101$

As rates for objects span a wider range of values, the zero–pronoun distinction is accordingly not an as clearly-cut mirror image. A number of languages (Northern Kurdish, Mandarin, Tabasaran, Nafsan, and Sanzhi Dargwa) essentially alternate between zero and lexical anaphoric objects, with very low rates of pronominal objects (cf. "avoid pronominal P", Haig et al. 2011b: 74–80). English and Tulil are the diametric opposite: Here, zero objects are rare, occurring in English only in certain constructions (e.g. VP echo deletion), so that a large majority of objects are pronominal. Notably, the corpora with the highest rates of pronominal objects (English, Tulil, Cypriot Greek) are also the ones with the lowest rates of lexical objects (and zero objects); conversely, the corpora with the most lexical objects (Mandarin, Nafsan, Tabasaran) have lowest rate of pronominal objects. That is, unlike with subject anaphors, the preference for either zero anaphors or pronouns does affect rates of lexical expression. If forms with intermediate informativity (i.e. pronouns) are dispreferred, then lexical expressions are used instead (if zero is not an option),

but conversely, if pronouns are the preferred reduced form for objects, then there is less reliance on lexical expressions.

As Figure 5.4 shows, obliques arguments are generally not zero, meaning speakers do not rely on contextual implicatures for non-core arguments, even if they are still implied by predicate. As such, form selection for these arguments is mostly a matter of choosing between pronominal and lexical forms.

(60) cross-corpus mean rates of form of oblique anaphors

a. zero $m = 0.05$; $\sigma = 0.055$

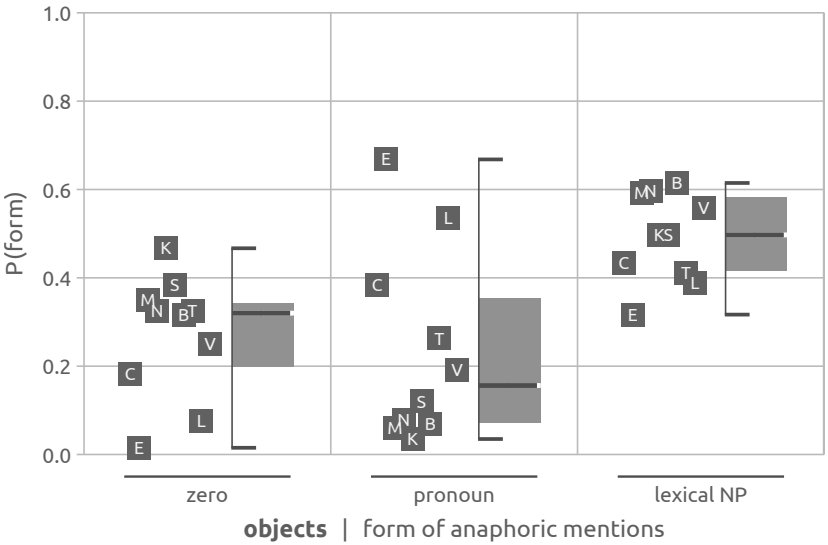
b. pronominal $m = 0.27$; $\sigma = 0.159$

c. lexical $m = 0.68$; $\sigma = 0.153$

We here find substantial variation between corpora, which are at least in part due to relatively low subsample sizes, especially in some corpora; obliques have less than half the data points of objects overall. Nevertheless, lexical forms preferred for obliques in all but one corpus (Cypriot Greek). In the Teop corpus, all oblique arguments are realized lexically; this is possibly an artefact of small number of cases, or else a quirk of the annotations (e.g. some annotators may be reluctant to annotate zero obliques).

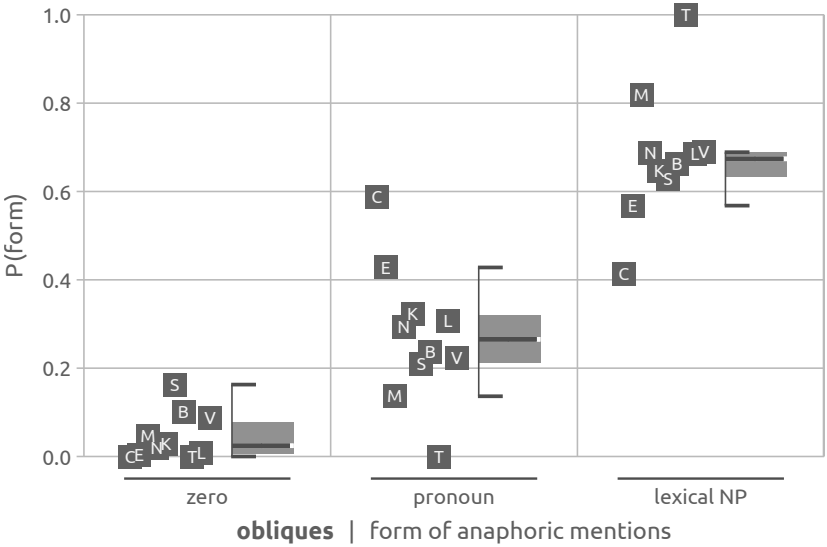
In sum, while objects and non-core arguments are generally associated most strongly with explicitness and indefiniteness (Comrie 1979), we here, perhaps surprisingly, also find that the actual rates of lexical expression in these roles vary considerably between corpora. This also means that much of the variation in the overall lexicality rates seen above in Figure 5.1 is due to variation among objects and obliques, as the rates for subjects are essentially the same across corpora. This raises the question of whether the form of object and oblique anaphors is in fact determined on the basis of same principles that determine the form of subjects (as per the claims of Ariel 1990).

As it stands, since lexical subject rates pattern the most similarly across corpora, they offer the greatest potential for useful typological generalizations, and will hence be the primary object of study. While we will address the question of role differences in the second part of this study (Chapters 6 and 7) by examining the factors influencing the selection of lexical expressions in both subject and object position, but the focus will be on the former. Since the lexicality profiles appear to be rather different between the roles – both in terms of magnitude and cross-corpus variability – it is best to examine them



		zero		pronoun		lexical NP		
	corpus	N(form)	P(form)	N(form)	P(form)	N(form)	P(form)	N(all)
C	C. Greek	43	0.18	90	0.38	102	0.43	235
E	English	11	0.02	483	0.67	229	0.32	723
M	Mandarin	71	0.35	12	0.06	120	0.59	203
N	Nafsan	74	0.32	18	0.08	136	0.60	228
K	N. Kurdish	120	0.47	9	0.04	128	0.50	257
S	S. Dargwa	48	0.38	15	0.12	62	0.50	125
B	Tabasaran	77	0.32	17	0.07	150	0.61	244
T	Teop	83	0.33	67	0.26	105	0.41	255
L	Tulil	18	0.08	127	0.54	92	0.39	237
V	Vera'a	153	0.25	118	0.19	343	0.56	614
	totals	698	—	956	—	1467	—	3121

Figure 5.3 | Form of third-person object anaphors across corpora from ten languages.



		zero		pronoun		lexical NP		
	corpus	N(form)	P(form)	N(form)	P(form)	N(form)	P(form)	N(all)
C	C. Greek	0	0.00	77	0.59	54	0.41	131
E	English	1	0.00	107	0.43	142	0.57	250
M	Mandarin	4	0.05	12	0.14	72	0.82	88
N	Nafsan	1	0.02	15	0.29	35	0.69	51
K	N. Kurdish	5	0.03	56	0.32	112	0.65	173
S	S. Dargwa	14	0.16	18	0.21	54	0.63	86
B	Tabasaran	15	0.10	35	0.24	98	0.66	148
T	Teop	0	0.00	0	0.00	33	1.00	33
L	Tulil	1	0.01	38	0.31	85	0.69	124
V	Vera'a	30	0.09	77	0.22	237	0.69	344
	totals	71	—	435	—	922	—	1428

Figure 5.4 | Form of third-person oblique anaphors across corpora from ten languages.

independently, as separate issues, rather than to consider role as a grouping variable. Oblique arguments and other positions are also of interest, but are much rarer than subjects and objects both grouped together and individually, and hence do not provide enough data for a robust analysis.

Before we move on to this part of the analysis, there are two more topics to touch on in this chapter: lexicality rates in referent introductions (as opposed to anaphoric mentions, Section 5.3), and the question of intra-corpus variation (Section 5.4).

5.3 | Referent introductions

While this study is focused on anaphoric mentions specifically, it would be remiss in not at least briefly touching on the referring expressions used for introductions of new referents. The lexicality rates among introductions also informs certain factors that will be tested later, such as sequence of mention (Section 6.9).

Shown in Figure 5.5 are the relative proportions of zero anaphors, pronouns, and lexical NPs among newly introduced referents, that is the first mention of each discourse referent in text (i.e. first instance of each referent index in linear order), including those inferable from frame semantics (cf. Fillmore 1982; Section 3.3.2); "new" is used here in terms of Chafe (1976), that is an entity has not been previously mentioned in the preceding discourse, irrespective of how familiar to the addressee it may be. Across corpora, new referents are almost categorically introduced via full, lexical NPs:

(61) cross-corpus mean rates of form of referent introductions

a.	zero	$m = 0.05;\ \sigma = 0.127$
b.	pronominal	$m = 0.10;\ \sigma = 0.316$
c.	lexical	$m = 0.85;\ \sigma = 0.324$

Most zero and pronominal introductions are of referents that are inferable (i.e. bridging anaphors), but a number of them, especially in Sanzhi Dargwa and Tabasaran, also occur in subordinate clauses, mostly relative clauses, that immediately precede first overt mention of the referent in question. Overall, bridging anaphors show more variation in forms than "brand new" referents,

which are essentially exclusively lexical (cross-corpus mean $P = 0.77$ lexical, $\sigma = 0.075$ vs. $P = 0.95$, $\sigma = 0.028$).

Notably, introductions make up only 17% ($\sigma = 5\%$) of all referring expressions (across roles) in the ten corpora – as Du Bois (1987b: 830) notes, if "a mention is new information, this typically entails that it will be realized with a full NP, but the converse is far from always true." This includes many referents that are mentioned only once in a text; their removal only marginally changes the proportions in Figure 5.5.

Ariel (1990) asserts that considerations of accessibility apply equally to both to introductions of discourse-new referents as well as to anaphoric mentions (see Section 2.2.4). While they in principle convey referential information of fundamentally different status, form selections for both should hence involve the same mechanisms. As such, those anaphors that are most like introductions (i.e. very low accessibility) should pattern similarly to Figure 5.5, with very high lexicality rates.

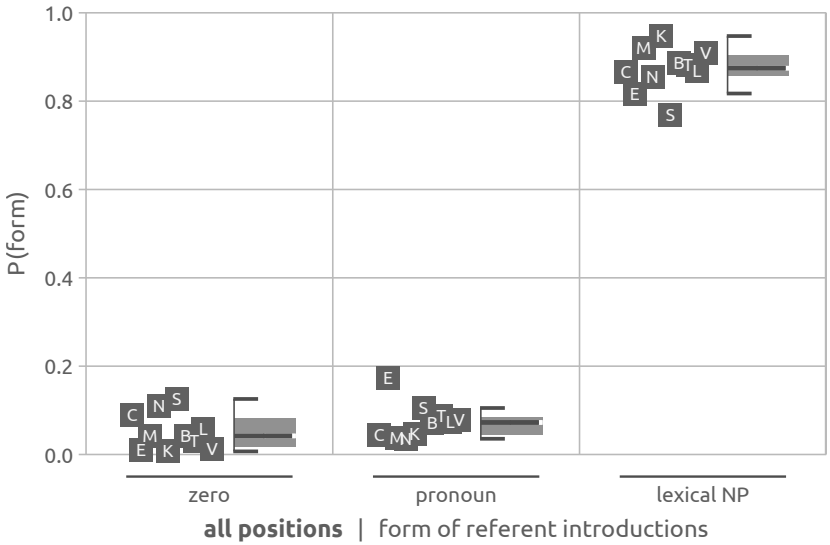
For further discussion of the processes involved in the introduction of new referents into discourse; see Haig et al. (2020) and Schnell et al. (2020, 2021b), among others.

5.4 | Intra-corpus variation

So far in this chapter, we have compared overall lexicality rates across corpora, and have found an appreciable degree of cross-corpus homogeneity. However, we also need to account for potential variation within each of the ten corpora in the sample, that is between speakers and individual texts.

One way of visualizing variability is by calculating for each corpus, speaker, and text in the sample the relative difference to the mean rate of lexical expression of the next higher scope, that is the entire sample, corpus, or speaker.[2] Figure 5.6 shows the resulting distribution: On the left, it shows the difference from the sample mean lexicality rate of each corpus in the sample; in the middle, the difference from the corpus mean of each speaker in that corpus; and on the right, the difference from the speaker mean of each text

2 The Cypriot Greek and Northern Kurdish corpora each contain data from a single speaker.



		zero		pronoun		lexical NP		
	corpus	N(form)	P(form)	N(form)	P(form)	N(form)	P(form)	N(all)
C	C. Greek	26	0.09	13	0.04	250	0.87	289
E	English	19	0.01	324	0.17	1534	0.82	1877
M	Mandarin	19	0.04	17	0.04	416	0.92	452
N	Nafsan	28	0.11	9	0.04	216	0.85	253
K	N. Kurdish	2	0.01	13	0.05	270	0.95	285
S	S. Dargwa	43	0.13	36	0.11	263	0.77	342
B	Tabasaran	16	0.04	27	0.07	338	0.89	381
T	Teop	8	0.03	23	0.09	230	0.88	261
L	Tulil	21	0.06	27	0.07	316	0.87	364
V	Vera'a	8	0.01	49	0.08	570	0.91	627
	totals	190	—	538	—	4403	—	5131

Figure 5.5 | Form of referent introductions across corpora from ten languages.

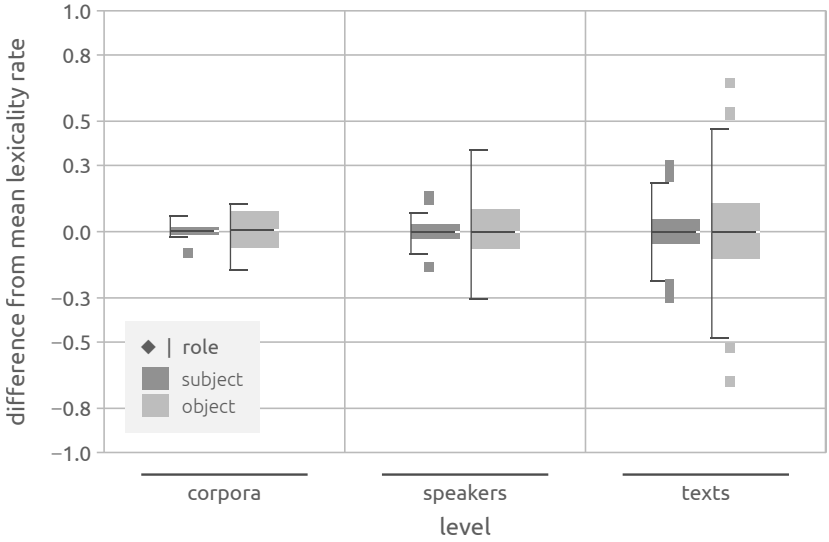


Figure 5.6 | Intra-corpus variation in the data, visualized as the difference of the proportion of lexical expressions at each level from the mean lexicality rate of the next higher level; for corpora, this is the entire collection. The small squares indicate outliers from the central distribution.

produced by that speaker. The values are split by mentions in subject and object position.[3]

While the spread of the distributions does increase from corpora to speakers and from speakers to individual texts, the differences between them are not statistically significant:

(62) subjects

a. inter-speaker variation vs. cross-corpus mean rate:
Levene's test: $p < 0.637$
ANOVA: $F(9, 27) = 1.25,\ p < 0.307$

3 Mentions in oblique positions are excluded here for the reasons given above.

b. inter-textual variation vs. cross-corpus mean rate:
Levene's test: $p < 0.721$
ANOVA: $F(9, 44) = 1.16,\ p < 0.344$

(63) objects

a. inter-speaker variation vs. cross-corpus mean rate:
Levene's test: $p < 0.769$
ANOVA: $F(9, 27) = 0.75,\ p < 0.658$

b. inter-textual variation vs. cross-corpus mean rate:
Levene's test: $p < 0.365$
ANOVA: $F(9, 44) = 1.07,\ p < 0.400$

Levene's test of homoscedasticity (i.e. equality of variances) accepts the null hypothesis (at $p > 0.05$) that the population variances are equal in all cases, which is a requirement for the ANOVA fits to the data, which in turn show that the differences between texts and speakers in each corpus (vis-à-vis the overall baseline rate) are not statistically significant (i.e. low *F*-value, high *p*-value). This is true for both mentions in subject and object position. Overall, this means that the lexical baseline rate of discourse is stable not only across corpora, but also across speakers and individual stretches of discourse. But as Figure 5.6 suggests, it is less stable for object mentions than for subject mentions. Note also that the variation at each level is approximately normally distributed, with values for individual speakers and texts clustering around a central tendency, and few extreme outliers. That being said, individual texts can still vary quite a bit in terms of lexicality rate. The divergence from the central tendency is slightly negatively correlated with text length ($r = -0.12$ for subjects, $r = -0.15$ for objects),[4] meaning shorter texts are somewhat more likely to deviate from the baseline rate than longer texts. This suggests that the cross-linguistically stable baseline only emerges across longer stretches of discourse or larger quantities of texts.[5]

4 Specifically, the correlation between the absolute difference between each text's lexicality rate and the cross-corpus mean rate, and the number of sampled subjects or objects in each text.

5 Which in turn suggests that the differences in lexicality rates observed between languages on the basis of Pear story data (e.g. in Stoll & Bickel 2009 and elsewhere) could possibly

Naturally, the overall lexicality rate of a text is only one metric affected by intra-speaker variation, and we should expect variability when carving up the data further along the various factors tested in this study, as is done in the next two chapters. For this reason, we will attempt to keep an eye on intra-corpus variation by including the identity of the speaker as a random effect in the regression models alongside the corpus, and by charting its relative magnitude from factor to factor.

To summarize, the figures presented in this chapter have shown that rates of lexical expressions are surprisingly stable not just across the ten corpora examined here, both overall and in subject in position in particular, but also across individual speakers and texts. That is, the variation in lexicality rate between speakers and texts within a single corpus is not much greater than the variation between corpora from different languages. This suggests that the overall explicitness of narrative discourse tends towards an ideal baseline level across languages. What these data do not show is whether the selection of lexical forms is also determined by same factors across languages (cf. accessibility theory, etc.), which is a question we will address over the next two chapters.

be an artefact of comparatively short text lengths.

6 | Examining individual factors

Where the previous chapter has noted remarkable cross-linguistic regularities in the rates of lexical expressions, especially for subject anaphors, this and the next chapter present data on the the circumstances under which lexical expressions are selected in the first place, in part in an effort to find explanations for the stable lexical baseline pattern.

We have discussed various approaches to the question of discourse anaphora – which, it bears repeating, have rarely focused on lexical anaphora in particular – in Chapter 2 above, and on their basis outlined a number of largely discourse-based factors to test as motivators for referential choices, which were already listed in Section 4.5 above. Since it can be helpful to have awareness of the individual effects contributing to the final outcome when interpreting relatively complex models, we will first examine the effects of the various factors on the selection of lexical expression one by one in this chapter, before moving on to the multifactorial models in Chapter 7, which forms the core of this part of the study. The specific factors to be tested for anaphoric mentions in subject and object position are repeated here for convenience:

1. humanness (Section 6.1),
2. total mention frequency (i.e. protagonisthood, Section 6.2),
3. anaphoric distance (Section 6.3),
4. number of recent co-referential mentions (Section 6.4),

5. number of recent mentions of related referents (Section 6.5),
6. number of recent mentions of competing referents (Section 6.6),
7. role of the antecedent (Section 6.7),
8. form of the antecedent (Section 6.8),
9. sequence of mention (Section 6.9),
10. clause type (Section 6.10),
11. clause length (Section 6.11), and
12. transitivity (for subject anaphors only, Section 6.12).

We will also be looking at a number of interactions between certain combinations of factors. The variation in the overall distribution of lexical expression between the ten corpora has already discussed in previous chapter.

6.1 | Animacy and humanness

6.1.1 | Definition and methodological issues

As mentioned earlier, animacy spans a fairly nuanced continuum, with human discourse participants at one end and most inanimate objects on the other (Silverstein 1976; Comrie 1989; Corbett 1991; Dixon 1994; Lockwood & Macaulay 2012), and the distinction between human and all non-human entities being by far the most fundamental (cf. the general animacy scale, Yamamoto 1999). For this reason, we will focus in the following on examining the binary distinction between

a. human, and
b. non-human referents.

In principle, a more nuanced analysis of animacy classes would be possible, as the corpus annotations distinguish a range of ontological animacy categories beyond humanness, in particular animate and inanimate non-humans; refer to the methods section on coding animacy below.

The human-centric nature of most texts, however, results in the overwhelming majority of eligible mentions in the sample being human ($m = 0.72$; $\sigma = 0.12$), while the frequency of non-human (and non-anthropomorphized, see below) animates is conversely much lower, and further dependent on the content of a text, making it highly variable. This is especially true for mentions in subject and object position, as a substantial proportion ($m = 0.34$;

$\sigma = 0.141$) of mentions of animal referents occur peripherally to events, for instance as goals (e.g. *she got on the horse*), instruments, or other oblique arguments, rather than in core argument roles. As such, the frequency of this type of referent in the sample differs greatly between individual texts and corpora, ranging from relatively common certain corpora ($P = 0.27$ in English, $P = 0.09$ in Nafsan) to nigh zero in others ($P = 0.01$ in Mandarin, Cypriot Greek, and Sanzhi Dargwa).

6.1.1.1 | Anthropomorphized entities

Since a major portion of the corpus data consists of folkloristic narratives, anthropomorphized entities such as talking animals and other supernatural entities like spirits and monsters are quite commonplace, though as with non-human animates discussed above, their presence is highly content-dependent, and as such they are unequally distributed across corpora and texts ($m = 0.08$; $\sigma = 0.12$). The example in (64) includes such a referent:

(64) Sanzhi Dargwa
K:urt:a dul at kumek birq'anda bik'ulcab.

	k:urt:a		*du-l*	*at*	*kumek*
	fox		1SG-ERG	2SG.DAT	help
#cv	np.d:s_ds	#ds	pro.1:a_cps	pro.2:obl	other:lvc
	0015		0015	0000	

b-irq'-an-da	*b-ik'-ul*	*ca-b*
N-do.IPFV-PTCP-1	N-say.IPFV-ICVB	be-N
v:pred %	v:pred	rv_aux

'The fox says, I will help you.' [mc_sanzhi_patima_0022]

Anthropomorphized entities constitute a liminal case between human and non-human referents. They generally possess capability for rational thought, planned action, and human speech, and often engage in certain aspects of human culture such as use of tools, houses, vehicles, and so on, sometimes in addition to the capabilities and behaviours usually associated with their non-human forms (e.g. flight). Narratively, they act as stand-ins for human characters in all but appearance. Whether speakers give precedence to their

human or non-human characteristics is likely dependent on the specifics of the narrative.

For the purposes of this study, anthropomorphized entities have been classified as human if they exhibit all or most of the aforementioned properties of human behaviour. This distinguishes them from clearly non-human entities (e.g. the magic talking horse vs. the hero's mundane mount).

6.1.1.2 | Body parts

Another boundary case concerns references to parts of human bodies – limbs, heads, internal organs, and so on, which here have been invariably coded as inanimate.

(65) Veraʹa
Dim ʼañ sarav ēn nōgōn e Wōwut ʼesegēn.

	di	*=m*	*ʼañ*	*sarav*	*ēn*	*nōgō-n*
	3SG	=TAM1	move.hand	rub	ART	face-CS
##	pro.h:a	=lv	v:pred	rv_v	ln	np:p

e	*Wōwut*	*esegēn*
PERS.ART	Wōwut	MAN.DEM2
rn	rn_pn_np.h:poss	dem_other:other

‘She reached for and rubbed Wowōt’s face like this.’

[mc_veraa_iswm_0342]

A point could be made for differentiating body parts that are (still) attached to a living body from those that are not, at least in terms of the inferability of their host and vice versa – see the discussion on bridging relations in Section 2.4.2.6 above and Section 6.5 further below – but the low overall frequency of this type of referent ($m = 0.02$; $\sigma = 0.02$) makes doing so unfeasible in practice. Deceased human bodies are always classified as inanimate unless magically reanimated.

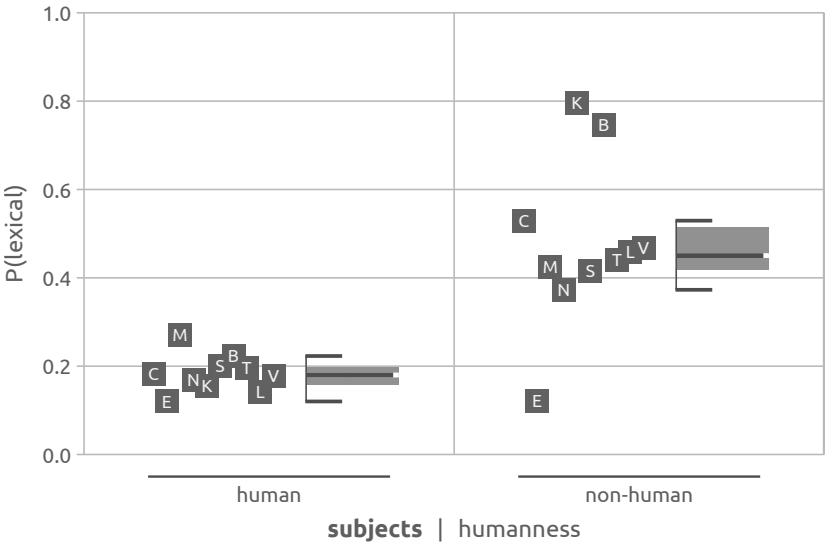
6.1.2 | Lexicality by humanness

This section explores the effect of humanness on the selection of lexical expressions, independently of the other factors examined in this study. The examination of factors in isolation informs the complex, multifactorial analysis that constitutes the core of this study (Chapter 7) by helping identify which associations to look out for. Where relevant, we will also examine selected interactions between two or more factors.

6.1.2.1 | Subjects

Figure 6.1 shows the proportion of lexical subjects among human and non-human referents across the ten corpora in the sample.[1] While human referents make up the vast majority of mentions in subject position (cross-corpus mean $P = 0.85$, with a standard deviation of $\sigma = 0.109$), human lexical subjects are quite rare overall ($P = 0.18$ of human subjects, $\sigma = 0.043$). Conversely, non-human referents, while less common, are noticeably more likely to be realized lexically; the median lexicality rate for non-humans ($M = 0.45$) is over twice that of humans ($M = 0.18$). For the latter, all corpora cluster relatively tightly together ($\sigma = 0.043$), indicating a high degree of homogeneity in the realization of human referents cross-linguistically. For non-humans, however, there are three notable outliers to an otherwise similarly narrow spread of values ($\sigma = 0.189$), suggesting that the association between form of subject and humanness may be parametricized across languages, a pattern we will encounter again numerous times. Specifically, Northern Kurdish ($P = 0.16$ for humans vs. $P = 0.80$ for non-humans) and Tabasaran ($P = 0.22$ vs. $P = 0.75$) appear to be substantially more sensitive to humanness than the corpus mean ($m = 0.18$ vs. $m = 0.48$), with lexicality rates for non-human referents almost twice as high as the mean. Conversely, the English data are essentially unaffected by humanness ($P = 0.12$ for both humans and non-humans). Among the remainder of corpora, Cypriot Greek shows a somewhat stronger association

1 A short guide to interpreting Tukey boxplots: The gray box shows the interquartile range from the first to the third quartile of the data; the black horizontal line within it indicates the median (i.e. the second quartile); the black brace on its left extends from the most extreme high and low data point not deemed an outlier, here defined as data greater (less) than the third (first) quartile plus (minus) 1.5 times the interquartile range.



		human			non-human				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	77	424	0.18	9	17	0.53	0.17	<0.001
E	English	103	858	0.12	60	497	0.12	0.00	0.971
M	Mandarin	172	634	0.27	36	85	0.42	0.11	0.004
N	Nafsan	98	580	0.17	44	118	0.37	0.19	<0.001
K	N. Kurdish	91	584	0.16	47	59	0.80	0.45	<0.001
S	S. Dargwa	94	471	0.20	22	53	0.42	0.16	<0.001
B	Tabasaran	163	731	0.22	41	55	0.75	0.30	<0.001
T	Teop	130	667	0.19	37	84	0.44	0.19	<0.001
L	Tulil	60	422	0.14	91	198	0.46	0.34	<0.001
V	Vera'a	397	2227	0.18	95	203	0.47	0.20	<0.001
	totals	1385	7598	—	482	1369	—	—	—

Figure 6.1 | Lexicality of anaphoric subjects by humanness, across corpora.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	humanness	(*human*, non-human)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.338
speaker	37	0.382

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.21	−1.551	0.134	−11.58	<0.001
(A_1)	humanness	= non-human	3.50	1.254	0.073	17.10	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−1.180	−0.528	−0.451	−0.324	4.721

d. | correlation of fixed effects

	(intercept)
(A_1)	−0.118

e. | model evaluation

observations	8967	AIC	8760
model deviance	8752	log-likelihood	−4376
residual d.f.	8963	conditional R^2	0.123
		marginal R^2	0.054

Table 6.1 | Summary of regression model results for the lexicality of anaphoric subjects by humanness, with corpus and speaker as random effects.

than the mean, and Mandarin a somewhat weaker one. For the latter, this is due to a comparatively higher rate of lexical subjects among human referents, rather than due to a higher rate among non-humans.

This distribution is reflected in the φ-coefficients for the corpora given in the table at the bottom of Figure 6.1. Pearson's φ is related to the χ^2-test statistic, but mapped to a range from −1 to 1 similar to a correlation coefficient,[2] with $\varphi = 0$ indicating no association between humanness and lexicality, and values closer to −1 and 1 indicating increasingly better associations. Unsurprisingly, Northern Kurdish ($\varphi = 0.45$, $p < 0.001$) and Tabasaran ($\varphi = 0.30$, $p < 0.001$) have some of the highest φ-coefficients in the sample, whereas the coefficient for English indicates that there is no association whatsoever ($\varphi = 0.00$, $p = 0.971$).

Complementing Figure 6.1, Table 6.1 summarizes a generalized linear mixed-effects regression model (GLMR) fit to the data, with humanness as the sole fixed effect and corpus and speaker as random effects. While such simplistic models are not particularly elucidating by themselves, they can aid in the identification of broad patterns in the data, and help by putting a number (or numbers, as it were) to the picture shown by Figure 6.1. The regression models are calculated using the `glmer` function from the *lme4* R package (Bates et al. 2020).

Corpus and speaker are included as random rather than fixed effects (i.e. "predictors"), since at this stage in the exploration of the data, we are more interested in the general degree of variability between corpora and speakers, rather than in their specific structural effect on the response variable. The higher the standard deviation σ of the intercepts of the random effects (in Table 6.1 subsection a.), the greater the variability between the various groups (corpora, or speakers) of that effect. As a rule of thumb, values below $\sigma < 0.5$ suggest an appreciable degree of homogeneity among the groups. With a values of $\sigma = 0.338$ and $\sigma = 0.382$ respectively, both corpora and speakers fall below this threshold, albeit towards the upper end – not surprising given the three outliers noted above.

Finally, the log odds e^{β} (i.e. the exponentiated model coefficients β) in subsection b. of Table 6.1 estimate how much of an effect differences in one

2 In fact, for two binary variables as in this case, the φ-coefficient is equal to the Pearson correlation coefficient r.

of the fixed effects have on the selection of one level of the response variable over the other when all other fixed effects in the model (of which there are none in this case of course) are held at their respective reference levels. The respective reference levels are given it italics at the top of the model summary table. For humanness specifically, a non-human referent has $e^{\beta} = 3.50$ times higher odds of being realized lexically compared to a human referent, which is a significant difference ($p < 0.001$) in line with our impressions of Figure 6.1.

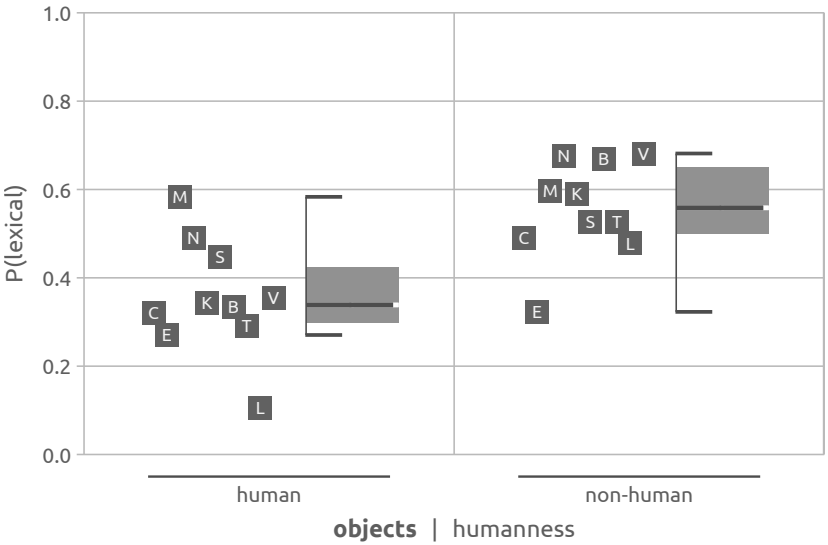
In sum, the well-known tendency for human subjects to be expressed non-lexically (cf. the light subject constraint, Chafe 1994) is consistently represented in all corpora in the sample. Non-human subjects conversely exhibit some degree of cross-corpus variation, ranging from high (Northern Kurdish, Tabasaran) to no sensitivity (English) to humanness.

6.1.2.2 | Objects

The complementary picture for objects is shown in Figure 6.2 and Table 6.2. Unlike subjects, only a minority of mentions in object position are of human referents (cross-corpus mean $P = 0.33$, $\sigma = 0.120$). But as with subjects, non-human referents are overall more likely to be expressed lexically than human referents ($P = 0.35$ for humans vs. $P = 0.56$ for non-humans; with $e^{\beta} = 2.75$ times higher odds if non-human), although the difference is smaller than for subjects, as lexical objects are more common in general.

Objects exhibit an overall greater degree of cross-corpus variation than subjects (random effect intercept $\sigma = 0.383$), and notably, also a substantial degree of inter-speaker variation ($\sigma = 0.603$), which may be attributable to differences in the content of individual texts. We might speculate that where the subject matter of a narrative is focused on interactions between people, rates of lexical human anaphora will tend to be lower than in those that are not.

As for differences between individual corpora, the English data again appear largely indifferent to the effects of humanness ($P = 0.27$ for humans vs. $P = 0.32$ for non-humans, $\varphi = 0.04$ with $p = 0.330$), but is here joined by Sanzhi Dargwa ($P = 0.45$ vs. $P = 0.53$, $\varphi = 0.08$ with $p = 0.393$) and Mandarin ($P = 0.58$ vs. $P = 0.60$, $\varphi = 0.01$ with $p = 0.849$) as notable breakaways from the central pattern, with lexicality rates higher for Sanzhi than for English, and higher for Mandarin than for Sanzhi. The high proportion of lexical objects in Mandarin, already noted earlier, is especially apparent here.



		human			non-human				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	25	78	0.32	77	157	0.49	0.16	0.013
E	English	23	85	0.27	206	638	0.32	0.04	0.330
M	Mandarin	49	84	0.58	71	119	0.60	0.01	0.849
N	Nafsan	48	98	0.49	88	130	0.68	0.19	0.004
K	N. Kurdish	33	96	0.34	95	161	0.59	0.24	<0.001
S	S. Dargwa	21	47	0.45	41	78	0.53	0.08	0.393
B	Tabasaran	13	39	0.33	137	205	0.67	0.25	<0.001
T	Teop	36	124	0.29	69	131	0.53	0.24	<0.001
L	Tulil	6	57	0.11	86	180	0.48	0.33	<0.001
V	Vera'a	82	231	0.35	261	383	0.68	0.32	<0.001
	totals	336	939	—	1131	2182	—	—	—

Figure 6.2 | Lexicality of anaphoric objects by humanness.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	humanness	(*human*, non-human)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.383
speaker	37	0.603

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.53	−0.639	0.182	−3.51	<0.001
(A_1)	humanness	= non-human	2.75	1.013	0.092	10.97	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−2.237	−0.789	−0.479	0.866	2.455

d. | correlation of fixed effects

	(intercept)
(A_1)	−0.329

e. | model evaluation

observations	3121	AIC	4028
model deviance	4020	log-likelihood	−2010
residual d.f.	3117	conditional R^2	0.181
		marginal R^2	0.054

Table 6.2 | Regression model results for the lexicality of anaphoric objects by humanness, with corpus and speaker as random effects.

Another notable observation is the strong disinclination towards lexical human objects in Tulil ($P = 0.11$). Pronouns in Tulil are relatively informative, distinguishing number, three genders, and a range of classifiers. Human objects in Tulil are almost exclusively pronominal, with zero found only among non-human referents:[3]

(66) Tulil
Io kəvənau laike itəra nəbo ba tipur.

	io	*kəvənau*	*laik*	*=e*	*i-tər*	*=a*
	then	rain	big	=SG.CL:FEM	3SG.F.PST-meet	=3SG.M.PAT
##	other	np:a	rn	=rn	v:pred	=pro.h:p
		0006				0003

nə-bo	*ba*	*tipur*
LOC-UP	in	forest
dem_other	adp	np:l
		0007

'Then heavy rain came to him there in the bush.'

[mc_tulil_all1_0004]

Compared to the mostly cross-corpus homogeneously distribution of lexical subjects by humanness – with the above-noted exceptions – the distribution of lexical objects shows much greater variation, with some languages (English, Sanzhi Dargwa, Mandarin) appearing insensitive to the effects of humanness, while others exhibit strongly skewed patterns (Tulil). There is also notable and presumably content-related variation between speakers, but from a broad cross-corpus angle, the difference between the rate of lexical expression for human and non-human mentions is the same for subjects and objects, irrespective of position.

In the next section, we briefly turn our attention to a more nuanced subdivision of non-human referents.

3 Compare Genetti & Crain (2003), who find that pronominal objects are restricted to human referents in narrative discourse in Nepali (Indo-Iranian, Nepal).

6.2 | Protagonisthood

6.2.1 | Definition and methodological issues

The centrality of a referent to narrative has been claimed to have influence on discourse structure and referential choice; in the literature, this idea is most commonly conceptual either in terms of protagonisthood (in the terminology of Kibrik et al. 2016 and others) or VIP'ness (in the terminology of Dooley & Levinsohn 2001). The more important a referent is from a narrative perspective, the higher its baseline level of accessibility is assumed to be, and hence the lower the likelihood of its anaphors being realized lexically is.

Which referents should be considered protagonists is left to the annotator's interpretation in other studies (e.g. Kibrik et al. 2016), but for this study we employ a programmatic approach to assigning protagonist status which is based on the overall frequency of mentions of a given referent in a text, normalized by the total number of mentions of all referents in said text. The resulting total mention frequency ratio represents the proportion of mentions taken up by a referent in a text. The higher this ratio, the more frequently mentioned and hence the more central to the narrative that referent is. All else being equal, higher frequency ratios should thus associate with lower rates of lexical expression. However, while these ratios allow for comparison of the referents in a given text, they cannot be directly compared between different texts and corpora, since the length of the texts in the sample varies considerably. The most frequent referent in a long text that contains hundreds of other referents would likely still be lower than a less frequent referent in a text with only a handful of other referents. To make this measure comparable across texts of differing lengths, in the following we will not be working with mention frequency ratios directly, but rather with their percentile of the distribution of all ratios in a text.

Figure 6.3 shows the number of observations (for both subjects and objects) in each percentile of the frequency ratio distribution across all texts. Unsurprisingly, the highest ratios account for the vast majority of data points in the sample; in this way, Figure 6.3 closely resembles a power law distribution. For example, in one of the Northern Kurdish texts (*muserz01*), the five most frequent referents account for over two thirds (68%) of all mentions in subject position.

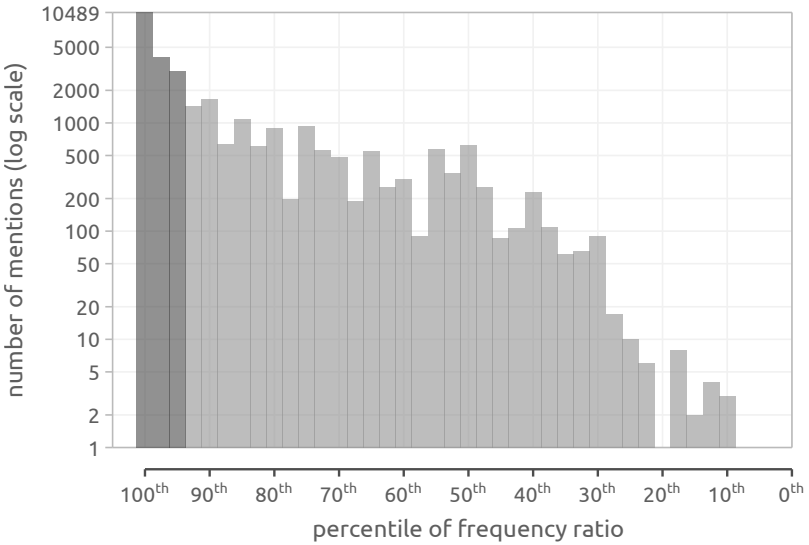


Figure 6.3 | Histogram of mention frequency ratios for all referents in the sample.

Related treatments of protagonisthood and VIP'ness in the literature are chiefly focused on the primacy of the most narratively central referents rather than their relative position on a cline of narrative centrality. This is a reasonable assumption, so we here simplify our percentile-based measure further, down to a binary distinction between

a. the 5% most frequent referents in a text,
b. the 95% least frequent referents, that is all others.

A viable alternative would be to include the frequency percentile as a scalar factor; for the sake of simplicity, this avenue is not pursued here. Note that this factor is agnostic of the actual position of an anaphor in a text, as it is calculated on the basis of all mentions, including those located in the future from the perspective of the anaphor and the speaker.

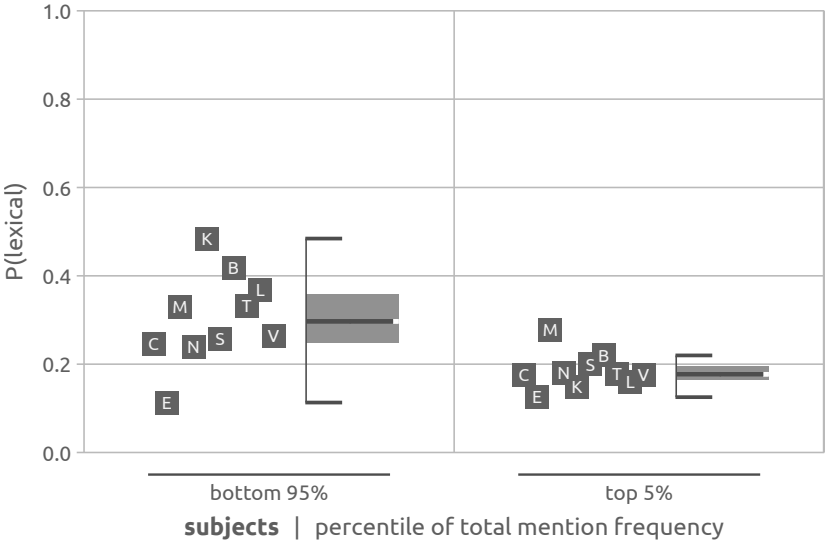
6.2.2 | Lexicality by total frequency

6.2.2.1 | Subjects

Figure 6.4 shows that from a broad cross-corpus perspective, highly frequent referents (i.e. the 5% most frequently mentioned in a given text) are less likely to be mentioned lexically in subject position than the remainder of less frequent referents is (cross-corpus $M = 0.18$ vs. $M = 0.30$). This suggests that greater narrative salience salient does indeed make referents more easily retrievable. The regression model summarized in Table 6.3 supports this observation, indicating that if the referent is among 5% most frequent, the odds of a lexical subject mention are roughly halved ($e^{\beta} = 0.55$, $p < 0.001$) relative to the base odds of $e^{\beta} = 0.39$. As already suggested by the distribution shown in the histogram in Figure 6.3 above, highly frequent referents make up the majority of all third-person mentions in subject position (cross-corpus mean $P = 0.69$; $\sigma = 0.085$).

The top 5% show a narrow spread of values across the ten corpora ($\sigma = 0.042$); comparatively greater variation is present among the bottom 95% ($\sigma = 0.105$), but overall there is still a relatively high degree of homogeneity across corpora (random effects intercept $\sigma = 0.234$). The strength of the association differs to some degree between corpora: In some, among them Northern Kurdish ($\varphi = 0.33$ with $p < 0.001$), Tulil ($\varphi = 0.24$ with $p < 0.001$), and Tabasaran ($\varphi = 0.18$ with $p < 0.001$), total mention frequency shows a stronger effect; in others the effect is weaker, most notably in English ($\varphi = 0.02$ with $p = 0.497$) and Mandarin ($\varphi = 0.05$ with $p = 0.193$), where total frequency makes essentially no difference. Total frequency is thus another factor that form choices in the English data appear to be insensitive to.

In sum, more frequent referents are less likely to be lexical when mentioned in subject position in most of the sample. Note however that higher total mention frequencies are correlated with lower antecedent distance (Section 6.3) and higher recent co-referential frequeny (Section 6.4), though not strongly; see Section 7.2.3.1 further below.



	corpus	bottom 95% N(lex)	bottom 95% N(all)	bottom 95% P(lex)	top 5% N(lex)	top 5% N(all)	top 5% P(lex)	φ	*p*-val.
C	C. Greek	29	118	0.25	57	323	0.18	0.08	0.104
E	English	63	557	0.11	100	798	0.13	0.02	0.497
M	Mandarin	55	167	0.33	153	552	0.28	0.05	0.193
N	Nafsan	67	280	0.24	75	418	0.18	0.07	0.054
K	N. Kurdish	62	128	0.48	76	515	0.15	0.33	<0.001
S	S. Dargwa	51	198	0.26	65	326	0.20	0.07	0.120
B	Tabasaran	66	158	0.42	138	628	0.22	0.18	<0.001
T	Teop	72	217	0.33	95	534	0.18	0.17	<0.001
L	Tulil	92	250	0.37	59	370	0.16	0.24	<0.001
V	Vera'a	193	730	0.26	299	1700	0.18	0.10	<0.001
	totals	750	2803	—	1117	6164	—	—	—

Figure 6.4 | Lexicality of anaphoric subjects by percentile of total mention frequency.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	total freq.	(*bottom 95%*, top 5%)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.234
speaker	37	0.341

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.39	−0.942	0.108	−8.74	<0.001
(A_1)	total freq.	= top 5%	0.55	−0.593	0.057	−10.48	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−0.868	−0.532	−0.455	−0.357	3.628

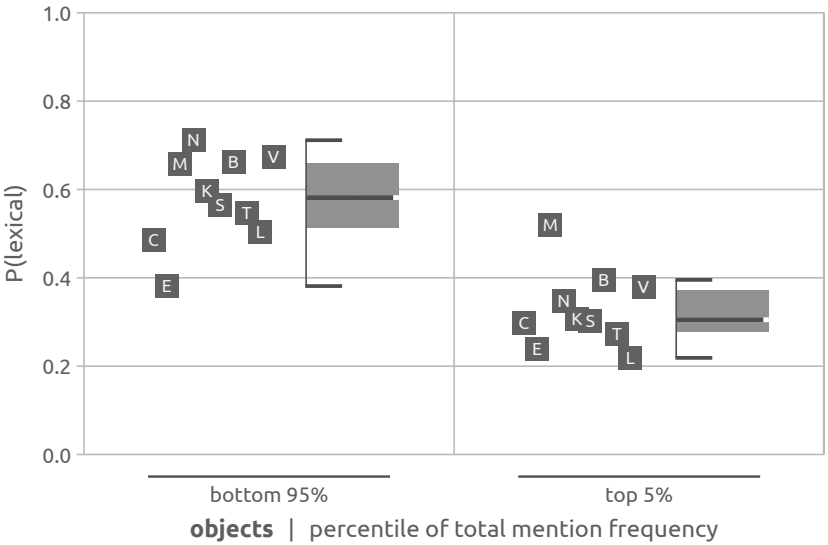
d. | correlation of fixed effects

	(intercept)
(A_1)	−0.318

e. | model evaluation

observations	8967	AIC	8939
model deviance	8931	log-likelihood	−4465
residual d.f.	8963	conditional R^2	0.070
		marginal R^2	0.021

Table 6.3 | Regression model results for the lexicality of anaphoric subjects by overall frequency, with corpus and speaker as random effects.



		bottom 95%			top 5%				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	83	171	0.49	19	64	0.30	0.17	0.009
E	English	151	396	0.38	78	327	0.24	0.15	<0.001
M	Mandarin	69	105	0.66	51	98	0.52	0.14	0.048
N	Nafsan	111	156	0.71	25	72	0.35	0.35	<0.001
K	N. Kurdish	101	169	0.60	27	88	0.31	0.28	<0.001
S	S. Dargwa	52	92	0.57	10	33	0.30	0.23	0.010
B	Tabasaran	133	201	0.66	17	43	0.40	0.21	0.001
T	Teop	71	130	0.55	34	125	0.27	0.28	<0.001
L	Tulil	71	141	0.50	21	96	0.22	0.29	<0.001
V	Vera'a	252	374	0.67	91	240	0.38	0.29	<0.001
	totals	1094	1935	—	373	1186	—	—	—

Figure 6.5 | Lexicality of anaphoric objects by percentile of total mention frequency.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	total freq.	(*bottom 95%*, top 5%)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.329
speaker	37	0.492

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	1.45	0.370	0.150	2.48	0.013
(A_1)	total freq.	= top 5%	0.36	−1.015	0.083	−12.19	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−1.980	−0.835	−0.503	0.849	2.036

d. | correlation of fixed effects

	(intercept)
(A_1)	−0.189

e. | model evaluation

observations	3121	AIC	4001
model deviance	3993	log-likelihood	−1997
residual d.f.	3117	conditional R^2	0.153
		marginal R^2	0.062

Table 6.4 | Regression model results for the lexicality of anaphoric objects by overall frequency, with corpus and speaker as random effects.

6.2.2.2 | Objects

For objects, the association between form and overall frequency is visualized in Figure 6.5, and a single-factor regression model fit to the data is summarized in Table 6.4. Where most subject anaphors are mentions of highly frequent referents, the majority of object anaphors are mentions of less frequent referents instead ($P = 0.36$ highly frequent; $\sigma = 0.104$). This is well in line with the higher degree of referential continuity carried by the subject role (cf. Givón 1983a; see also Section 6.7).

From a broad cross-corpus perspective, the association between total frequency and lexicality is more pronounced for objects than it is for subjects (cross-corpus mean $\varphi = 0.24$; $\sigma = 0.069$): The 5% most frequent referents in each text, when mentioned in the third person in object position, are roughly half as likely to be realized lexically than less common referents ($M = 0.30$ vs. $M = 0.58$). For these referents, the odds of a lexical anaphor in object position are about a third of the odds of a non-lexical one ($e^{\beta} = 0.36$, $p = 0.001$).

Taking into account the greater dispersal of values among corpora, there are no notable outliers among corpora; the strength of the association is weakest in Mandarin cross-corpus mean ($\varphi = 0.14$ with $p = 0.048$), which also showed weaker effect for subjects as seen above, and strongest in Nafsan ($\varphi = 0.35$ with $p = 0.001$). Variation between speakers is comparatively high (random effect intercept $\sigma = 0.492$); as noted before, this is a common pattern for object mentions, and likely the result of content-related differences between texts.

In sum, the more central a referent is to the narrative in a text as quantified by its total frequency ratio, the less likely it is to be realized lexically in object position, with an effect slightly stronger but less cross-linguistically stable than the corresponding effect observed for subjects in Section 6.2.2.1.

6.2.3 | Interaction with humanness

Given the human-centric nature of the texts in the corpora, it does not come as a surprise that the overwhelming majority of highly frequent referents are human: The proportion of humans among the 5% most frequently mentioned referents in the sample is over twice of that among the 95% least frequent (cross-corpus mean $P = 0.84$ with $\sigma = 0.151$ vs. $P = 0.35$ with $\sigma = 0.063$). The corpus with the lowest proportion of human referents in the former group is English ($P = 0.43$ vs. $P = 0.32$).

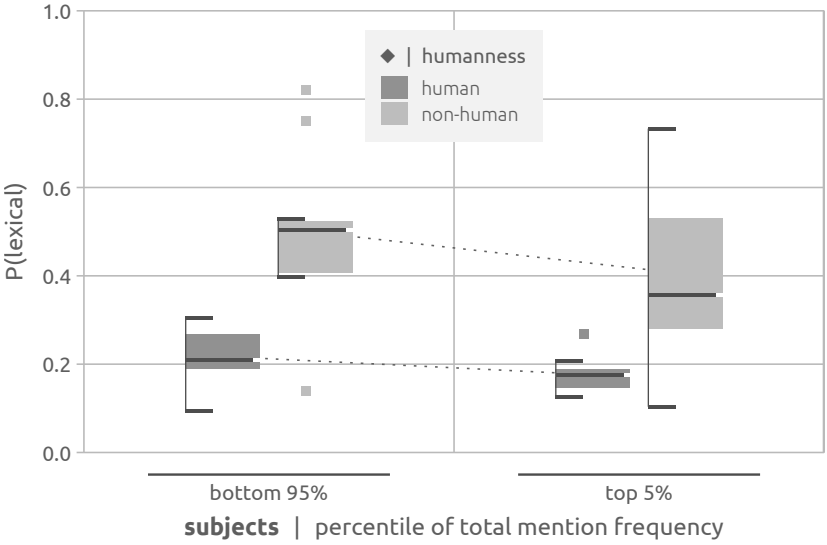


Figure 6.6 | Interaction of the effect of overall frequency and humanness on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

Figure 6.6 shows the association between humanness and total frequency for subjects, and Figure 6.7 does so for objects. For subjects, there is only a small difference between the lexicality rates of the top 5% and bottom 95% of human referents; the association is comparatively stronger, but still marginal, for non-human referents (mean $\varphi = 0.06$ and $\sigma = 0.038$ for humans; $\varphi = 0.11$ and $\sigma = 0.056$ for non-humans).

For objects, there is very wide spread of values for less-frequent human referents among the ten corpora in the sample, ranging from less than 10% lexical to almost 90% lexical. In comparison, the lexicality rates of highly frequent human referents are clustered tightly together. This suggests that corpora differ substantially on how those human referents that are not among the most narratively central in the text are realized in object position. As with subjects, the human referents are less likely to be realized lexically, slightly

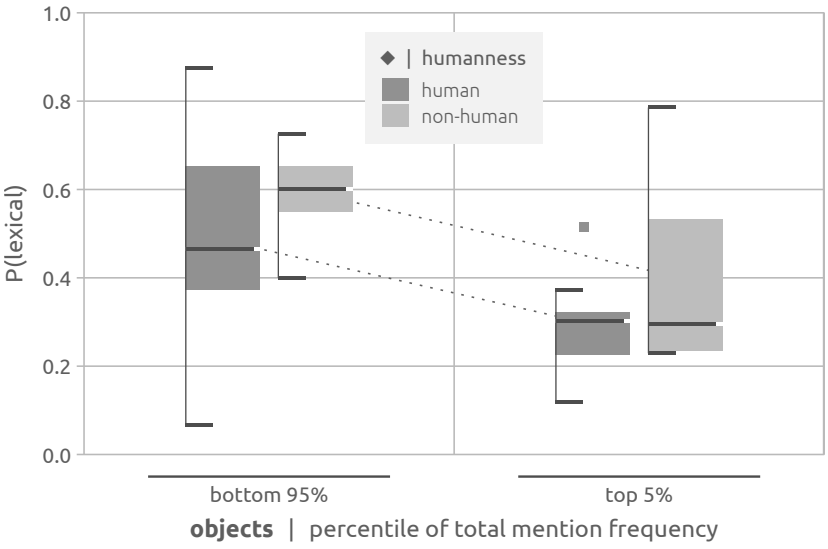


Figure 6.7 | Interaction of the effect of overall frequency and humanness on the lexicality of anaphoric objects, by corpus.
The small squares indicate outliers from the central distribution.

more so if they are also among the most frequent ($\varphi = 0.21$ and $\sigma = 0.151$ for humans; $\varphi = 0.17$ and $\sigma = 0.151$ for non-humans).

The relative lack of an association between total frequency and humanness suggests that the inherent salience of human referents is not substantially boosted by the added salience of protagonisthood. Non-human referents conversely see more of a benefit from high frequency, but this combination is quite rare.

6.3 | Anaphoric distance

6.3.1 | Definition and methodological issues

Anaphoric distance, and more generally recency effects, can be quantified in a variety of manners, some of which have been discussed in above in Section 4.6.1.1. Perhaps the most straightfoward is the figurative "textual distance" between a given mention of a referent and its antecedent. Three elements are needed for calculation of this anaphoric (or antecedent) distance, all of which have already been touched upon in previous sections: First, a means of identifying which referring expressions in a text share the same underlying referent, which is provided by the corpus annotations, as outlined earlier in Section 3.3.2. Second, a rule for selecting specific co-referential mentions as the antecedent to which to measure distances, which has been defined in Section 4.6.2 as the most recent NP-level mention occuring in linear order, irrespective of its form. And finally, we need a unit to measure anaphoric distance in; as discussed in Section 4.6.1, we here use clause units as the basic unit of discourse segmentation, and hence as a unit of measurement.

Example (67) illustrates the principle used for the calculation of anaphoric distances. By sequentially assigning a numerical index to each clause unit (independent or otherwise), we can quite trivially calculate anaphoric distances by subtracting the index of the clause containing the antecedent *he* (⟨0530⟩) in (67a) from the index of the clause with the anaphor in question in (67c):

(67) English

a.

	and	*he*	*slammed*	*the*	*bell*	*off*
	and	3SG.M	slam.PST	the	bell	off
##	other	pro.h:a	v:pred	ln_det	np:p	rv
		0530			0524	

'And he slammed the bell off, …'

b.

	and	*it*	*broke*	*the*	*stay*
	and	3SG.N	break.PST	the	stay
##	other	pro:a	v:pred	ln_det	np:p
		0524			0531

'and it broke the stay, …'

c. *and* *he* *went* *up* *along* *with* *it*
and 3SG.M go.PST up along with 3SG.N.OBL
other pro.h:s v:pred rv adp adp pro:obl
0530 0531

'and he went up along with it.' [mc_english_kent03_0500]

As 3 − 1 = 2, the anaphor is exactly two clause units away from its immediate antecedent. Similarly, *it* (⟨0524⟩) in (67c) has a distance of 2 − 1 = 1 clauses from its antecedent *the bell* in (67b). A distance of $d = 0$ clause, then, indicates an antecedent in the same clause, and a distance of $d = 1$ an antecedent in the previous clause. Naturally, the calculations are not always as straightforward as this, but in the vast majority of cases the distance measurements have a very high degree of accuracy.

The treatment of same-clause anaphors is particularly relevant for the calculation of anaphoric distances; as discussed in Section 4.6.2.1, subject and object anaphors following a previous co-referential mention in the same clause (such as a topic NP or a possessor, etc.) are excluded from the sample, as their realizations are likely subject to different constraints than other anaphors (cf. Arnold 2010: 190).

Lastly, as mentioned earlier in Section 4.5, a side effect of how referents are distinguished in the annotations is that the distance measures are agnostic of any existing semantic relationships between referents and their mentions. To compensate, potential priming effects from mentions of related referents are discussed in a little later in Section 6.5 below.

6.3.1.1 | Capping extreme distances

Figure 6.8 shows a histogram of the measured antecedent distances in clause units, for both subject and object mentions, across all ten corpora in the sample. The vast majority of anaphora occur at very low distances to the antecedent, in particular in the next clause ($d = 1$). Frequencies drop off rapidly with increasing distance, leaving the distribution of distances with a long tail that stretches into higher distances, each with only a small number of mentions. While the distances on the x-axis in Figure 6.8 cut off at $d = 70$ clauses for presentational reasons, the actual maximum distance in the sample is over $d > 1000$ clauses, which is found in one of the longer texts in the English corpus.

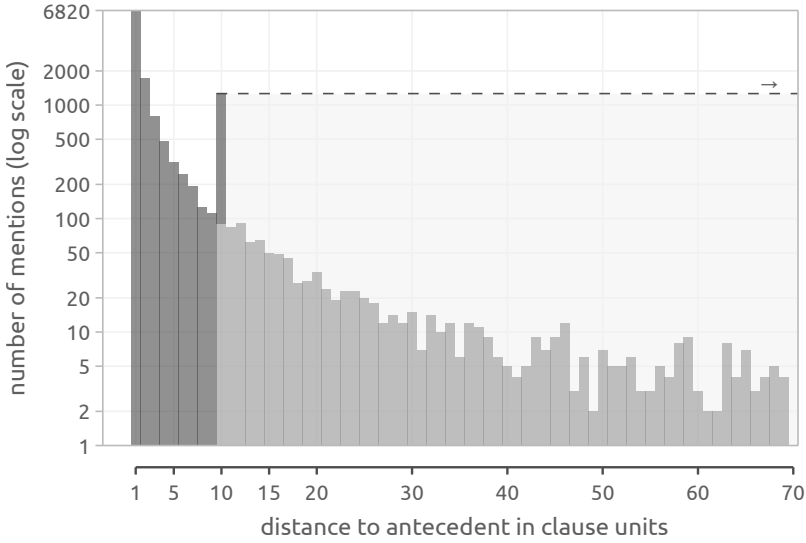


Figure 6.8 | Histogram of anaphors at various anaphoric distances, including both subjects and objects.
The dark grey bars on the left side of the graph show the effective sample, winsorized at $d = 10$ clause units; the light grey bars mentions at distances above the winsorization threshold.

Figure 6.9 offers another perspective on these data. Here, the vertical axis shows the maximum distance in clause units at various quantiles of the data. The maximum distance value at the 90th percentile, for instance, is $d = 10$, meaning that 90% of the mentions in the sample occur at distances of $d = 10$ or less clause units from their antecedent. Conversely, only 10% of anaphors, the aforementioned tail end of the distribution, have an anaphoric distance greater than $d = 10$ clauses.

Both from a practical and a descriptive perspective, it makes sense to curtail the influence of extremly high distance values in the data. It is reasonable to assume that above a given threshold, changes in anaphoric distance cease to have an appreciable effect on referential choice, as activation levels will

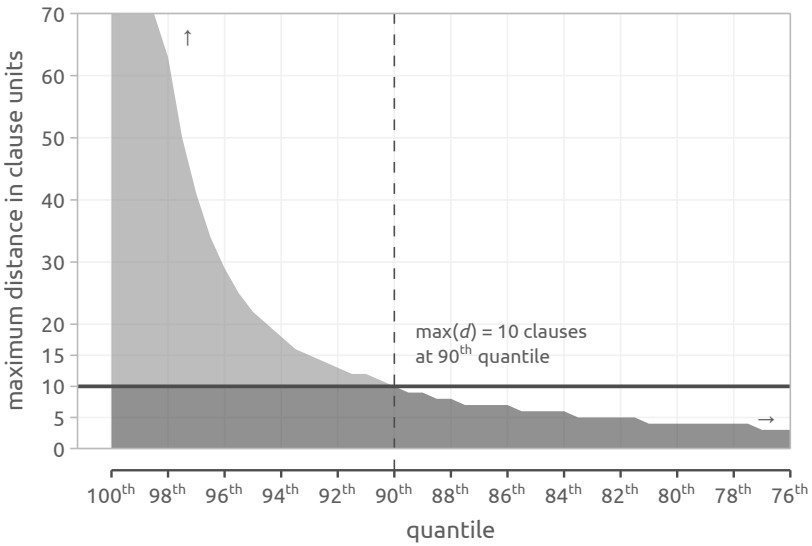


Figure 6.9 | Maximum anaphoric distance at various percentiles of the data, for mentions in both subject and object position.
The dark grey area delineates the maximum distance of $d = 10$ at the 90th percentile.

have plateaued at a global minimum, though it is unfortunately not possible to systematically test this assumption as part of this study, given the low frequency of extremely high-distance anaphors. From a statistical point of view, an excessive number of outliers far from the median can skew the mean of the distribution, hurting predictive models. Furthermore, since maximum distance is correlated with text length and text length is not uniformly distributed across corpora in the sample (as seen earlier in Table 4.1 in Section 4.4), the skewness differs between languages, running the risk of further distorting the picture.

With that in mind, it is sensible to not leave observations with high distance values as they are. But rather than discard them, I have chosen to instead winsorize all values above the 90th percentile, that is above a threshold

of $d = 10$ clause units. In other words, all observations with distances $d > 10$ clause units are treated as occuring at exactly $d = 10$ clause units. Winsorization limits the adverse effects of outliers while preserving the weight high values exert on the overall distribution, which outright removal of data points does not. For comparison, here are the median (M), mean (m), and standard deviation (σ) for the distance values in the sample, before and after winsorization, in clause units from the antecedent:

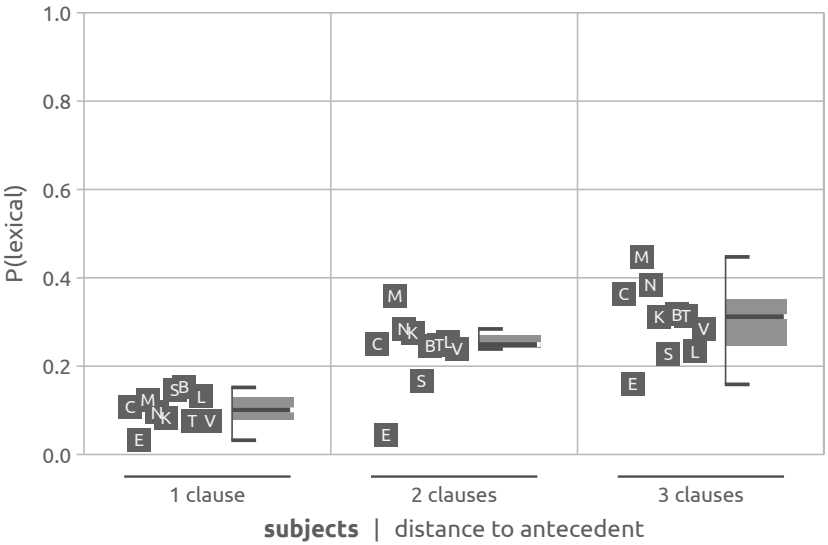
- before winsorization: $M = 1.00$; $m = 7.12$, $\sigma = 35.303$
- after winsorization: $M = 1.00$; $m = 2.79$, $\sigma = 2.942$

Note the vastly different dispersal of values, indicated by the change in the difference between median and mean and the standard deviations. The winsorization threshold of $d = 10$ at the 90th percentile has been chosen more or less arbitrarily, and arguments could be made for a higher or lower value. If we estimate a sentence to contain about two clauses on average and a paragraph of written text about five sentences, the "ten clauses and above" category could be taken as a rough approximation of Ariel's (1990) and others' "across paragraphs" category in terms of distance, if not as regards the effect of episodic boundaries (cf. Fox 1987a).

6.3.2 | Lexicality by anaphoric distance

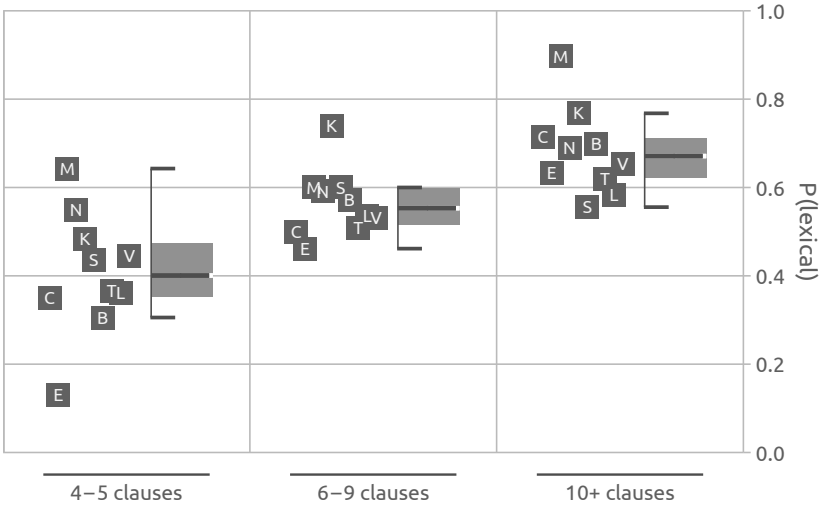
In the graphs presented in this section, certain distances are grouped together for the purpose of visualization. In particular, the graphs merge distance values between four and five clauses into a "4–5" category, and values between six and nine clauses into a "6–9" category. The rationale for this simplification is that since the vast majority of anaphors occur at low distances from their antecedent, the intermediate distance categories below the winsorization threshold are left with low subsample sizes when split across individual corpora and syntactic positions. The grouping of selected categories reduces the resolution of the graphs, but in turn makes broader patterns easier to identify. Importantly, this grouping affects only the graphical representations; the regression model and any calculated statistics use the ungrouped data instead.

A further preliminary note concerns the interpretation of the regression coefficients of scalar variables such as anaphoric distance. To calculate the effective model coefficients for a specific distance, the given coefficient β in the model summary tables needs to be multiplied by the desired distance



		1 clause			2 clauses			3 clauses		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	32	300	0.11	15	60	0.25	8	22	0.36
E	English	23	713	0.03	11	253	0.04	20	126	0.16
M	Mandarin	56	452	0.12	32	89	0.36	17	38	0.45
N	Nafsan	45	473	0.10	27	95	0.28	15	39	0.38
K	N. Kurdish	37	442	0.08	17	62	0.27	9	29	0.31
S	S. Dargwa	40	275	0.15	19	114	0.17	10	44	0.23
B	Tabasaran	70	461	0.15	32	130	0.25	13	41	0.32
T	Teop	32	416	0.08	25	101	0.25	16	51	0.31
L	Tulil	46	350	0.13	23	90	0.26	7	30	0.23
V	Vera'a	120	1563	0.08	59	247	0.24	38	134	0.28
	totals	501	5445	—	260	1241	—	153	554	—

Figure 6.10 | Lexicality of anaphoric subjects by distance to their antecedent, measured in clause units.



4–5 clauses			6–9 clauses			10+ clauses				
N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	r_{pb}	*p*-val.
7	20	0.35	9	18	0.50	15	21	0.71	0.39	<0.001
12	92	0.13	30	65	0.46	67	106	0.63	0.54	<0.001
27	42	0.64	24	40	0.60	52	58	0.90	0.52	<0.001
22	40	0.55	13	22	0.59	20	29	0.69	0.42	<0.001
15	31	0.48	17	23	0.74	43	56	0.77	0.54	<0.001
17	39	0.44	15	25	0.60	15	27	0.56	0.32	<0.001
11	36	0.31	20	35	0.57	58	83	0.70	0.41	<0.001
19	52	0.37	28	55	0.51	47	76	0.62	0.44	<0.001
17	47	0.36	23	43	0.53	35	60	0.58	0.36	<0.001
61	137	0.45	60	113	0.53	154	236	0.65	0.48	<0.001
208	536	—	239	439	—	506	752	—	—	—

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	ante. distance	(*1*–10+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.248
speaker	37	0.397

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.13	–2.025	0.115	–17.66	<0.001
(A)	ante. distance	* [0, 9]	1.42	0.352	0.010	36.33	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–2.508	–0.428	–0.359	–0.263	5.247

d. | correlation of fixed effects

	(intercept)
(A)	–0.187

e. | model evaluation

observations	8967	AIC	7491
model deviance	7483	log-likelihood	–3742
residual d.f.	8963	conditional R^2	0.258
		marginal R^2	0.208

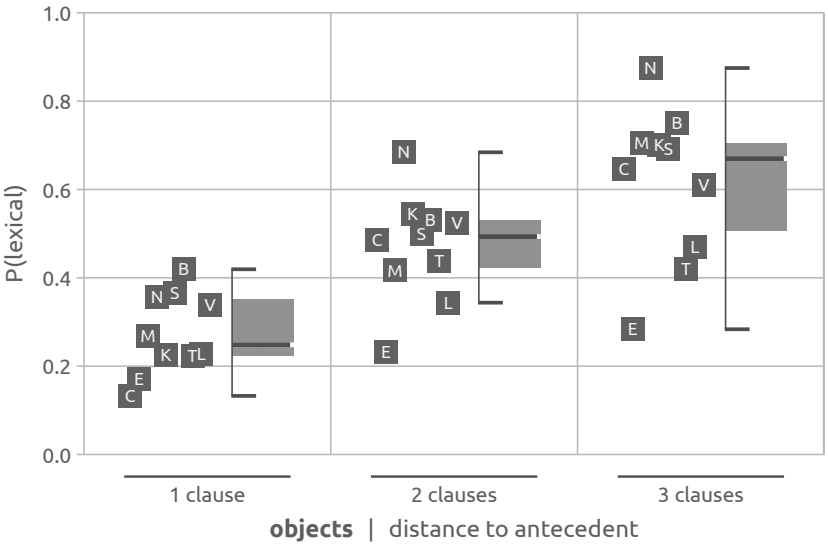
Table 6.5 | Regression model results for the lexicality of anaphoric subjects by distance to the antecedent, with corpus and speaker as random effects.

minus one (i.e. $\beta_d = \beta \times (d-1)$, where d is the distance in clause units). Distance values need to be shifted down by one since in the model formulae, the reference level of continuous factors is by definition zero, but the shortest distance is $d = 1$ clause units; continuous variables in later sections whose lowest level is already zero obviously do not require this adjustment. As the odds ratios e^{β} are equivalent to the exponentiated coefficients, the odds for a specific distance value can be calculated directly by raising the odds ratio of the reference level to the power of the desired distance minus one (i.e. $e^{\beta}_{d} = e^{\beta\,(d\,-\,1)}$).

6.3.2.1 | Subjects

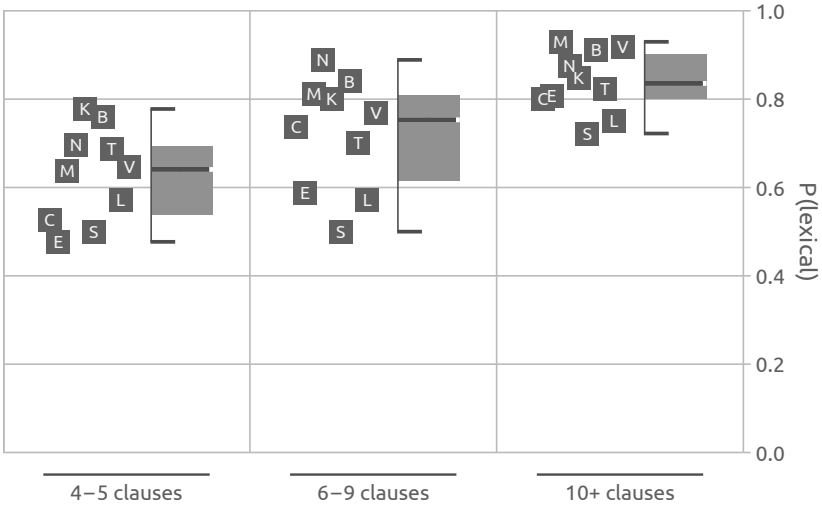
Figure 6.10 shows the rates of lexical expression of subjects by distance to the antecedent in clause units. First, a few general observations. As noted above in Section 6.3.1.1, low-distance anaphors vastly outnumber higher-distance anaphors: Across all corpora, well over half of all subject anaphora (cross-corpus mean $P = 0.61$, $\sigma = 0.064$) occur in the clause immediately following their antecedent. With only a small number of exceptions, proportions of lexical expressions across all corpora increase monotonically with distance to the antecedent (cross-corpus mean $\rho = 0.82$, $\sigma = 0.108$). As the regression model in Table 6.5 shows, a one-clause difference in anaphoric distance increases the odds of a lexical expression $e^{\beta} = 1.42$ times ($p < 0.001$). Rates start at around a tenth lexical for antecedents in the preceding clause ($P = 0.10$, $\sigma = 0.037$), and end up at well over half lexical at distances of $d \geq 10$ and higher ($P = 0.68$, $\sigma = 0.099$). Across corpora, lexical NPs become the preferred form for subjects (i.e. with an incidence rate above 50%) at distance to the antecedent of six clauses and above. With an antecedent in the preceding clause, subjects are lexical at a rate lower than the baseline level identified earlier in Section 5.2.1. The graded effect of anaphoric distance on lexicality that is apparent in these data also highlights the importance of conceptualizing referential choices as a continuum, rather than in terms of categorical cut-offs (as, e.g. in Givón 1995).

As the regression model suggests, anaphoric distance by itself already manages to explain a substantial proportion of the variance in the data, with good estimates for goodness-of-fit (cf. Nakagawa's R^2 in Table 6.5, subsection e.) compared to the corresponding values for humanness and protagonisthood in the sections above and most of the factors examined further below. Furthermore, the relatively small difference between the conditional and marginal R^2



		1 clause			2 clauses			3 clauses		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	13	98	0.13	18	37	0.49	11	17	0.65
E	English	55	321	0.17	34	146	0.23	19	67	0.28
M	Mandarin	18	67	0.27	10	24	0.42	12	17	0.71
N	Nafsan	36	101	0.36	26	38	0.68	14	16	0.88
K	N. Kurdish	27	120	0.22	18	33	0.55	14	20	0.70
S	S. Dargwa	19	52	0.37	7	14	0.50	9	13	0.69
B	Tabasaran	39	93	0.42	25	47	0.53	15	20	0.75
T	Teop	30	134	0.22	14	32	0.44	8	19	0.42
L	Tulil	28	123	0.23	11	32	0.34	8	17	0.47
V	Vera'a	90	266	0.34	44	84	0.52	25	41	0.61
	totals	355	1375	—	207	487	—	135	247	—

Figure 6.11 | Lexicality of anaphoric objects by distance to their antecedent, measured in clause units.



4–5 clauses			6–9 clauses			10+ clauses				
N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	r_{pb}	*p*-val.
10	19	0.53	14	19	0.74	36	45	0.80	0.49	<0.001
31	65	0.48	27	46	0.59	63	78	0.81	0.45	<0.001
14	22	0.64	13	16	0.81	53	57	0.93	0.52	<0.001
16	23	0.70	16	18	0.89	28	32	0.88	0.38	<0.001
14	18	0.78	16	20	0.80	39	46	0.85	0.47	<0.001
6	12	0.50	8	16	0.50	13	18	0.72	0.21	0.021
19	25	0.76	21	25	0.84	31	34	0.91	0.36	<0.001
11	16	0.69	14	20	0.70	28	34	0.82	0.44	<0.001
8	14	0.57	4	7	0.57	33	44	0.75	0.41	<0.001
31	48	0.65	40	52	0.77	113	123	0.92	0.45	<0.001
160	262	—	173	239	—	437	511	—	—	—

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	ante. distance	(*1*–10+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.257
speaker	37	0.526

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.48	–0.733	0.139	–5.28	<0.001
(A)	ante. distance	* [0, 9]	1.37	0.315	0.015	21.31	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–4.666	–0.677	–0.476	0.772	2.116

d. | correlation of fixed effects

	(intercept)
(A)	–0.207

e. | model evaluation

observations	3121	AIC	3548
model deviance	3540	log-likelihood	–1770
residual d.f.	3117	conditional R^2	0.307
		marginal R^2	0.235

Table 6.6 | Regression model results for the lexicality of anaphoric objects by distance to the antecedent, with corpus and speaker as random effects.

means that the random effects in the models do not substantially influence the outcome, indicating a comparatively high degree of homogeneity across corpora and speakers. This is true for both subjects and objects both, as we will see further below, though less so for the latter, at least as regards inter-speaker variation. Even so, certain patterns of inter-corpus variability are noticeable. At $d = 1$ clauses distance, all corpora are tightly clustered, suggesting a shared preference for non-lexical expressions. The spread of values widens with increasing anaphoric distance, but as mentioned above, not substantially so (random effect intercept $\sigma = 0.248$).

However, there are two consistent outliers from the general trend, English and Mandarin, with characteristically lower and higher than average rates of lexical expression, respectively, as observed earlier. In English, lexicality rates rise only gradually below the $d < 6$ threshold, indicating a greater tolerance for pronominal forms at low and intermediate distances. Mandarin conversely displays higher-than-average lexicality rates in all contexts, except if the antecedent is in the preceding clause, where it is instead well in line with the general trend. Among the corpora in the sample, English and Mandarin hence appear to exhibit opposing trajectories, with all other corpora falling somewhere inbetween: While both strongly disprefer lexical expressions at distances of $d = 1$ clause units, English extends this disprefence for longer than average, whereas Mandarin more quickly transitions to using lexical expressions. A tentative explanation might point to differences in the preferred reduced form for subjects in these languages – with English strongly favouring pronouns (85% pronouns) and Mandarin zero (83% zero) in equal measure – yet other zero-favouring corpora such as Cypriot Greek and Northern Kurdish do not show the same pattern as Mandarin, and neither do the relatively more pronoun-heavy Vera'a and Teop corpora follow the English pattern.

Another notable observation concerns Sanzhi Dargwa, which, not unlike English, shows little change in lexicality rates below $d < 4$ clauses distance, and also has the lowest correlation coefficient across distances in the sample with $r_{pb} = 0.32$ ($p < 0.001$, vs. the cross-corpus mean of $r_{pb} = 0.44$, $\sigma = 0.076$). The latter is likely the result of the lexicality rate in the highest ($d \geq 10$) category being quite low ($P = 0.56$), lower than that of the previous category ($P = 0.60$). The Sanzhi texts are on the short end of the spectrum, which makes long-distance anaphora less frequent overall; this in turn can exaggerate the effects of random fluctuations in the data. Another factor at play here is that in Sanzhi Dargwa, it is common for an anaphoric subject to be preceded by a

dependent clause with a co-referential subject. This pattern is especially frequent with converb clauses as in (68), although it may also occur with other types of subordinate clauses.

In the third person, the subjects of these clauses are predominantly, but not categorically, zero ($P = 0.74$), and only occasionally lexical NPs ($P = 0.21$) or pronouns ($P = 0.06$). In the Sanzhi Dargwa corpus, converb clauses in general are very common: Almost half of all anaphoric subjects occur within them ($P = 0.41$). Notably, this is true even at high anaphoric distances ($P = 0.40$ for $d \geq 6$ clauses distance from the antecedent), indicating that speakers give no consideration to recency when planning converb clauses, though the rate of zero subjects in converb clauses does halve in high-distance contexts ($P = 0.38$). A similar pattern is also found in Tabasaran, a related Nakh-Daghestanian language, but is not as strongly pronounced there.

(68) Sanzhi Dargwa
c'il helka duc'rik'ul t:uraričib car Pat'ima, k:aza k'at'ara haq:ible.

	c'il		*hel-ka*	*duc'*	*r-ik'-ul*	
	then		that-down	run	F-move.IPFV-ICVB	
##	other	#cv	0.h:s 0001	other:g	lv_v	v:pred	%

t:ura-r-ič-ib	*ca-r*	*Pat'ima*
outside-F-occur.PFV-PRET	be-F	Patima
v:pred	rv_aux	pn_np.h:s 0001

	k:aza	*k'at'a*	*=ra*	*h-aq:-ib-le*
	fork	spade	=and	upwards-carry-PRET-CVB
#cv	0.h:a 0001	np:p 0022	rn_np 0023	=rn v:pred

'Then, running down there, Patima arrived carrying a fork and a spade.'
[mc_sanzhi_patima_0035]

6.3.2.2 | Objects

Figure 6.11 shows the rates of lexical expression of objects anaphors at different distances from the antecedent. As with subjects, a large proportion of object anaphora occur at one clause distance from their antecedent ($P = 0.44$, $\sigma = 0.059$), but longer-distance anaphors are comparatively more common. Objects also exhibit a similarly gradual rise in lexicality rates from low to high distances (cross-corpus mean Spearman's rank correlation coefficient $\rho = 0.69$, $\sigma = 0.152$), with the biggest changes found between the low-distance categories. If the antecedent is in the previous clause (i.e. $d = 1$ clause), lexicality rates for objects are substantially lower than in other contexts, roughly on the level of the baseline rate for subjects. However, the threshold at which lexical expressions become the preferred choice (i.e. > 50% incidence) is situated at much lower distances than for subjects, being crossed at $d = 2$ clauses distance for most corpora. The regression model fit to the data and summarized in Table 6.6 confirms these overall impressions: The odds of a lexical expression increase $e^{\beta} = 1.37$ times for each clause unit increase in distance ($p < 0.001$).

The ten corpora do not cluster nearly as tightly as they do for subjects, with English and Nafsan being noticeable outliers. English again charts only a slow rise in lexicality rates at low distances, then a sudden jump at $d \geq 4$ clauses distance. Nafsan shows the inverse picture, with a sharp increase in rates at low distances, which flattens out above $d > 3$ clauses; however, the subsamples at intermediate distance for Nafsan and a number of other corpora are on the small side, which can exacerbate random variation in the data.

6.3.3 | Interaction with humanness

Figure 6.12 shows interaction of anaphoric distance and humanness (see Section 6.1) on the form of subject anaphors, Figure 6.13 on that of objects. In both roles, human referents across corpora are more likely to be expressed non-lexically at higher distances from the antecedent than non-human referents. This effect is most pronounced for subjects, where humans have a lower rate of lexical expression even at very high distances ($d \geq 10$ clauses) than non-humans; for objects the difference seems to disappear above $d \geq 6$ clause units distance. Both human subjects and objects exhibit a much narrower spread of values across corpora. Comparatively greater variation is found

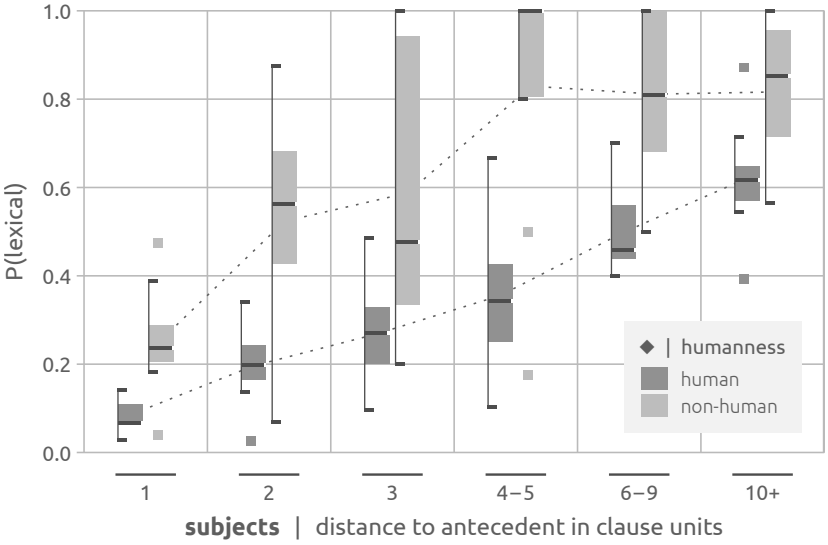


Figure 6.12 | Interaction of the effect of anaphoric distance and humanness on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

among non-humans; especially in subject position this is likely a result of low subsample sizes, since as noted above in Section 6.1, the vast majority of subject mentions are human (cross-corpus mean $P = 0.85$, $\sigma = 0.109$ of subjects vs. $P = 0.33$, $\sigma = 0.120$ of objects).

In Figure 6.12, the two outliers at distances $d = 1$ and $d = 2$, one for non-human and one for human referents, are from the English data. As noted above, English appears largely indifferent to humanness distinctions (Section 6.1), and separately seems to tolerate non-lexical (i.e. in the case of English, pronominal) subjects at longer distances than the other corpora in the sample. But as the interaction between these factors shows, English is in fact insensitive towards humanness at any distance. The other notable outlier is Northern Kurdish, where humanness makes an especially large difference.

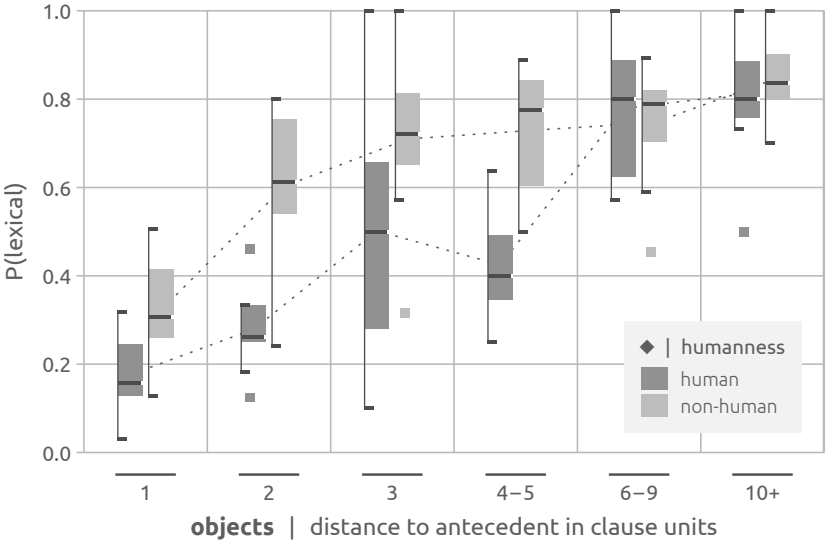


Figure 6.13 | Interaction of the effect of anaphoric distance and humanness on the lexicality of anaphoric objects, by corpus.
The small squares indicate outliers from the central distribution.

6.4 | Local discourse prominence

6.4.1 | Definition and methodological issues

Next we will examine the effect on lexical choice of a referent's prominence in its local context, which we will quantify as the frequency of earlier mentions of that referent. Counted for each anaphor is the number of co-referential mentions (including zero anaphors) in any position (i.e. not limited to just subjects or objects) within a more or less arbitrarily chosen interval of the preceding six clause units, measured starting from the clause containing the anaphor in question. In the example in (69), for instance, three co-referential mentions precede the object anaphor (with the index ⟨0004⟩) at the end.

(69) Teop

a. *Paa huun a suinnae, paa muraka, me paa taverete tahii, …*

	paa	*huun*	*a*	*suin*	*=na*	*=e*
	TAM3	liquid	ART2.SG	body	=3SG.POSS	=3SG.PRON
##	lv	v:pred	ln	np:s	=rn	=rn_pro.h:poss
				0099		0004

	paa	*muraka*	*me*	*paa*	*ta-verete*	*tahii*
	TAM3	soft	and	TAM3	CAUS-turn	saltwater
##	0:s	lv v:pred	## other	0:s	lv v:pred	np:other
	0099			0099		

‘Her body turned into liquid, [it] became soft and turned into saltwater, …’

b. *… me paa rova komana tahii.*

	me	*paa*	*rova*	*koma*	*=n*
	and	TAM3	disappear	inside	=3SG.POSS
##	other	0:s	lv	v:pred	np:l	=rn
		0099				

=a	*tahii*
=ART2.SG	sea
=rn	rn_np:poss
	0100

‘… and disappeared in the sea.’ [mc_teop_iar_0294-0296]

This interval of six clauses, which we will henceforward refer to by the shorthand “recent discourse”, is also used for several other factors discussed later (Sections 6.5 and 6.6). An interval of six clauses contains approximately three to four sentences, which in turn is roughly the length of a paragraph in written English (cf. Section 4.6.1.1). Refer back to Section 4.6.1.1 for a discussion of studies that use paragraphs and other textual units to subdivide discourse. The scope of recent discourse was chosen in part to align with the winsorization threshold for anaphoric distances; see Section 6.3.1.1. A related measure employed in Kibrik et al. (2016) counts the total number of

mentions of a referent, starting from its initial introduction up to the anaphor in question; this measure captures the intermediate point between local discourse prominence (as in this section) and global prominence (as in our measure of protagonisthood, Section 6.2).

For the majority of mentions (cross-corpus mean $P = 0.91$, $\sigma = 0.031$), the frequency of recent co-referential mentions is well below $N \leq 5$, that is, the distribution of frequency values has a long tail similar to that anaphoric distance. For the same reasons stated in that section, mention frequencies have been winsorized at a threshold of $N = 5$; in other words, in the following, frequencies above that threshold are treated as if belonging to the $N = 5$ category.

Lastly, note that certain values of this factor logically entail certain anaphoric distances and vice versa: Since recent discourse is defined as the previous $d = 6$ clause units, the distance to their antecedent is necessarily $d \geq 6$ for all anaphoric mentions with $N = 0$ co-referential mentions in that interval. Conversely, if there is at least one co-referential mention in recent discourse, then anaphoric distance is necessarily $d < 6$.

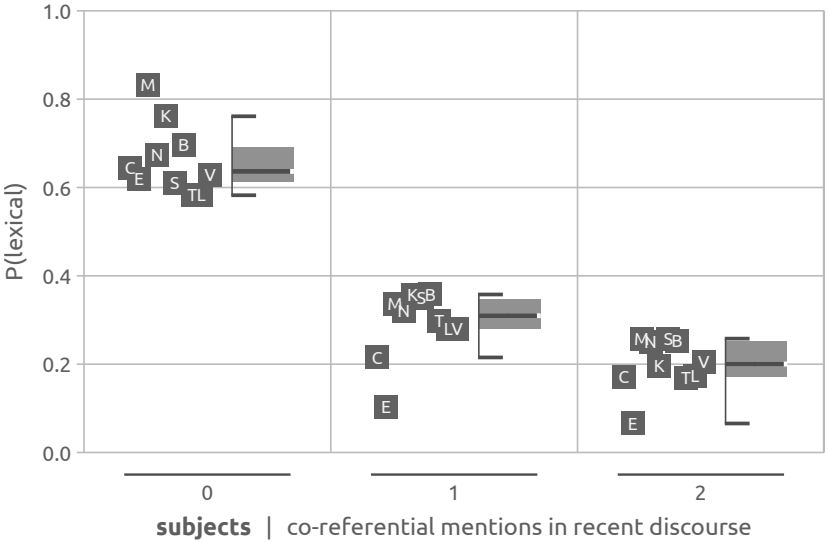
6.4.2 | Lexicality by recent co-referential mentions

6.4.2.1 | Subjects

Figure 6.14 shows the effect of the frequency of recent co-referential mentions on the selection of lexical expressions for subjects.[4] Note, first of all, that a substantial proportion of subject mention have quite high frequencies of recent co-referential mentions, especially at the very high end of the scale. This suggests a high degree of local persistence of referents mentioned in subject position.

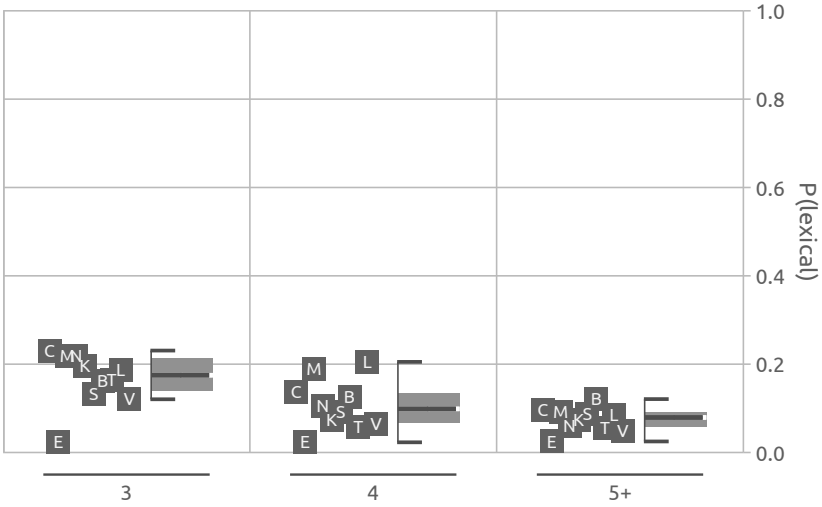
As expected from the association with anaphoric distance and our earlier observations in Section 6.3.2, the majority of subjects in all corpora (cross-corpus mean $P = 0.66$, $\sigma = 0.080$) are lexical if there are no co-referential mentions in the previous $d = 6$ clause units. If there is even one mention in recent discourse, the likelihood of lexical subjects is reduced by over half ($P = 0.29$,

4 Disregarding for now the possible effects of role continuity across clauses, see Section 6.7 further on.



		0			1			2		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	20	31	0.65	17	79	0.22	13	76	0.17
E	English	88	142	0.62	42	411	0.10	22	335	0.07
M	Mandarin	69	83	0.83	46	137	0.34	33	128	0.26
N	Nafsan	29	43	0.67	35	109	0.32	26	104	0.25
K	N. Kurdish	51	67	0.76	30	84	0.36	16	82	0.20
S	S. Dargwa	25	41	0.61	35	100	0.35	25	97	0.26
B	Tabasaran	73	105	0.70	44	123	0.36	27	107	0.25
T	Teop	66	113	0.58	45	151	0.30	23	137	0.17
L	Tulil	53	91	0.58	38	136	0.28	17	98	0.17
V	Vera'a	194	309	0.63	118	421	0.28	83	404	0.21
	totals	668	1025	—	450	1751	—	285	1568	—

Figure 6.14 | Lexicality of anaphoric subjects by the frequency of co-referential mentions in recent discourse.



3			4			5+				
N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	r_{pb}	*p*-val.
15	65	0.23	9	66	0.14	12	124	0.10	−0.24	<0.001
5	217	0.02	3	130	0.02	3	120	0.02	−0.34	<0.001
27	123	0.22	20	106	0.19	13	142	0.09	−0.39	<0.001
27	123	0.22	14	133	0.11	11	186	0.06	−0.35	<0.001
17	87	0.20	7	95	0.07	17	228	0.07	−0.45	<0.001
17	129	0.13	7	76	0.09	7	81	0.09	−0.33	<0.001
20	124	0.16	14	112	0.12	26	215	0.12	−0.38	<0.001
20	122	0.16	6	104	0.06	7	124	0.06	−0.37	<0.001
16	86	0.19	16	78	0.21	11	131	0.08	−0.30	<0.001
47	390	0.12	22	337	0.07	28	569	0.05	−0.40	<0.001
211	1466	—	118	1237	—	135	1920	—	—	—

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	co-ref. ment.	(*0*–5+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.328
speaker	37	0.377

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	1.10	0.093	0.136	0.69	0.491
(A)	co-ref. ment.	* [0, 5]	0.53	–0.633	0.020	–31.84	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–1.504	–0.514	–0.302	–0.186	9.219

d. | correlation of fixed effects

	(intercept)
(A)	–0.270

e. | model evaluation

observations	8967	AIC	7765
model deviance	7757	log-likelihood	–3879
residual d.f.	8963	conditional R^2	0.298
		marginal R^2	0.245

Table 6.7 | Regression model results for the lexicality of anaphoric subjects by the frequency of co-referential mentions in recent discourse, with corpus and speaker as random effects.

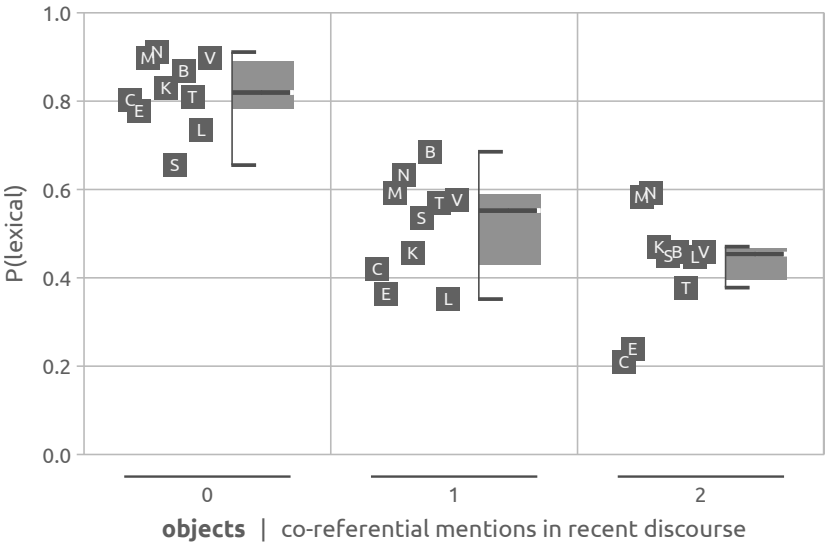
$\sigma = 0.080$). Further co-referential mentions beyond the first have a diminishing, but consistent, effect, with lexical subjects becoming gradually less likely as the number of mentions increases. At the far end of the frequency spectrum (i.e. for $N \geq 5$ co-referential mentions), lexicality rates are exceedingly low in all corpora in the sample ($P = 0.07$, $\sigma = 0.028$). The regression model in Table 6.7 indicates that, averaged out, every additional co-referential mention decreases the odds of a lexical expression in subject position by a factor of $e^{\beta} = 0.53$ ($p < 0.001$), that is roughly by half. But as seen in Figure 6.14, most of this difference is in fact found between $N = 0$ and $N = 1$ co-referential mentions.

Overall, this pattern is reflected in all ten corpora, resulting in a fairly narrow spread of values (random effects intercept $\sigma = 0.328$), with two exceptions. Subjects in English precipitously drop to a much lower rate of lexical expression than most other corpora at $N = 1$ mentions, resulting in a much flatter curve. That is, after the first recent co-referential mention, subjects in English are realized lexically at a much lower rate than those in other corpora, with their trajectories not converging until higher frequencies. This rate is also not majorly affected by additional co-referential mentions beyond the first. Cypriot Greek shows a similarly if not quite as prominently flattened trajectory. Note that both of these corpora are outliers at $N \geq 1$ mentions, but are close to the cross-corpus median for $N = 0$ mentions.

In sum, the picture these data paint is quite distinctive: There exists a strong association between a low frequency of co-referential mentions (i.e. low local prominence) and lexical expressions on the one hand, and a high frequency and non-lexical expressions on the other, for all ten tested corpora, though the strength of the association for $N = 1$ mentions and the trajectory for further mentions differ somewhat for two of the corpora.

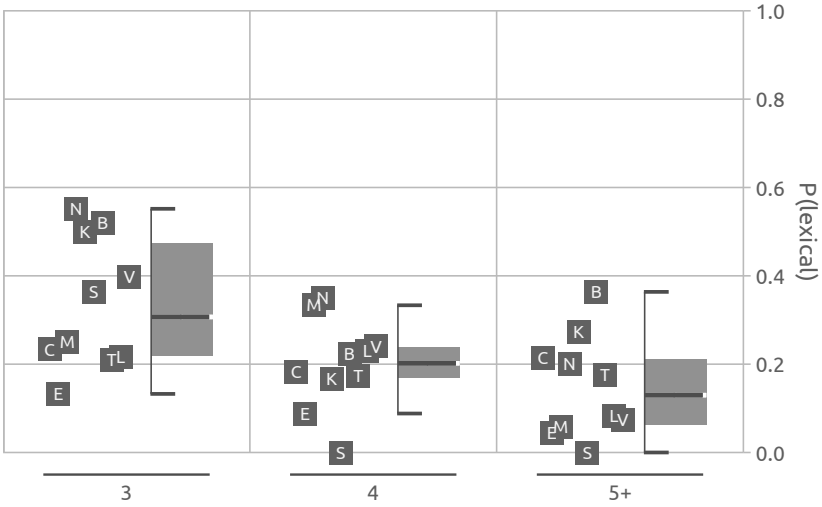
6.4.2.2 | Objects

The corresponding data for direct objects are shown in Figure 6.15. Here we see much the same trajectory as among subjects, in that the frequency of recent co-referential mentions is inversely proportional with lexicality, and while not as pronounced, the largest difference is again found between $N = 0$ and $N = 1$ mentions. However, higher mention frequencies are much rarer than they are for subjects.



		0			1			2		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	49	61	0.80	32	76	0.42	9	43	0.21
E	English	77	99	0.78	95	261	0.36	36	151	0.24
M	Mandarin	62	69	0.90	32	54	0.59	14	24	0.58
N	Nafsan	41	45	0.91	38	60	0.63	29	49	0.59
K	N. Kurdish	49	59	0.83	37	81	0.46	16	34	0.47
S	S. Dargwa	19	29	0.66	30	56	0.54	9	20	0.45
B	Tabasaran	46	53	0.87	61	89	0.69	22	48	0.46
T	Teop	38	47	0.81	29	51	0.57	17	45	0.38
L	Tulil	36	49	0.73	19	54	0.35	21	47	0.45
V	Vera'a	142	158	0.90	109	189	0.58	54	118	0.46
	totals	559	669	—	482	971	—	227	579	—

Figure 6.15 | Lexicality of anaphoric objects by the frequency of co-referential mentions in recent discourse.



3			4			5+				
N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	r_{pb}	*p*-val.
7	30	0.23	2	11	0.18	3	14	0.21	−0.39	<0.001
13	98	0.13	6	68	0.09	2	46	0.04	−0.40	<0.001
6	24	0.25	5	15	0.33	1	17	0.06	−0.53	<0.001
16	29	0.55	7	20	0.35	5	25	0.20	−0.40	<0.001
13	26	0.50	4	24	0.17	9	33	0.27	−0.34	<0.001
4	11	0.36	0	6	0.00	0	3	0.00	−0.31	<0.001
13	25	0.52	4	18	0.22	4	11	0.36	−0.36	<0.001
8	38	0.21	4	23	0.17	9	51	0.18	−0.45	<0.001
8	37	0.22	6	26	0.23	2	24	0.08	−0.37	<0.001
23	58	0.40	12	50	0.24	3	41	0.07	−0.47	<0.001
111	376	—	50	261	—	38	265	—	—	—

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	co-ref. ment.	(*0*–5+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.332
speaker	37	0.481

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	3.15	1.148	0.156	7.34	<0.001
(A)	co-ref. ment.	* [0, 5]	0.52	–0.662	0.032	–20.68	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–2.576	–0.795	–0.294	0.736	4.508

d. | correlation of fixed effects

	(intercept)
(A)	–0.333

e. | model evaluation

observations	3121	AIC	3607
model deviance	3599	log-likelihood	–1800
residual d.f.	3117	conditional R^2	0.294
		marginal R^2	0.221

Table 6.8 | Regression model results for the lexicality of anaphoric objects by the frequency of co-referential mentions in recent discourse, with corpus and speaker as random effects.

At least in part due to relatively small subsample sizes, there is greater variation between corpora for objects, that is, unlike subjects, where most corpora cluster together into a relative stable pattern, here the spread of values is somewhat wider (cross-corpus mean point-biserial correlation coefficient $r_{pb} = -0.40$, $\sigma = 0.065$). The regression model in Table 6.8 indicates that a one-unit increase in mention frequency still roughly halves the odds of a lexical objects ($e^{\beta} = 0.52$ with $p < 0.001$), but that there is a noticeable degree of variation between speakers (random effect intercept $\sigma = 0.481$) in particular.

6.5 | Priming from bridging relations

6.5.1 | Definition and methodological issues

Besides co-referential mentions, a referent's accessibility is expected to also be affected indirectly by mentions of referents related to it, that is those that are not fully co-referential, but overlap the referent in question in referential space, in essence priming expectation of its retrieval. Thus, the higher the local frequency of thusly related mentions, the lower the rate of lexical expression should be. There are three basic types of mereological relations between referents captured by the annotations, described in more detail in Section 3.3.4.2 above. A referent is deemed related to the referent in question (i) if it contains it via split antecedence (*the mother and father*, *the parents*); (ii) if it is being contained by it via partial co-reference (*the children*, *the daughter*); or (iii) if it is in a part-whole relationship with it (*the body*, *the head*).

While it would be possible to adapt the measurements of anaphoric distance to take bridging relations into account by altering the selection criteria for antecedents, it is unclear if and how fully and partially co-referential antecedents would be graded relative to one another in such a system, and whether measurements should be made from any mention of a set member in the case of split antecedence. As such, we will instead resort to simply counting mentions of related referents in recent discourse, that is over a stretch of $d = 6$ clause units, counting backwards from the anaphor in question. For the sake of consistency, this interval of 'recent discourse' is the same for all factors in this study that evaluate the local properties of discourse in some way (Sections 6.4 and 6.6).

Since related mentions are actually relatively rare, the counts for subjects are winsorized at a threshold of $N \geq 2$; that is, frequencies above that threshold are treated as occurring at it – similar to anaphoric distances above. Frequencies for object mentions are likewise winsorized, but at an even lower threshold of $N \geq 1$, in essence reducing this variable to a binary one (i.e. has there been a mention of a related referent in recent discourse?), though it continues to be coded as scalar for the purposes of statistical modeling. As mentioned above, the reasoning behind this decision is tied to small subsample sizes at higher mention frequencies: For the the majority of anaphors ($P = 0.76$), there are no related referents mentioned in the preceding $d = 6$ clauses. Consequently, mentions with recent bridging relations are relatively uncommon, less so for subject mentions ($P = 0.27$ with $N \geq 1$), but especially so for object mentions ($P = 0.15$).[5] The maximum number of related mentions for anaphors in any position is $N = 5$.

Example (70) is a case of a relatively long-distance anaphor for which a pronominal form is chosen due to the recent mention of another referent related via partial co-reference. Here the actual distance to most recent coreferential mention for the referent with the index ⟨0025⟩, expressed via the pronoun *he* in (70b), is $d = 41$ clauses, well above the threshold at which lexical forms become the preferred form of reference in all corpora in the sample, including English (see Section 6.3). But since this referent is contained in that expressed by *we* (⟨0034⟩), which is mentioned numerous times in the preceding discourse, the speaker nevertheless selects a pronoun for its expression.

(70) English

a. *We used to [...] unyoke at four [...].*

	we	*used*	*to*	*unyoke*	*at*	*four*
	1PL	used	to	unyoke.INF	at	four
##	pro.1:s	lv_aux	lv	v:pred	adp	np:other
	0034					

‘We used to [...] unyoke at four [...].’

[mc_english_kent03_0044-0045]

5 It is unclear how come recent related mentions are so much less frequent for objects than for subjects. A possible explanation might invoke the overall lower frequency of object anaphors, which hence tend to be more dispersed compared to subject anaphors, and the lower narrative persistence of non-human entities, which make up a large proportion of object anaphors.

b. *And then he would stop there till six o'clock.*

	and	*then*	*he*	*would*	*stop*	*there*	*till*
	and	then	3SG.M	would	stop.INF	there	until
##	other	other	pro.h:s	lv_aux	v:pred	other:l	adp
			0025				

six	*o'clock*
six	o'clock
np:other	rn

'And then he would stop there till six o'clock.'

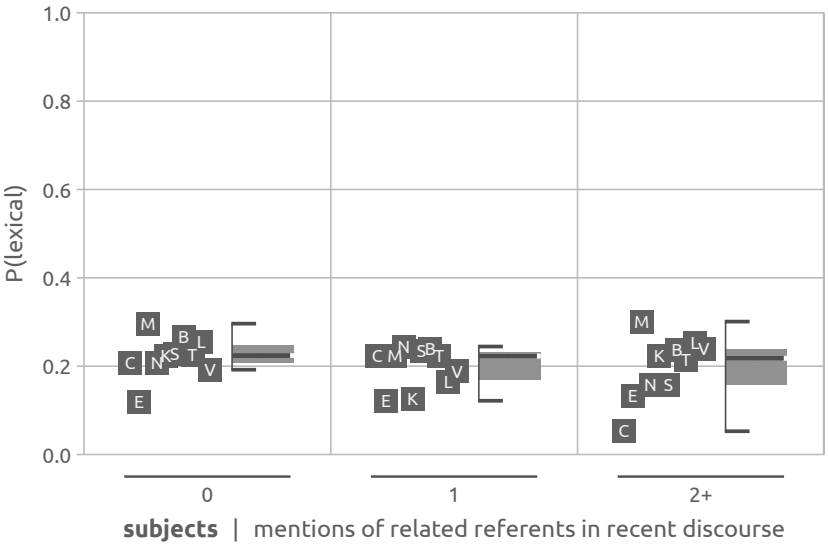
[mc_english_kent03_0047]

6.5.2 | Lexicality by recent related mentions

6.5.2.1 | Subjects

As Figure 6.16 shows for subject anaphors, the cross-corpus effect of related mentions on form of expression is marginal (cross-corpus mean rank correlation coefficient $r_{pb} = -0.02$; $\sigma = 0.037$). Rates of lexical expression at mention frequencies $N \geq 1$ appear to be either slightly lower than or no more than roughly equal to the rates at $N = 0$. The only corpora with a slightly more pronounced drop in lexicality rate are in Cypriot Greek ($r_{pb} = -0.10$) and Sanzhi Dargwa ($r_{pb} = -0.04$), and there only for $N \geq 2$ related mentions. Others, such as English ($r_{pb} = 0.01$) and Teop ($r_{pb} = -0.01$), exhibit virtually no association. There exists some small degree of variation between corpora (random effect intercept $\sigma = 0.169$), but the tenability of any identifiable pattern is uncertain due to the relatively small subsample sizes in some of the groups.

What is perhaps most noteworthy about these data, however, is that this ostensibly weak association appears to run counter to the effect observed for a similar measure, the frequency of competing (i.e. unrelated and not coreferential) mentions in recent discourse, which we will examine independently later in Section 6.6. There, frequencies above $n \geq 1$ and beyond show a steady increase in the rate of lexicality for subject anaphors, indicating that the



		0			1			2+		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	78	376	0.21	6	27	0.22	2	38	0.05
E	English	124	1049	0.12	18	148	0.12	21	158	0.13
M	Mandarin	157	530	0.30	17	76	0.22	34	113	0.30
N	Nafsan	116	558	0.21	11	45	0.24	15	95	0.16
K	N. Kurdish	115	515	0.22	7	56	0.12	16	72	0.22
S	S. Dargwa	96	422	0.23	12	51	0.24	8	51	0.16
B	Tabasaran	161	605	0.27	21	88	0.24	22	93	0.24
T	Teop	117	521	0.22	19	85	0.22	31	145	0.21
L	Tulil	102	402	0.25	11	67	0.16	38	151	0.25
V	Vera'a	305	1589	0.19	50	266	0.19	137	575	0.24
	totals	1371	6567	—	172	909	—	324	1491	—

Figure 6.16 | Lexicality of anaphoric subjects by the frequency of mentions of related referents in recent discourse.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	related ment.	(*0*–2+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.169
speaker	37	0.353

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.27	–1.315	0.090	–14.63	<0.001
(A)	related ment.	* [0, 2]	0.99	–0.012	0.036	–0.34	0.737

c. | scaled residuals

min.	lower	median	upper	max.
–0.702	–0.549	–0.478	–0.319	3.175

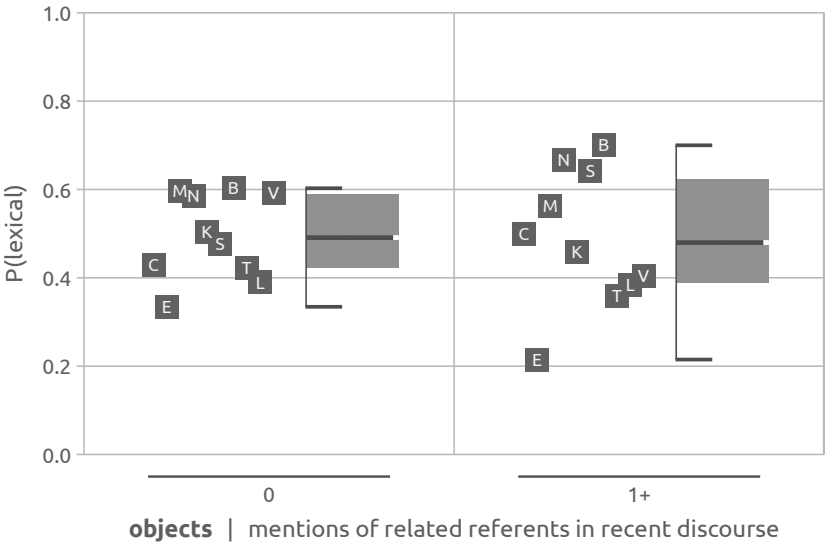
d. | correlation of fixed effects

	(intercept)
(A)	–0.175

e. | model evaluation

observations	8967	AIC	9046
model deviance	9038	log-likelihood	–4519
residual d.f.	8963	conditional R^2	0.045
		marginal R^2	0.000

Table 6.9 | Regression model results for the lexicality of anaphoric subjects by the frequency of mentions of related referents in recent discourse, with corpus and speaker as random effects.



	corpus	0 N(lex)	0 N(all)	0 P(lex)	1+ N(lex)	1+ N(all)	1+ P(lex)	φ	*p*-val.
C	C. Greek	95	221	0.43	7	14	0.50	0.03	0.608
E	English	206	616	0.33	23	107	0.21	0.09	0.014
M	Mandarin	102	171	0.60	18	32	0.56	0.03	0.720
N	Nafsan	116	198	0.59	20	30	0.67	0.06	0.401
K	N. Kurdish	111	220	0.50	17	37	0.46	0.03	0.612
S	S. Dargwa	53	111	0.48	9	14	0.64	0.10	0.244
B	Tabasaran	129	214	0.60	21	30	0.70	0.07	0.306
T	Teop	91	216	0.42	14	39	0.36	0.05	0.467
L	Tulil	74	190	0.39	18	47	0.38	0.01	0.935
V	Vera'a	299	505	0.59	44	109	0.40	0.14	<0.001
	totals	1276	2662	—	191	459	—	—	—

Figure 6.17 | Lexicality of anaphoric objects by the frequency of mentions of related referents in recent discourse.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	related ment.	(*0*–1+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.276
speaker	37	0.574

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	1.06	0.062	0.145	0.43	0.668
(A)	related ment.	* [0, 1]	0.76	–0.276	0.109	–2.54	0.011

c. | scaled residuals

min.	lower	median	upper	max.
–1.807	–0.888	–0.610	0.984	1.781

d. | correlation of fixed effects

	(intercept)
(A)	–0.122

e. | model evaluation

observations	3121	AIC	4149
model deviance	4141	log-likelihood	–2071
residual d.f.	3117	conditional R^2	0.112
		marginal R^2	0.003

Table 6.10 | Regression model results for the lexicality of anaphoric objects by the frequency of mentions of related referents in recent discourse, with corpus and speaker as random effects.

increased cognitive load associated with juggling multiple concurrently active referents has some influence on referential choice. As we will see in Section 6.6, more lexical expressions are chosen as a compensatory strategy as the number of candidate antecedents in local discourse increases (Arnold 2010: 195; see Section 2.4.2.4), so as to more efficiently disambiguate between them, though the effect is not strongly pronounced. From the observable lack of any such effect for related referents, we might surmise that referents which share a bridging relation with the referent in question might contribute less to cognitive load. Instead, the presence of semantic links between referents might facilitate identification of the intended referent, hence easing the retrieval process – if only minimally – which in turn might explain the differences in rates of lexical expression between the two measures.

6.5.2.2 | Objects

As mentioned above, object anaphors with recent related mentions are of considerable rarity, accounting for just $P = 0.15$ of object mentions in the sample. Unfortunately, this renders any observations made on the basis of the data shown in Figure 6.17 and Table 6.10 highly tenuous. From a cross-linguistic perspective, the presence or absence of mentions of related differences appears to make little difference for lexical choice (cross-corpus mean $r_{pb} = -0.01$; $\sigma = 0.076$), similar to what we have observed for subjects above, but with noticeably greater dispersion among corpora (random effect intercept $\sigma = 0.276$) as well as speakers ($\sigma = 0.574$).

6.6 | Local competition

6.6.1 | Definition and methodological issues

The previous two sections have discussed the effects of co-referential mentions (Section 6.4) and mentions of referents with overlapping referential spaces (Section 6.5) in recent discourse. The third and last class of referents, then, are those that are neither co-referential nor directly related (see Section 2.4.2.4). Unlike the first two, which amplify the accessibility of the referent in question, mentions of unrelated referents are claimed to have an adversary effect, both by making identification of the intended antecedent more

difficult (Ariel 1990; 2001) as well as by stressing constraints on short-term memory capacity (Arnold 2010: 195). According to accessibility theory (Ariel 1990), the higher the number of mentions of other referents in recent discourse, the greater the competition between candidate antecedents, and hence the greater the need for speakers to disambiguate between them via use of more explicit referring expressions.

The definition of competing referents used here is somewhat crude, as it simply includes all non-co-referential referents. As Givón (1983a: 14) notes, ambiguity between candidate antecedents can only arise where referents share certain distinctive semantic properties, such as their "humanity, agentivity or semantic plausibility as object or subject", which also are likely to differ from language to language. However, maintaining lists of all potentially relevant properties for (every mention of) every referent and operationalizing language-specific rules for which combinations of properties interfere with one another is impractical at this stage.[6] While the simpler measure of competition used hence cannot capture instances of semantic ambiguity, it can nevertheless serve as a measure of overall memory load as mentioned above.

As such, we here count for each anaphor the number of mentions of referents that are neither co-referential nor share a bridging relation to the referent in question in recent discourse, which as before is defined as the previous six clauses, counted backwards from the anaphor in question. Since frequencies of competing mentions above $N \geq 6$ become increasingly rare, frequencies are winsorized at that threshold. In this example from English, the anaphor in (71a) (with referent index ⟨0090⟩) is preceded by over a dozen mentions of other referents, but is nevertheless realized as a pronoun:

(71) English

a. *He said, …*

	he	*said*
	3SG.M	say.PST
##	pro.h:s_ds 0090	v:pred

'He said, …'

6 Gipper (2016: 150), in investigating referential choices in Yurakaré, likewise notes difficulty in finding a suitable operationalization for competition.

b. *You give'im what money you've got and tell'im you'll pay the rest when you've saved it up. I went down there and I asked him for this pair of boots, and he wouldn't hear of it. Well, I went back home again, up to where I lodged.*

c. *And he said, What's the matter with you, boy?*

	and	*he*	*said*	*what*	*='s*
	and	3SG.M	say.PST	what	=be.PRS.3SG
##	other	pro.h:s_ds 0090	v:pred	##ds other:pred	=cop

the	*matter*	*with*	*you*	*boy*
the	matter	with	2SG.OBL	boy
ln_det	np:s	rn_adp	rn_pro.2 0000	np.h:voc 0000

'And he said, What's the matter with you, boy?'

[mc_english_kent03_0096-0099]

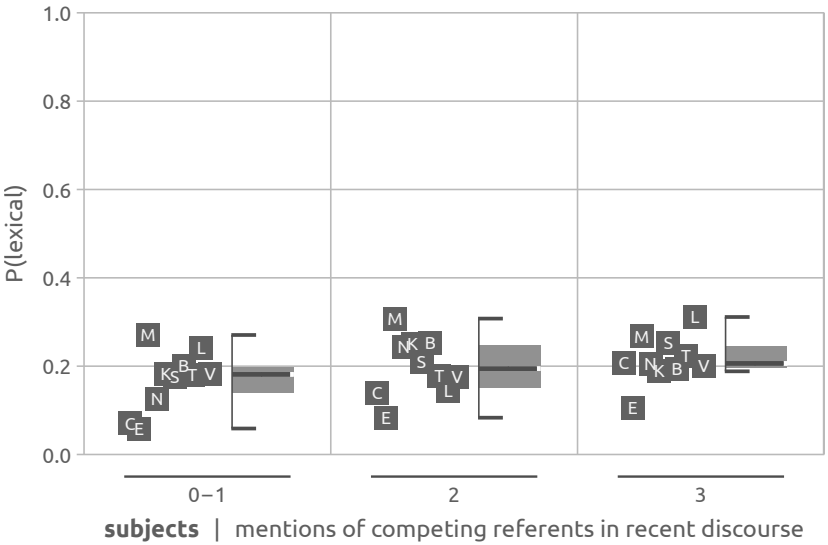
The measure of competition used here is superficially related to notion of referent pressure in Du Bois (2003b), which is there calculated as the total number of unique referents mentioned a text divided by the number of clauses. Though more complicated to implement, measures limited to a specific window of discourse, such as done here, are likely to offer better representations of referent activation and memory states as speakers can hardly be assumed to possess awareness of the full extent of what they have and are going to have said eventually. It can be argued that in general, speakers' decisions, be they about referential choices or other aspects of discourse management, should ideally be modelled on the basis of evaluations of the local discourse at the time of making said decisions, rather than of larger text-sized chunks of discourse.

6.6.2 | Lexicality by recent competing mentions

For the purposes of visualization only, frequencies of $n = 0$ and $n = 1$ mentions are grouped together into a single category. The former in particular has a relatively low incidence rate as few mentions occur in complete isolation from other, unrelated referents. As always, statistics calculated from the data, such as the regression models and any analyses later on, use the ungrouped values.

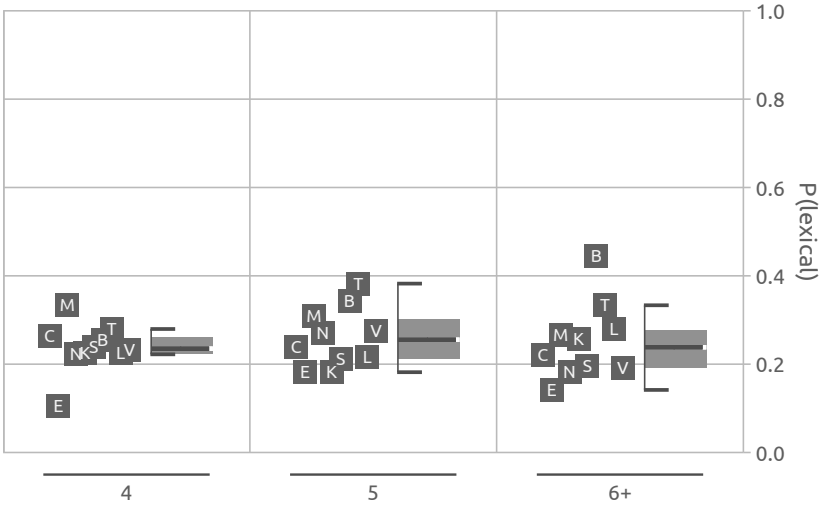
6.6.2.1 | Subjects

Figure 6.18 shows that competition between candidate antecedents has only a weak effect on the selection of lexical expressions for subjects, contrary to the claims of accessibility theory (Ariel 1990). The cross-corpus mean point-biserial correlation coefficient for lexicality and number of competing mentions is $r_{pb} = 0.07$ ($\sigma = 0.052$), indicating a near absence of an association. The three corpora with the strongest associations are Tabasaran ($r_{pb} = 0.14$ with $p < 0.14$), Cypriot Greek ($r_{pb} = 0.13$ with $p < 0.13$), and Teop ($r_{pb} = 0.12$ with $p < 0.12$), and even for these there is little tangible effect. The regression model fit to the data, summarized in Table 6.11, likewise indicates that the odds of a lexical subject increase $e^{\beta} = 1.12$ times ($p < 0.001$) for every competing mention in recent discourse, relative to an intercept of $e^{\beta} = 0.19$, which is only a marginal increase. These data suggest that lexical expressions are no more or less likely in subject position irrespective of how many other referents are active at the same time, that is, there is no discernible effect of competition on the lexicality of subjects. It should be noted, however, that especially for subjects there is a potential negative correlation between the number of competing referents and the frequency of the referent in question: If a referent is mentioned continuously in the examination window, it thereby 'blocks' involvement of other referents in the position or positions it occupies.



	corpus	0–1 N(lex)	0–1 N(all)	0–1 P(lex)	2 N(lex)	2 N(all)	2 P(lex)	3 N(lex)	3 N(all)	3 P(lex)
C	C. Greek	4	57	0.07	10	72	0.14	22	106	0.21
E	English	6	102	0.06	13	156	0.08	29	274	0.11
M	Mandarin	23	85	0.27	28	91	0.31	35	131	0.27
N	Nafsan	20	158	0.13	40	165	0.24	34	166	0.20
K	N. Kurdish	14	77	0.18	30	120	0.25	29	154	0.19
S	S. Dargwa	13	74	0.18	25	119	0.21	38	150	0.25
B	Tabasaran	30	149	0.20	34	135	0.25	35	180	0.19
T	Teop	38	210	0.18	33	185	0.18	39	175	0.22
L	Tulil	28	116	0.24	9	63	0.14	33	106	0.31
V	Vera'a	113	619	0.18	87	493	0.18	110	547	0.20
	totals	289	1647	—	309	1599	—	404	1989	—

Figure 6.18 | Lexicality of anaphoric subjects by the frequency of mentions of competing referents in recent discourse.



4			5			6+				
N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	r_{pb}	*p*-val.
20	76	0.26	17	71	0.24	13	59	0.22	0.13	0.006
28	267	0.10	37	203	0.18	50	353	0.14	0.09	0.001
39	117	0.33	31	100	0.31	52	195	0.27	0.00	0.935
26	117	0.22	16	59	0.27	6	33	0.18	0.07	0.066
30	133	0.23	14	77	0.18	21	82	0.26	0.01	0.764
21	88	0.24	11	52	0.21	8	41	0.20	0.03	0.510
38	149	0.26	34	99	0.34	33	74	0.45	0.14	<0.001
26	93	0.28	13	34	0.38	18	54	0.33	0.12	0.001
35	156	0.22	14	65	0.22	32	114	0.28	0.02	0.598
97	419	0.23	58	211	0.27	27	141	0.19	0.05	0.011
360	1615	—	245	971	—	260	1146	—	—	—

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	comp. ment.	(*0*–6+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.181
speaker	37	0.384

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.19	–1.682	0.109	–15.39	<0.001
(A)	comp. ment.	* [0, 6]	1.12	0.113	0.016	6.82	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–0.820	–0.544	–0.461	–0.340	3.680

d. | correlation of fixed effects

	(intercept)
(A)	–0.497

e. | model evaluation

observations	8967	AIC	8999
model deviance	8991	log-likelihood	–4496
residual d.f.	8963	conditional R^2	0.062
		marginal R^2	0.011

Table 6.11 | Regression model results for the lexicality of anaphoric subjects by the frequency of mentions of competing referents in recent discourse, with corpus and speaker as random effects.

6.6.2.2 | Objects

For object anaphors, Figure 6.19 shows much the same picture as seen for subjects above. Overall, the number of competing mentions in the previous six clauses has only a weak, if positive, effect on lexicality rates (point-biserial correlation coefficient $r_{pb} = 0.11$; $\sigma = 0.081$). Table 6.12 indicates that in a simple single-factor regression model fit to the data, the odds of lexical object increase only $e^{\beta} = 1.15$-fold ($p < 0.001$) for each additional competing mention. Variation between corpora (random effect intercept $\sigma = 0.294$) and especially speakers ($\sigma = 0.560$) is greater than for subjects, as it generally is for objects for the factors tested in this study. There are three corpora for which the association is stronger than for the rest, this being Mandarin ($r_{pb} = 0.23$ with $p < 0.23$), Teop ($r_{pb} = 0.21$ with $p < 0.21$), and Vera'a ($r_{pb} = 0.21$ with $p < 0.21$).

6.7 | Role of the antecedent

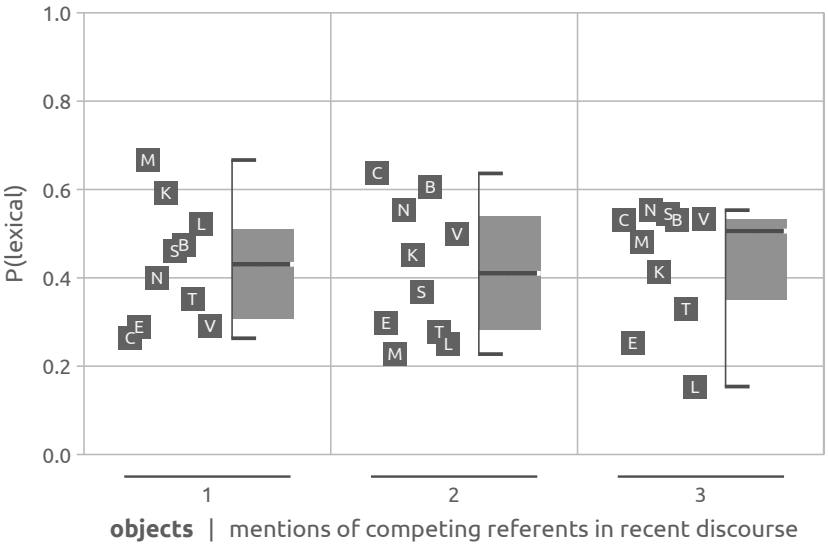
6.7.1 | Definition and methodological issues

Next we will examine the effect that the grammatical role of the antecedent has on referential choices. Where similar studies, such as Kibrik et al. (2016), employ relatively differentiated classification systems for grammatical role, for the present study the complexity of the categorization is deliberately kept to a minimum, distinguishing only whether the immediate antecedent is in

a. subject position
b. object position, or
c. in some other position.

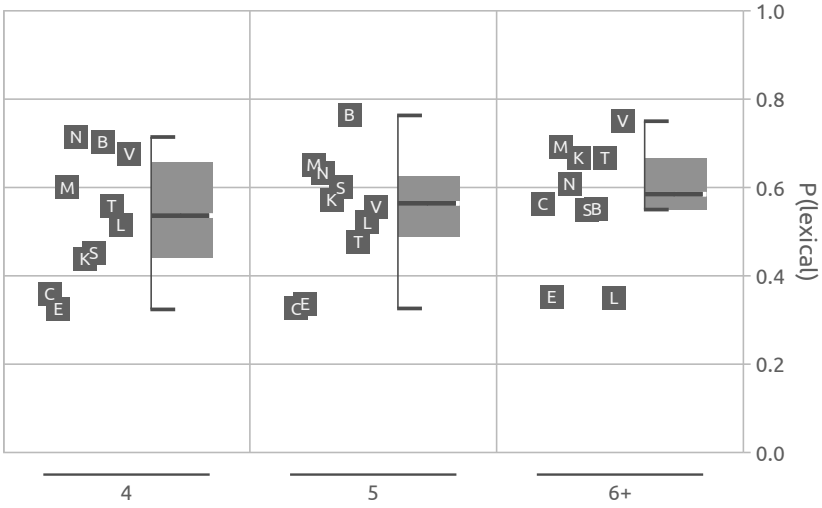
The classification of subjects and objects is based on the definitions given in Section 4.3. The third category, 'other', captures mentions in all other syntactic positions, including as oblique arguments, as adjuncts, as well as NP-internal constituents such as possessive expressions and adpositional modifiers, and others. For an elaboration of the methods used in the identification of antecedents, refer to Section 4.6.2 above.

Before moving on to examining the effects of antecedent role on choice of form in Section 6.7.3, the next section will briefly touch on the association between the role of an anaphor and the role of its antecedent, and how this association informs the selection of lexical expressions.



		1			2			3		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	5	19	0.26	7	11	0.64	25	47	0.53
E	English	11	38	0.29	22	74	0.30	34	135	0.25
M	Mandarin	4	6	0.67	5	22	0.23	13	27	0.48
N	Nafsan	8	20	0.40	26	47	0.55	26	47	0.55
K	N. Kurdish	16	27	0.59	19	42	0.45	24	58	0.41
S	S. Dargwa	6	13	0.46	7	19	0.37	12	22	0.55
B	Tabasaran	9	19	0.47	23	38	0.61	35	66	0.53
T	Teop	12	34	0.35	15	54	0.28	24	73	0.33
L	Tulil	12	23	0.52	3	12	0.25	6	39	0.15
V	Vera'a	18	62	0.29	59	118	0.50	79	148	0.53
	totals	101	261	—	186	437	—	278	662	—

Figure 6.19 | Lexicality of anaphoric objects by the frequency of mentions of competing referents in recent discourse.



4			5			6+				
N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	r_{pb}	*p*-val.
23	64	0.36	15	46	0.33	27	48	0.56	0.04	0.496
46	142	0.32	37	110	0.34	79	224	0.35	0.06	0.125
18	30	0.60	26	40	0.65	54	78	0.69	0.23	0.001
40	56	0.71	19	30	0.63	17	28	0.61	0.11	0.101
25	57	0.44	28	49	0.57	16	24	0.67	0.05	0.447
14	31	0.45	12	20	0.60	11	20	0.55	0.07	0.468
38	54	0.70	29	38	0.76	16	29	0.55	0.08	0.186
24	43	0.56	10	21	0.48	20	30	0.67	0.21	0.001
18	35	0.51	25	48	0.52	28	80	0.35	0.00	0.994
96	142	0.68	49	88	0.56	42	56	0.75	0.21	<0.001
342	654	—	250	490	—	310	617	—	—	—

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	comp. ment.	(*0*–6+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.294
speaker	37	0.560

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.61	–0.487	0.172	–2.82	0.005
(A)	comp. ment.	* [0, 6]	1.15	0.136	0.025	5.52	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–1.957	–0.877	–0.599	0.950	1.986

d. | correlation of fixed effects

	(intercept)
(A)	–0.530

e. | model evaluation

observations	3121	AIC	4125
model deviance	4117	log-likelihood	–2058
residual d.f.	3117	conditional R^2	0.120
		marginal R^2	0.013

Table 6.12 | Regression model results for the lexicality of anaphoric objects by the frequency of mentions of competing referents in recent discourse, with corpus and speaker as random effects.

6.7.2 | Role persistence

For the role of a given antecedent, Figure 6.20 shows the mean proportions of the roles of the subsequent anaphor across corpora. The roles of antecedent and anaphor are not randomly distributed; rather, there is a substantial degree of role persistence. The majority of anaphors in subject and object position have antecedents in same position (and similarly for non-core roles), and transfer between roles from one mention of a referent to another is the less common case. From one mention to the next, almost three quarters of subjects stay subjects, and only about a fifth is moved outside the clausal core into other positions. Notably, in all corpora in the sample, subject-to-object transfer is quite rare (cross-corpus $m = 0.08$, $\sigma = 0.019$), as is other-to-object transfer ($m = 0.14$, $\sigma = 0.025$), indicating a general dispreference for shifting into object position. Among objects, about half persist as objects from mention to mention, while about a quarter are promoted to subject, a noticeably higher proportion than is moved the opposite direction. A similar pattern is found for mentions in oblique and other roles, that is those that are neither subjects nor objects: The majority stay in some position outside the clausal core, a third are raised to subject position, and less than a fifth become objects. (72) exemplifies the most common case, subject-to-subject persistence:

(72) Cypriot Greek

a. *Aniksen ta dhixtia tu, …*

	aniksen	*ta*	*dhixtia*	*tu*
	open.PST.PFV.3SG	DEF.PL.N	net.PL	3SG.M.POSS
##	0.h:a	v:pred	ln	np:p	rn_pro.h:poss
	0002			0004	0002

‘(He) spread his nets, …’

b. *esiren ta mes stin thalassa, …*

	esiren	*=ta*	*mes*	*stin*	*thalassa*
	throw.PST.PFV.3SG	=3PL.N.ACC	in	to	sea
##	0.h:a	v:pred	=pro:p	adp	ln	np:g
	0002		0004			

‘(he) threw them into the sea, …’

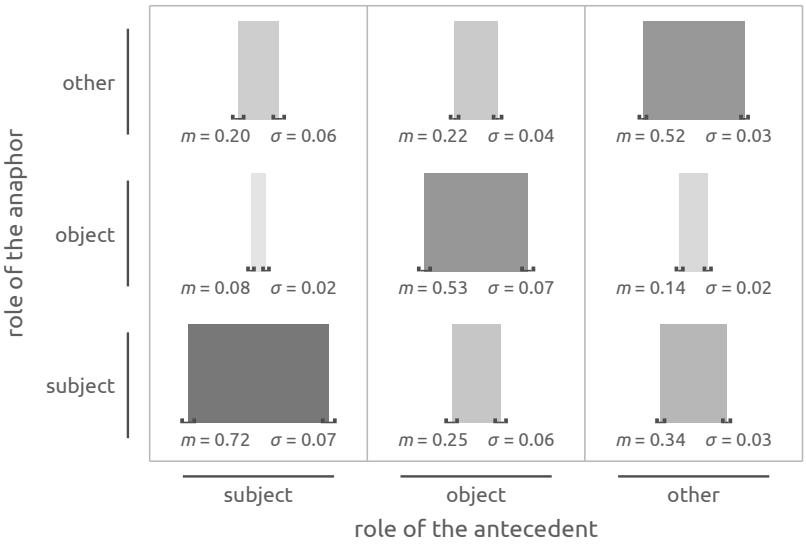


Figure 6.20 | Areal mosaic plot of the mean rate of role changes from antecedent to anaphor across corpora.

The width of each box indicates the relative proportions of anaphor roles from the perspective of the antecedent; the braces underneath the boxes show the 95% confidence intervals around the mean (bootstrapped with $n = 5000$ replicates, BCa).

c. *etavrisen mian volan, …*

	etavrisen	*mian*	*volan*
	pull.PST.PFV.3SG	one.F	time
## 0.h:s 0002	v:pred	ln	np:other

‘(he) pulled once, …’ [mc_cypgreek_psarin_0023]

d. *efkalen kamboson.*

	efkalen	*kamboson*
	take_out.PST.PFV.3SG	many
## 0.h:a 0002	v:pred	other:p

‘(he) took out many [fish].’ [mc_cypgreek_psarin_0023]

Conversely, subject-to-object transfers as in (73) is comparatively rare:

(73) Nafsan

a. *Kori ifit paakor.*

kori	*i=*	*fit*	*paakor*
dog	3S.RS=	run	arrive
## pro:s 0007	=lv	v:pred	rv

‘The dog came running.’

b. *Imtaki kori.*

	i=	*mtak-ki*	*kori*
	3S.RS=	fear-TR	dog
## 0.d:a 0002	=lv	v:pred	np:p 0007

‘(He) was scared of the dog.’ [mc_nafsan_kori_0032]

Notably, there is fairly limited variation between corpora in this regard, which suggests that patterns of role persistence are quite similar cross-linguistically.

In sum, successive mentions of a referent in the same syntactic position are fairly common; for subjects, they constitute a large majority of cases. This coupling of co-reference with syntactic continuity is a significant contributor to the overall coherence of discourse (Givón 1983a, see also Grosz et al. 1983), in essence facilitating comprehension by associating certain syntactic positions – particularly subjects, but to a lesser degree also other roles – with known and narratively central information. This is especially apparent in cases where the antecedent is located in the previous clause, that is in anaphoric same-role chains (Givón 2017), which we will look at later in Section 6.7.4 below. In contexts with role persistence from antecedent to anaphor, we would expect substantially lower rates of lexical expression than in other contexts. As we will see in the next sections, this expectation is borne out for subjects irrespective of distance to the antecedent, and for both subjects and objects in same-role chains.

6.7.3 | Lexicality by antecedent role

6.7.3.1 | Subjects

As discussed above, Figure 6.21 shows that the vast majority of subject anaphors stay subjects on subsequent mentions (cross-corpus mean $P = 0.72$, $\sigma = 0.074$). Subject anaphors with subjects antecedents are highly unlikely to be lexical in all corpora in the sample ($P = 0.17$; $\sigma = 0.030$). Instead, rates of lexical expression are higher if the immediate antecedent was not a subject, that is either an object ($P = 0.39$; $\sigma = 0.146$) or a mention in some other position ($P = 0.31$; $\sigma = 0.31$). While the association between non-subject antecedents and higher lexicality rates is present in almost all corpora, the magnitude of the effect varies substantially between corpora; part of this variation can be explained by relatively small subsample sizes, since as mentioned above, subject anaphors with non-subject antecedents are relatively uncommon. The single notable outlier from the cross-corpus pattern is once again English, where the role of the antecedent has little appreciable effect on the form of subject anaphors (for subject vs. non-subject antecedents, $\varphi = 0.06$ with $p < 0.038$). Other corpora with comparatively small effects are Cypriot Greek ($\varphi = 0.09$ with $p < 0.049$) and Northern Kurdish ($\varphi = 0.12$ with $p < 0.003$).

The logistic regression model summarized in Table 6.13 provides further evidence to support these impressions: Relative to an antecedent in subject

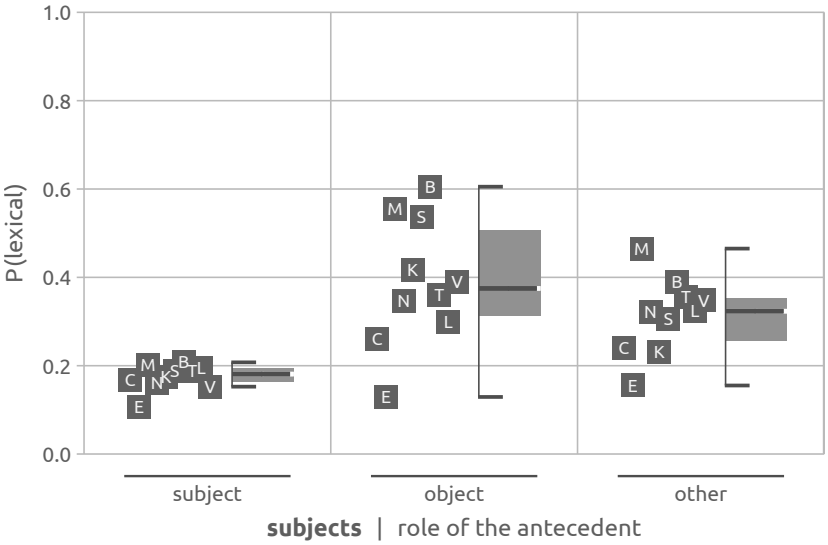
position, which yields lexical subjects at a rate of $e^{\beta} = 0.20$, anaphors with object and other antecedents respectively have $e^{\beta} = 2.93$ ($p < 0.001$) and $e^{\beta} = 2.32$ ($p < 0.001$) times higher odds of being expressed lexically. The standard deviation among the intercepts of the random effect for corpora, $\sigma = 0.222$, is towards the low end of the spectrum of variability encountered among the factors discussed so far, a consequence of the high frequency of the large and noticeably homogeneous subject-antecedent group. Variation between speakers, on the other end, is somewhat higher ($\sigma = 0.369$), but still comparatively limited.

Since functional continuity from one co-referential mention to the next is the norm for subjects, as noted above, this establishes a high degree of discourse coherence (Givón 1983a; Givón 2017). In other words, the subject of the previous clause is expected to re-occur as the subject of the next clause, and this effect, and the concomitant use of reduced forms is likely to strengthen as the length of the anaphoric chain increases. Since promotions from object and other positions to subject are comparativey rare and hence break the expected pattern of same-subject chaining, they tend to require more explicit forms to ensure correct identification of the intended referent, though as noted above, how pronounced this requirement is appears to differ from corpus to corpus.

6.7.3.2 | Objects

The data for objects summarized in Figure 6.22 show a markedly different picture from that for subjects. While, as noted above, objects also exhibit a fairly high degree of role persistence (cross-corpus $P = 0.53$, $\sigma = 0.066$), this evidently does not bear on the rate at which subsequent mentions are realized lexically. In most corpora in the sample, object anaphors with antecedents in subject position are less likely to be lexical than those with object antecedents, though the difference is not large. Lexical objects are most likely with antecedents outside the clausal core in all corpora but two, Nafsan and Tulil, where rates tie with those for object antecedents. The latter notably shows only a marginal association between lexicality and antecedent role for objects; objects in English, on the other hand, follow the general cross-corpus trend.

The regression model fit to the data summarized in Table 6.14 reflects the heterogeneous distribution of the corpora (random effect intercept $\sigma = 0.293$, and further shines light on a substantial degree of inter-speaker variability



		subject			object			other		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	49	291	0.17	12	46	0.26	25	104	0.24
E	English	91	856	0.11	27	209	0.13	45	290	0.16
M	Mandarin	103	509	0.20	45	81	0.56	60	129	0.47
N	Nafsan	85	527	0.16	27	78	0.35	30	93	0.32
K	N. Kurdish	67	384	0.17	25	60	0.42	46	199	0.23
S	S. Dargwa	82	434	0.19	15	28	0.54	19	62	0.31
B	Tabasaran	126	607	0.21	23	38	0.61	55	141	0.39
T	Teop	112	597	0.19	23	64	0.36	32	90	0.36
L	Tulil	73	375	0.19	18	60	0.30	60	185	0.32
V	Vera'a	282	1849	0.15	73	187	0.39	137	394	0.35
	totals	1070	6429	—	288	851	—	509	1687	—

Figure 6.21 | Lexicality of anaphoric subjects by role of antecedent.
The 'other' category encompasses oblique arguments and all non-core positions such as adjuncts, as well as possessives and other types of modifiers.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	ante. role	(*subject*, object, other)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.222
speaker	37	0.369

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.20	−1.626	0.105	−15.53	<0.001
(A_1)	ante. role	= object	2.93	1.075	0.083	13.01	<0.001
(A_2)		= other	2.32	0.842	0.065	12.97	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−1.044	−0.519	−0.428	−0.337	3.990

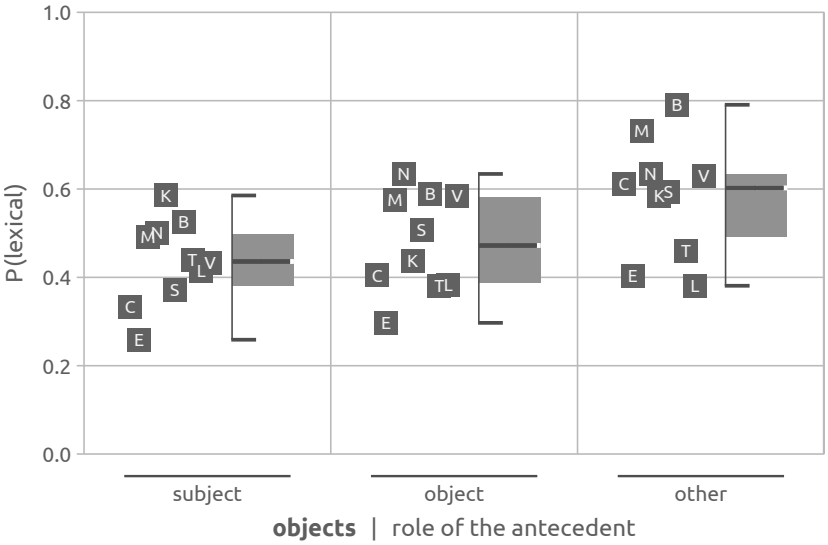
d. | correlation of fixed effects

	(intercept)	(A_1)
(A_1)	−0.140	
(A_2)	−0.181	0.234

e. | model evaluation

observations	8967	AIC	8781
model deviance	8771	log-likelihood	−4385
residual d.f.	8962	conditional R^2	0.099
		marginal R^2	0.048

Table 6.13 | Regression model results for the lexicality of anaphoric subjects by the role of the antecedent, with corpus and speaker as random effects.



		subject			object			other		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	15	45	0.33	57	141	0.40	30	49	0.61
E	English	37	143	0.26	117	394	0.30	75	186	0.40
M	Mandarin	28	57	0.49	54	94	0.57	38	52	0.73
N	Nafsan	32	64	0.50	78	123	0.63	26	41	0.63
K	N. Kurdish	24	41	0.59	66	151	0.44	38	65	0.58
S	S. Dargwa	10	27	0.37	36	71	0.51	16	27	0.59
B	Tabasaran	20	38	0.53	96	163	0.59	34	43	0.79
T	Teop	29	66	0.44	53	139	0.38	23	50	0.46
L	Tulil	17	41	0.41	51	133	0.38	24	63	0.38
V	Vera'a	64	148	0.43	187	320	0.58	92	146	0.63
	totals	276	670	—	795	1729	—	396	722	—

Figure 6.22 | Lexicality of anaphoric objects by the role of the antecedent.
The 'other' category encompasses oblique arguments and all non-core positions such as adjuncts, as well as possessives and other types of modifiers.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	ante. role	(*subject*, object, other)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.293
speaker	37	0.586

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.79	−0.234	0.165	−1.42	0.157
(A_1)	ante. role	= object	1.22	0.195	0.097	2.01	0.044
(A_2)		= other	1.90	0.643	0.114	5.64	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−2.106	−0.855	−0.593	0.950	1.787

d. | correlation of fixed effects

	(intercept)	(A_1)
(A_1)	−0.420	
(A_2)	−0.353	0.614

e. | model evaluation

observations	3121	AIC	4122
model deviance	4112	log-likelihood	−2056
residual d.f.	3116	conditional R^2	0.127
		marginal R^2	0.013

Table 6.14 | Regression model results for the lexicality of anaphoric objects by the role of the antecedent, with corpus and speaker as random effects.

($\sigma = 0.586$), which, as we have seen and will see again, is not uncommon for objects. Compared to a subject antecedent, the odds of a lexically realized object are $e^{\beta} = 1.22$ ($p = 0.044$) times higher for object antecedents, and $e^{\beta} = 1.90$ ($p < 0.001$) times higher for antecedents in other roles.

Similar to promotions to subject, role changes to object position incur a penalty in identifiability that favours a higher rate of lexical expression, though this effect appears more subdued due to the already higher baseline rate of lexicality among objects. The associated effect of same-role chains reducing lexicality rates seen above for subjects is notably absent here – as a whole, objects benefit less from highly cohesive contexts than subjects. We will examine the effect of role persistence in anaphoric chains in more detail in the next section.

6.7.4 | Interaction with anaphoric distance

For the interaction between antecedent role and antecedent distance, we are chiefly interested in same-role clause chains (cf. Givón 2017; Travis & Torres Cacoullos 2018), that is, cases where the antecedent is located in the previous clause (i.e. $d = 1$ clause) and the roles of the antecedent and the anaphor in question are the same. This is a fairly restrictive definition of anaphoric chains; a more inclusive alternative might consider mentions in any position, so long as they occur in adjacent clauses. In other words, we are looking at for instances of a referent being mentioned in subject or object position, then being mentioned in same position again in next clause, as in (72) above and (74) below:

(74) English

a. They used to get rope, …

	they	*used*	*to*	*get*	*rope*
	3PL	used	to	get.INF	rope
##	pro.h:a	lv_aux	lv	v:pred	np:p
	0268				0271

'They used to get rope, …'

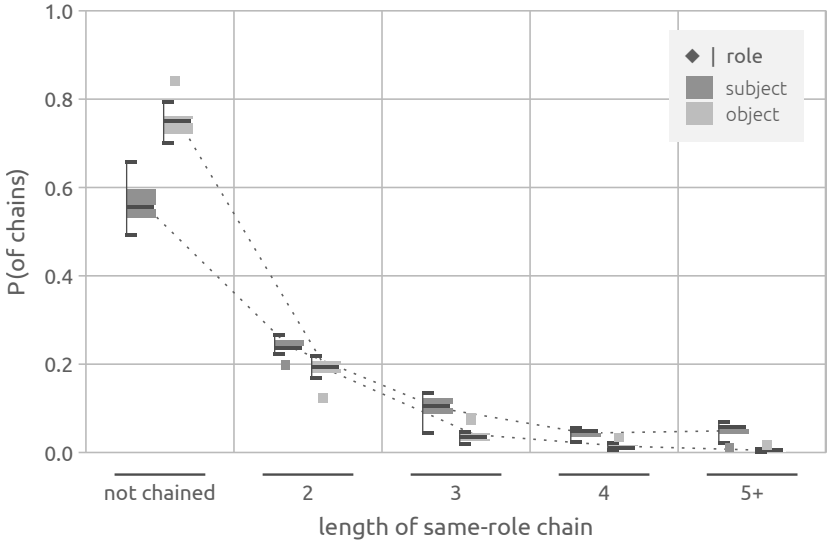


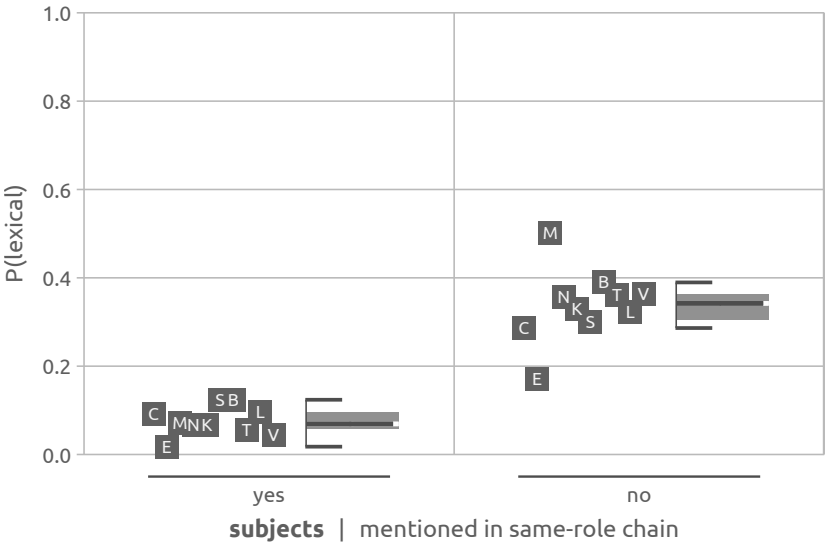
Figure 6.23 | Lengths of same-subject and same-object chains across the ten corpora.
An anaphor is considered to be in a same-role context if its immediate antecedent was mentioned in the same position in the previous clause, i.e. at a distance of $d = 1$ clause unit. Also shown here is the proportion of anaphors that do not start clause chains, i.e. with a length of $l = 1$ link.

b. *and pick it all to pieces.*

	and	*pick*	*it*	*all*	*to*	*piece-s*
	and	pick.INF	3SG.N.OBL	all	to	piece-PL
##	other 0.h:a 0268	v:pred	pro:p 0271	rn	rv	rv

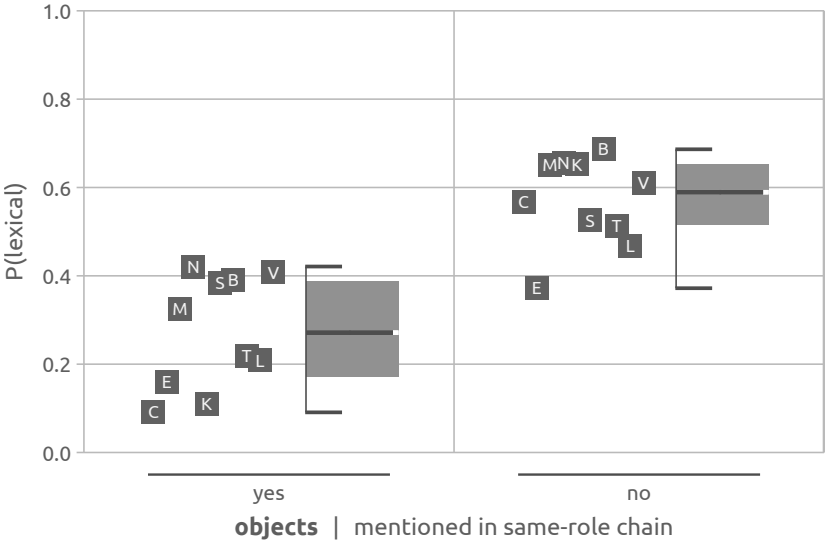
'and pick it all to pieces.' [mc_english_devon01_0125]

This definition places no length limitations on what does and does not constitute a clause chain, so that chains can be as short as two mentions, minimally consisting of the anaphor in question and its antecedent. As Figure 6.23 shows,



		yes			no				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	19	207	0.09	67	234	0.29	0.25	<0.001
E	English	8	452	0.02	155	903	0.17	0.22	<0.001
M	Mandarin	25	354	0.07	183	365	0.50	0.47	<0.001
N	Nafsan	25	369	0.07	117	329	0.36	0.36	<0.001
K	N. Kurdish	19	281	0.07	119	362	0.33	0.32	<0.001
S	S. Dargwa	29	234	0.12	87	290	0.30	0.21	<0.001
B	Tabasaran	47	383	0.12	157	403	0.39	0.30	<0.001
T	Teop	19	341	0.06	148	410	0.36	0.37	<0.001
L	Tulil	21	216	0.10	130	404	0.32	0.25	<0.001
V	Vera'a	53	1225	0.04	439	1205	0.36	0.40	<0.001
	totals	265	4062	—	1602	4905	—	—	—

Figure 6.24 | Lexicality of anaphoric subjects in a same-role chain.
A subject mention is considered to be in a same-role context if its immediate antecedent was the subject of the previous clause, i.e. at a distance of $d = 1$ clause unit.



		yes			no				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	6	66	0.09	96	169	0.57	0.43	<0.001
E	English	30	188	0.16	199	535	0.37	0.20	<0.001
M	Mandarin	12	37	0.32	108	166	0.65	0.26	<0.001
N	Nafsan	24	57	0.42	112	171	0.65	0.21	0.002
K	N. Kurdish	8	73	0.11	120	184	0.65	0.49	<0.001
S	S. Dargwa	10	26	0.38	52	99	0.53	0.11	0.202
B	Tabasaran	23	59	0.39	127	185	0.69	0.26	<0.001
T	Teop	19	87	0.22	86	168	0.51	0.28	<0.001
L	Tulil	15	72	0.21	77	165	0.47	0.24	<0.001
V	Vera'a	64	157	0.41	279	457	0.61	0.18	<0.001
	totals	211	822	—	1256	2299	—	—	—

Figure 6.25 | Lexicality of anaphoric objects in a same-role chain.
An object mention is considered to be in a same-role context if its immediate antecedent was the object of the previous clause, i.e. at a distance of $d = 1$ clause unit.

most same-role chains are in fact fairly short (cross-corpus mean length of $l = 2.92$ clauses for subjects, $\sigma = 0.238$, and $l = 2.33$, $\sigma = 0.141$ for objects), with proportions that are remarkably stable across corpora, in particular for objects. Clause chains are substantially more likely to occur and tend to be slightly longer in subject position than in object position, which leads to a substantial proportion of subject anaphors to be situated in clause-chaining contexts ($P = 0.45$ with $\sigma = 0.064$ for subjects, compared to $P = 0.26$ with $\sigma = 0.046$ for objects). As per Givón (Givón, 2017), this pattern is a major driver of discourse coherence and hence of referential choices. The strength of the effects discussed in the following presumably increases proportionally with the length of the clause chain; since the effect is already essentially unambigious when charted across chains of any length, as we shall see in the following, however, this is not looked into more closely here, and will be left as a possible avenue for future research.

We now move on to examining the effect of clause chaining on lexicality. Note that in the following, the first element in a chain (i.e. the one that starts it) is not considered part of it, only the subsequent links are. We will return to the role that chain starters play in structuring discourse later in Section 9.2. As Figure 6.24 shows, lexical subjects in clause chains are exceedingly uncommon in all ten corpora in the sample, having close to half the overall lexicality rate for subjects (cf. Figure 5.2 in Section 5.2.1). At the same time, this is where zero subjects in particular are an extremely frequent occurrence (cross-corpus mean $P = 0.71$, $\sigma = 0.212$ in chains vs. $P = 0.43$, $\sigma = 0.203$ outside), including in English ($P = 0.28$ vs. $P = 0.06$) and Vera'a ($P = 0.51$ vs. $P = 0.18$), where it is in fact the context in which the majority of zero subjects occur (cf. Huddleston & Pullum 2002; Schnell & Barth 2020).

The corresponding picture for objects is shown in Figure 6.25. Here as with subjects, mentions in same-role chains are substantially less likely to be realized lexically than those in other contexts, though as is usual with objects, variation between corpora is greater. In Cypriot Greek and Northern Kurdish, for instance, the association between chained and non-chained mentions is particularly strong ($\varphi = 0.43$ with $p < 0.001$ in Greek; $\varphi = 0.49$ with $p < 0.001$ in Kurdish), while in Sanzhi Dargwa ($\varphi = 0.11$ with $p < 0.202$) and to lesser degree English ($\varphi = 0.20$ with $p < 0.001$), it is much weaker. As noted above, however, a much smaller fraction of object mentions occur in same-role chains compared to subject mentions, though this fraction is still remarkably stable across corpora.

Lastly, note that since same-role chains are in practice the interactions of two other factors included in our model – anaphoric distance and antecedent role – they will not be independently represented in the multifactorial analysis in Chapter 7. However, as will become apparent later, the interaction that underpins them will come out as highly significant there as well, with substantial ramifications for our account of the mechanisms of referential choice and lexicality rates.

6.8 | Form of the antecedent

6.8.1 | Definition and methodological issues

Another factor claimed to have influence on referential choice is the form of the immediate antecedent (e.g. Ariel 2001, see Section 2.4.2.3), in that preceding reducing mentions yield subsequently reduced mentions, and vice versa. This section examines the rate at which lexical expressions are selected for anaphors in subject and object position depending on whether the immediate antecedent is

- lexical or
- non-lexical.

As before, pronominal and zero antecedents are combined into a single category here, as these form types are not equally common across the ten corpora (see Figures 5.2 and 5.3 in Section 5.2). This imbalance could otherwise adversely affect the statistical robustness of the results. In particular, zero is infamously rare in English ($P = 0.08$ zero antecedents to subject anaphors), but conversely very frequent in the Nafsan ($P = 0.62$) and Sanzhi Dargwa ($P = 0.57$) corpora, among others (cross-corpus mean $P = 0.41$; $\sigma = 0.162$). For an explanation of how antecedence is defined for the purposes of this study, refer to Section 4.6.2.

Different antecedent forms are likely to have different effects depending on the distance to the anaphor. The interaction between these factors is tested in Sections 6.8.3 and 6.8.4, the one with antecedent role specifically in the context of same-role clause chains.

6.8.2 | Lexicality by antecedent form

6.8.2.1 | Subjects

As Figure 6.26 shows, non-lexical antecedents usually lead into non-lexical subject anaphors in all corpora in the sample (cross-corpus mean $P = 0.16$; $\sigma = 0.037$). Conversely, if the antecedent is lexical, then the rate of lexical expression for the anaphor is higher ($P = 0.37$; $\sigma = 0.100$). In all but two of the corpora, it is substantially higher, being roughly double that for non-lexical antecedents. The exceptions are Cypriot Greek and English, where the form of antecedent by itself has comparatively little influence on the form of the anaphor ($\varphi = 0.06$ with $p < 0.206$ for the former, and $\varphi = 0.11$ with $p < 0.001$ for the latter). These associations are echoed by the regression model summarized in Table 6.15: The odds of a lexical expression are appreciably higher ($e^{\beta} = 3.21$ with $p < 0.001$) if the antecedent is also lexical compared to the base rate for a non-lexical antecedent, relative to the odds of the intercept ($e^{\beta} = 0.18$). The model further indicates a certain level of cross-corpus homogeneity (random effects intercept for corpus $\sigma = 0.219$), meaning that even though the association is not equally strong in all corpora, it is nevertheless recognizable in all of them.

These observations suggest that form persistence plays an appreciable role in the selection of referring expressions for subjects: Once a referent has been established via a pronominal or zero mention, it is far more likely than not to remain non-lexical on subsequent mentions. Notably, lexical antecedents do not appear to transition into non-lexical anaphors at a rate that would equalize the effect of antecedent form on anaphor form; a confounding factor is the interaction between antecedent form and antecedent distance, which we will briefly examine further below.

6.8.2.2 | Objects

Figure 6.27 shows that in terms of their general association, the forms of antecedent and anaphor pattern in much the same way for objects as they do for subjects. Lexical antecedenta are more likely to result in a lexical object anaphor with $e^{\beta} = 3.01$ times higher odds ($p < 0.001$) compared to non-lexical antecedents, as the regression model in Table 6.16 suggests. The difference between median lexicality rates for non-lexical and lexical antecedents is almost the same as that for subjects ($M_{diff} = 0.29$ for objects vs. $M_{diff} = 0.24$

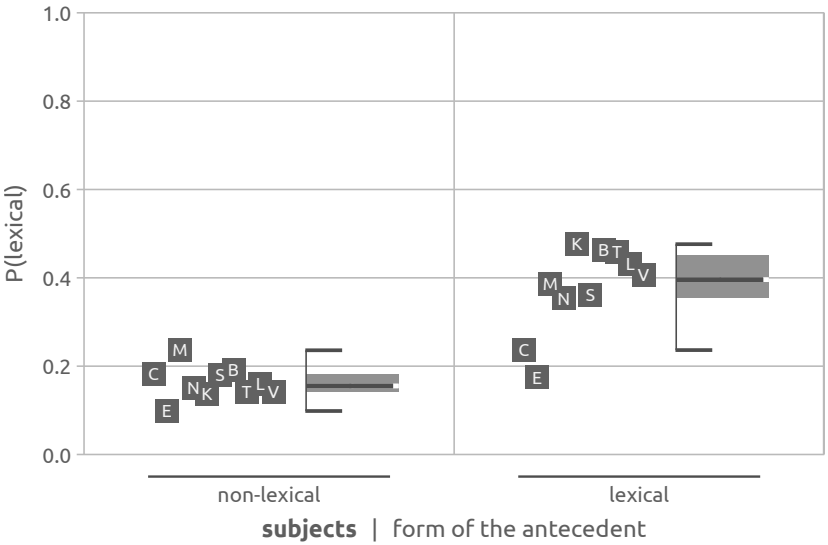
for subjects) but shifted upwards, as objects are more likely to be lexical in general. However, there is a noticeably greater spread of values among the ten corpora, especially for non-lexical antecedents (random effect intercept $\sigma = 0.241$).

Notably, almost half of all objects in the sample have lexical antecedents (cross-corpus mean $P = 0.52$, $\sigma = 0.092$); of these, more than half ($P = 0.62$, $\sigma = 0.069$) are themselves realized lexically, indicating a substantial level of form persistence that is well in line with the general tendency for objects to be realized as full NPs.

6.8.3 | Interaction with anaphoric distance

Figure 6.28 visualizes the interaction of antecedent form with antecedent distance for subject anaphors. For all corpora, non-lexical antecedents at short distances ($d \leq 2$ clauses) lead to non-lexical subject anaphors at a rate lower than the lexical baseline rate. This is especially true for anaphors with non-lexical antecedents in the preceding clause (i.e. $d = 1$ clause), which are almost invariably non-lexical themselves (cross-corpus mean $P = 0.08$, $\sigma = 0.038$). From a broad cross-corpus perspective, the difference between non-lexical and lexical antecedents is greatest at intermediate distances, but appears to narrow slightly at either extreme of the distance scale, especially at its lower end. The extension of the influence of the formal properties of the antecedent to long anaphoric distances is surprising, as we would expect form persistence and the circumstances under which the antecedent was mentioned to become less important as distance increases. The smallest difference between lexical and non-lexical antecedents being at a distance of $d = 1$ clause follows from the effect of tight anaphoric chains on referential choice, which we already remarked on earlier in Section 6.7.4. This effect complements the fact that certain a classes of referents that are inherently more likely to be non-lexically expressed, irrespective of distance, specifically human (Section 6.1) and highly frequent referents (Section 6.2).

The corresponding interaction for object anaphors is illustrated in Figure 6.29, which largely shows the same picture, albeit shifted upwards and with a somewhat greater disparity between antecedent forms at distances of $d = 1$ clause units. Note further that for both object and subjects anaphors, there is a degree of variability between corpora, particularly in the intermediate distance categories where subsample sizes are relatively small.



	corpus	non-lexical N(lex)	non-lexical N(all)	non-lexical P(lex)	lexical N(lex)	lexical N(all)	lexical P(lex)	φ	*p*-val.
C	C. Greek	60	331	0.18	26	110	0.24	0.06	0.206
E	English	95	965	0.10	68	390	0.17	0.11	<0.001
M	Mandarin	109	462	0.24	99	257	0.39	0.16	<0.001
N	Nafsan	77	514	0.15	65	184	0.35	0.22	<0.001
K	N. Kurdish	68	496	0.14	70	147	0.48	0.35	<0.001
S	S. Dargwa	73	405	0.18	43	119	0.36	0.18	<0.001
B	Tabasaran	112	587	0.19	92	199	0.46	0.27	<0.001
T	Teop	80	561	0.14	87	190	0.46	0.33	<0.001
L	Tulil	69	430	0.16	82	190	0.43	0.29	<0.001
V	Vera'a	267	1876	0.14	225	554	0.41	0.28	<0.001
	totals	1010	6627	—	857	2340	—	—	—

Figure 6.26 | Lexicality of anaphoric subjects by the form of the antecedent.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	ante. form	(*non-lexical*, lexical)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.219
speaker	37	0.295

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.18	−1.704	0.096	−17.69	<0.001
(A_1)	ante. form	= lexical	3.21	1.167	0.056	20.86	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−0.989	−0.480	−0.414	−0.309	3.803

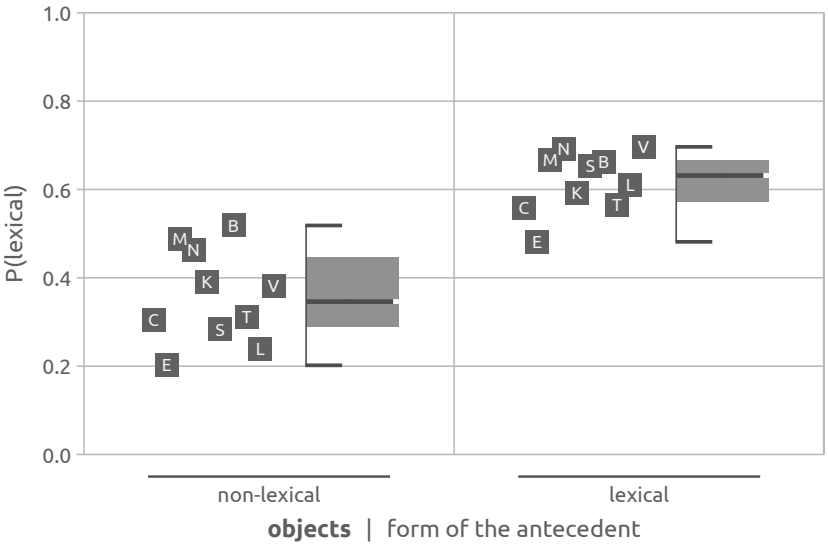
d. | correlation of fixed effects

	(intercept)
(A_1)	−0.226

e. | model evaluation

observations	8967	AIC	8621
model deviance	8613	log-likelihood	−4306
residual d.f.	8963	conditional R^2	0.108
		marginal R^2	0.071

Table 6.15 | Regression model results for the lexicality of anaphoric subjects by the form of the antecedent, with corpus and speaker as random effects.



		non-lexical			lexical				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	35	115	0.30	67	120	0.56	0.26	<0.001
E	English	86	426	0.20	143	297	0.48	0.30	<0.001
M	Mandarin	42	86	0.49	78	117	0.67	0.18	0.011
N	Nafsan	44	95	0.46	92	133	0.69	0.23	0.001
K	N. Kurdish	47	120	0.39	81	137	0.59	0.20	0.001
S	S. Dargwa	15	53	0.28	47	72	0.65	0.37	<0.001
B	Tabasaran	42	81	0.52	108	163	0.66	0.14	0.029
T	Teop	48	154	0.31	57	101	0.56	0.25	<0.001
L	Tulil	34	142	0.24	58	95	0.61	0.37	<0.001
V	Vera'a	102	268	0.38	241	346	0.70	0.32	<0.001
	totals	495	1540	—	972	1581	—	—	—

Figure 6.27 | Lexicality of anaphoric objects by the form of the antecedent.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	ante. form	(*non-lexical*, lexical)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.241
speaker	37	0.464

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.57	−0.567	0.130	−4.35	<0.001
(A_1)	ante. form	= lexical	3.01	1.103	0.078	14.14	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−2.113	−0.765	−0.520	0.837	1.923

d. | correlation of fixed effects

	(intercept)
(A_1)	−0.318

e. | model evaluation

observations	3121	AIC	3950
model deviance	3942	log-likelihood	−1971
residual d.f.	3117	conditional R^2	0.149
		marginal R^2	0.079

Table 6.16 | Regression model results for the lexicality of anaphoric objects by the form of the antecedent, with corpus and speaker as random effects.

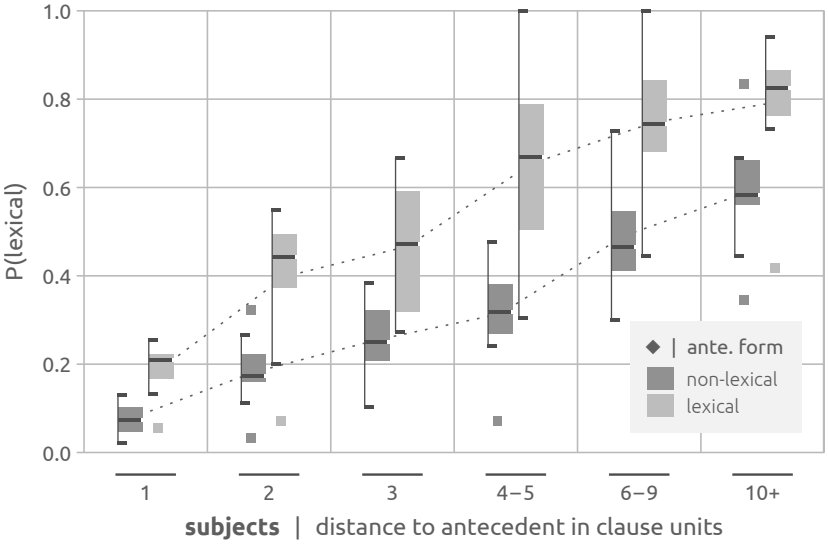


Figure 6.28 | Interaction of the effect of anaphoric distance and form of the antecedent on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

6.8.4 | Form of antecedents in same-role chains

Lastly, we will briefly examine a special case of the interaction discussed in the previous section, the four-way interplay between lexicality of anaphors, and the form and role of and distance to the antecedent. Specifically, Figures 6.30 and 6.31 show, for subjects and objects respectively, the interaction between antecedent form and whether or not the anaphor in question is part of a same-role chain (i.e. having the same role as its antecedent in the preceding clause).

For subject anaphors, if a same-role antecedent is non-lexical, then a lexical realization is extremely unlikely in all corpora in the sample. It is only slightly more likely if the antecedent is lexical, with the exception English (i.e. the low outlier), where antecedent form makes no difference on the form

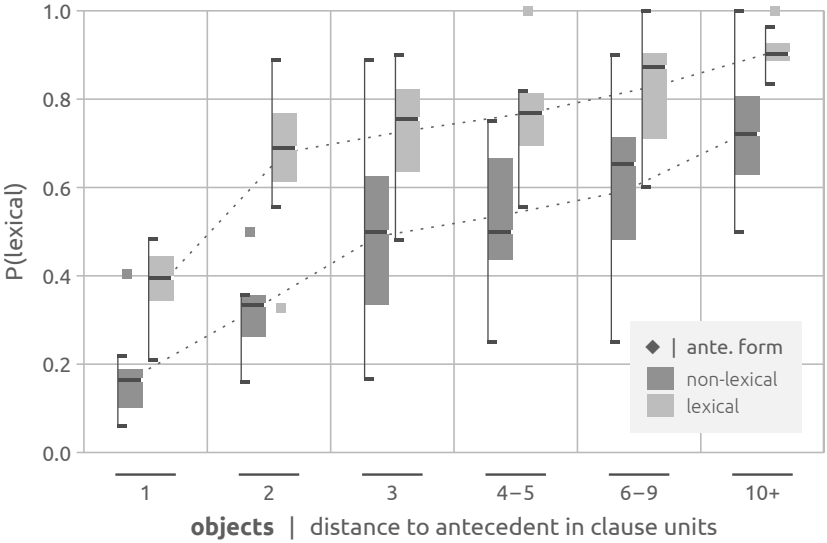


Figure 6.29 | Interaction of the effect of anaphoric distance and form of the antecedent on the lexicality of anaphoric objects, by corpus.
The small squares indicate outliers from the central distribution.

of chained subjects. Also note that the vast majority of antecedents in clause chains are non-lexical (cross-corpus mean $P = 0.81$, $\sigma = 0.036$). For mentions outside of same-role chain contexts, conversely, differences in the form of the antecedent have much greater impact on form of anaphor, in essence reflecting picture in Figure 6.26 above.

Among objects, the interaction is largely additive, in that there is no specific association of antecedent forms with same-role contexts as regards lexicality. In other words, objects with lexical antecedents have higher lexicality rates than those with non-lexical antecedents, but to the same degree in and outside of same-role chains. Nevertheless, the pattern seen in (75) and (76), of lexical object mentions followed by non-lexical ones in the next clause, is not an uncommon one.

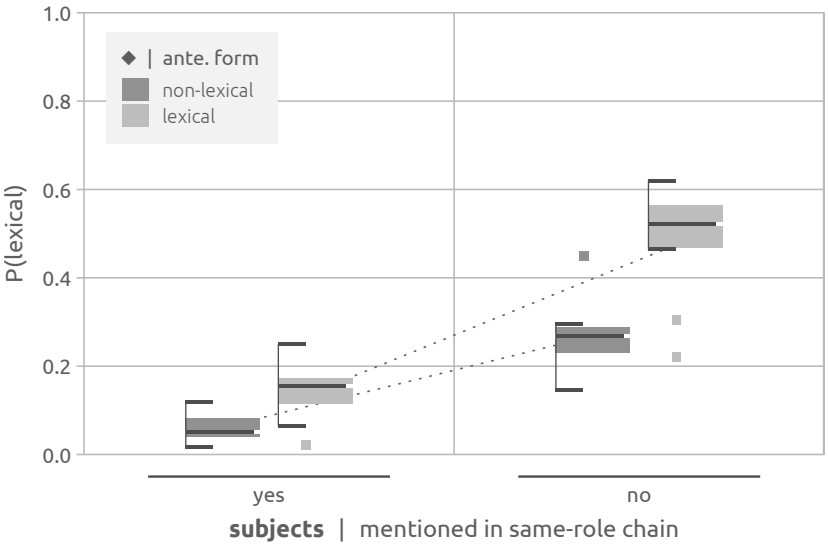


Figure 6.30 | Interaction of mention in same-role chains and form of the antecedent on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

(75) Tabasaran

a. *aˁru gülaˁli,*

	aˁ-ru	*gülaˁli*
	go-FUT	Gulali
##	v:pred	pn_np.h:s
		0005

'Gulali went,'

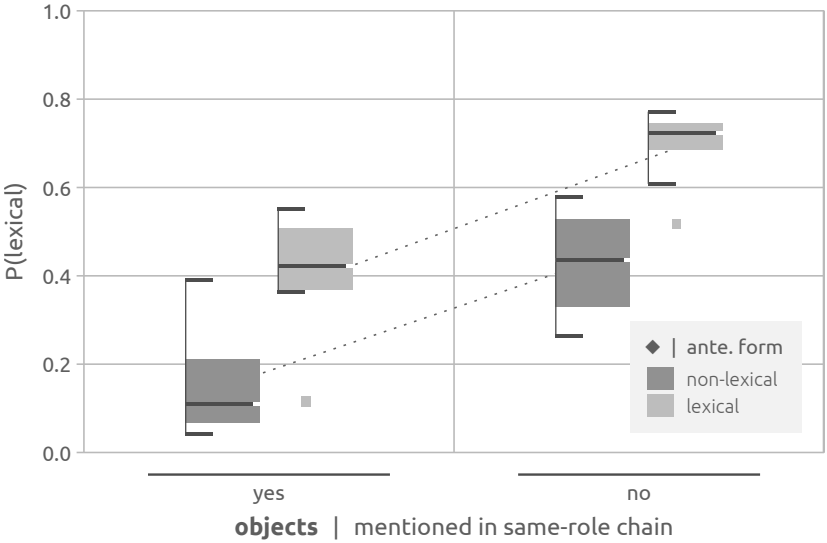


Figure 6.31 | Interaction of mention in same-role chains and form of the antecedent on the lexicality of anaphoric objects, by corpus.
The small squares indicate outliers from the central distribution.

b. *χuru murtir,*

	χ-uru	*murt-ir*
	bring-FUT	egg-PL(ABS)
##	0.h:a	np:p
	0005	0040
	v:pred	

'(he) brought eggs,'

c. *durxnu murtir,*

	d-u<r>x-nu	*murt-ir*
	PFV-<PL>boil-PST	egg-PL(ABS)
##	0.h:a v:pred	np:p
	0005	0040

‘(he) boiled the eggs,’

d. *ǯivru,*

	ǯiv-ru
	put-FUT
##	0.h:a v:pred 0:p
	0005 0040

‘(he) served (them),’

e. *ip’uru murari.*

	ip’-uru	*mu-rari*
	<NSG>eat-FUT	PROX-PL(ERG)
##	v:pred	dem_pro.h:a 0:p
		0036 0040

‘and they ate (them).’ [mc_tabasaran_naz_0053]

(76) a. *And he went up between’em […] and put the plug in.*

	and	*he*	*went*	*up*	*between*	*=’em*
	and	3SG.M	go.PST	up	between	=3PL.OBL
##	other	pro.h:s	v:pred	rv	adp	=pro:g
		0128				0127

	and	*put*	*the*	*plug*	*in*
	and	put.PST	DEF	plug	in
##	other	0.h:a v:pred	ln_det	np:p	rv
		0128		0143	

b. *Now, he said, […] when you put that plug in, […] tie it in with that bit of thong.*

	now	*he*	*said*							
	now	3SG.M	say.PST							
##	other	pro.h:s_ds	v:pred							
		0128								

		when	*you*	*put*	*that*	*plug*	*in*	
		when	2SG	put.PRS	DIST.SG	plug	in	
##ds	#ds_ac	adp	pro.2:a	v:pred	ln_dem	np:p	rv	%
			0000			0143		

tie	*it*	*in*	*with*	*that*	*bit*	*of*	*thong*
tie.IMP	3SG.N	in	with	DIST.SG	bit	of	thong
0.2:a	v:pred	pro:p	rv	adp	ln_dem	np:obl	rn	rn_np
0000		0143				0145		

[mc_english_kent03_0140-0141]

6.9 | Sequence of mention

6.9.1 | Definition and methodological issues

In Section 2.4.2.6, we have proposed the notion, based on Kibrik (2000: 78), that newly introduced referents require repeated mentions before they reach a "working level" of accessibility, which should be reflected in the rate of lexical expression among subsequent mentions. In the following, we will chart this gradual establishment of referents by charting the linear order of occurrence of the first few mentions of each referent, starting from its introduction into discourse, and examine the effect that the position of an anaphor in this sequence has on its form. To this end, we distinguish whether the anaphor in question is

a. the second mention of the referent in the text, with its antecedent being the first;

b. the third mention of the referent in the text, with its antecedent being the second; or
c. a subsequent mention.

This sequence may include mentions in any position that meet the selection criteria given in Section 4.1, not just subjects or objects. In other words, what we are charting here is the number of times a referent has been mentioned in any position *prior* to being mentioned in subject or object position, rather than the number of times a referent has been mentioned *in* subject or object position. The following is an example from the Cypriot Greek corpus that includes the first and second mentions of a referent (with the index ⟨0066⟩), the latter in this case being a definite NP:

(77) Cypriot Greek

a. *O jiris mu inda ghul'an ekamnen?*

	o	*jiris*	*mu*	*inda*	*ghul'an*	*ekamnen*
	DEF.M.NOM	father	1SG.POSS	what	work	do.IPFV.3SG
#ds	ln	np.h:a	rn_pro.1:poss	ln	np:p	v:pred
		0066	0002		0067	

'[He asked,] What work did my father do?'

b. *O jiris su itan psaras ye mu.*

	o	*jiris*	*su*	*itan*	*psaras*
	DEF.M.NOM	father	2SG.POSS	be.PST.3SG	fisherman
##ds	ln	np.h:s	rn_pro.2:poss	cop	np.h:pred
		0066	0002		

ye	*mu*
son	1SG.POSS
np.h:voc	rn_pro.1:poss
0002	0001

'Your father was a fisherman, my son.'

[mc_cypgreek_psarin_0011-0012]

As noted earlier in Section 5.3, first mentions of referents are overwhelmingly lexical (cross-corpus mean $P = 0.85$, $\sigma = 0.049$) across all corpora, and mostly

occur in object position ($P = 0.31$, $\sigma = 0.064$) or else in various positions outside of the clausal core ($P = 0.46$, $\sigma = 0.059$).

Because of how mention sequences are coded, there exists a structural association between them and another factor, the frequency of recent co-referential mentions (Section 6.4): If the antecedent of a given anaphor is the first mention of that referent in the text, then the frequency of co-referential mentions is necessarily $n \leq 1$, and analogously $n \leq 2$ if it is the second mention. As later mentions tend to vastly outnumber earlier ones, however, this association is unlikely to affect results in a noticeable way.

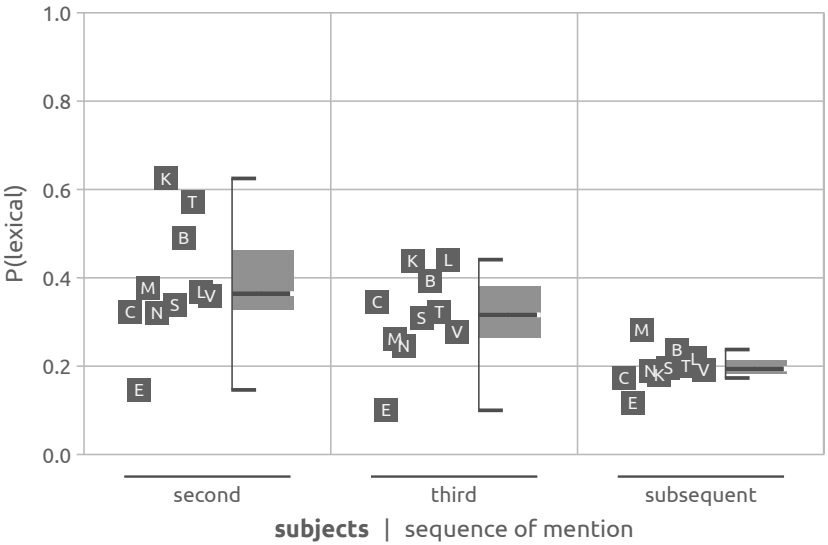
Note further that only 28% ($\sigma = 8\%$) of all referents in the corpus data are mentioned four or more times; 19% ($\sigma = 2\%$) are mentioned twice, 9% ($\sigma = 2\%$) three times. This means a substantial proportion of referents (44%, $\sigma = 8\%$) are only ever mentioned once in a text; as discussed earlier, these cases are excluded from analysis due to our specific focus on anaphoric mentions.

6.9.2 | Lexicality by sequence of mention

A note on the regression model summaries in this section: Unlike the previously examined categorical factors, form and role of the antecedent, the three levels of sequence of mention have an inherent order ('second' < 'third' < 'subsequent'). Because of this, the model coefficients and odds ratios in Tables 6.17 and 6.18 should not be interpreted relative to the base level by default, but to the next lower level, which in turn depends on the level below, and so on.

6.9.2.1 | Subjects

Figure 6.32 shows the association between sequence of mentions and lexicality for subjects. From a broad cross-corpus view, the displayed pattern matches up with the expectation of rates of lexical expression gradually decreasing as referents are mentioned repeatedly following their initial introduction into discourse ($\rho = -0.65$ with $p < 0.001$). This broadly supports similar findings in Lichtenberk (1996) on the basis of narratives in To'aba'ita (Oceanic, Solomon Islands). After $N \geq 3$ mentions, subjects approach the baseline lexicality rate of around 20% (Section 5.2.1). Adding the lexicality



	corpus	second N(lex)	second N(all)	second P(lex)	third N(lex)	third N(all)	third P(lex)	subsequent N(lex)	subsequent N(all)	subsequent P(lex)
C	C. Greek	11	34	0.32	9	26	0.35	66	381	0.17
E	English	37	253	0.15	17	170	0.10	109	932	0.12
M	Mandarin	23	61	0.38	12	46	0.26	173	612	0.28
N	Nafsan	18	56	0.32	12	49	0.24	112	593	0.19
K	N. Kurdish	20	32	0.62	14	32	0.44	104	579	0.18
S	S. Dargwa	20	59	0.34	13	42	0.31	83	423	0.20
B	Tabasaran	25	51	0.49	11	28	0.39	168	707	0.24
T	Teop	20	35	0.57	10	31	0.32	137	685	0.20
L	Tulil	21	57	0.37	15	34	0.44	115	529	0.22
V	Vera'a	41	114	0.36	26	94	0.28	425	2222	0.19
	totals	236	752	—	139	552	—	1492	7663	—

Figure 6.32 | Lexicality of anaphoric subjects by sequence of mention.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	sequence	(*second* < third < subsequent)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.255
speaker	37	0.353

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.37	−1.001	0.114	−8.78	<0.001
(A_1)	sequence	= third	0.56	−0.586	0.063	−9.37	<0.001
(A_2)		= subsequent	0.93	−0.077	0.089	−0.87	0.386

c. | scaled residuals

min.	lower	median	upper	max.
−1.008	−0.538	−0.471	−0.354	3.643

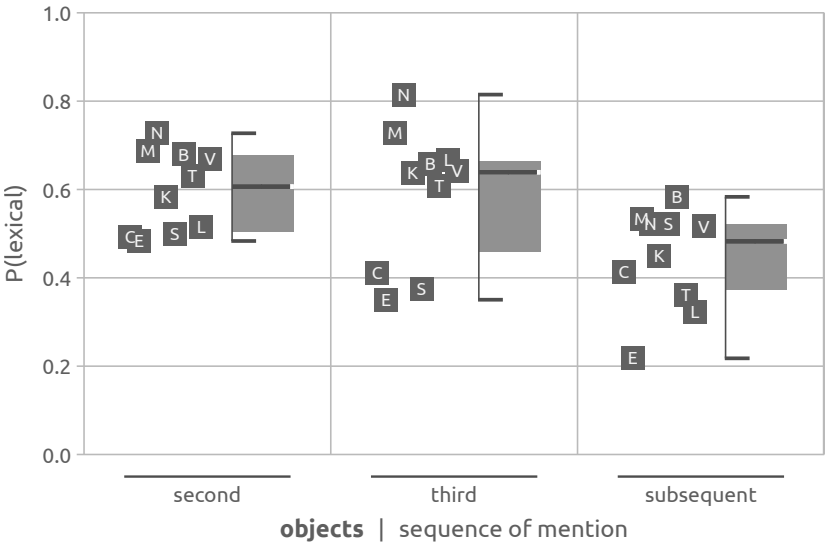
d. | correlation of fixed effects

	(intercept)	(A_1)
(A_1)	−0.193	
(A_2)	−0.170	−0.277

e. | model evaluation

observations	8967	AIC	8953
model deviance	8943	log-likelihood	−4471
residual d.f.	8962	conditional R^2	0.071
		marginal R^2	0.018

Table 6.17 | Regression model results for the lexicality of anaphoric subjects by sequence of mention, with corpus and speaker as random effects.



		second			third			subsequent		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	32	65	0.49	16	39	0.41	54	131	0.41
E	English	102	211	0.48	41	117	0.35	86	395	0.22
M	Mandarin	33	48	0.69	16	22	0.73	71	133	0.53
N	Nafsan	32	44	0.73	22	27	0.81	82	157	0.52
K	N. Kurdish	28	48	0.58	21	33	0.64	79	176	0.45
S	S. Dargwa	19	38	0.50	6	16	0.38	37	71	0.52
B	Tabasaran	36	53	0.68	23	35	0.66	91	156	0.58
T	Teop	17	27	0.63	14	23	0.61	74	205	0.36
L	Tulil	17	33	0.52	18	27	0.67	57	177	0.32
V	Vera'a	77	115	0.67	43	67	0.64	223	432	0.52
	totals	393	682	—	220	406	—	854	2033	—

Figure 6.33 | Lexicality of anaphoric objects by sequence of mention.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	sequence	(*second* < third < subsequent)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.315
speaker	37	0.564

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	1.26	0.228	0.154	1.48	0.140
(A_1)	sequence	= third	0.59	−0.532	0.068	−7.83	<0.001
(A_2)		= subsequent	0.84	−0.180	0.093	−1.93	0.054

c. | scaled residuals

min.	lower	median	upper	max.
−2.190	−0.878	−0.558	0.915	1.792

d. | correlation of fixed effects

	(intercept)	(A_1)
(A_1)	−0.112	
(A_2)	−0.128	−0.184

e. | model evaluation

observations	3121	AIC	4083
model deviance	4073	log-likelihood	−2036
residual d.f.	3116	conditional R^2	0.139
		marginal R^2	0.029

Table 6.18 | Regression model results for the lexicality of anaphoric objects by sequence of mention, with corpus and speaker as random effects.

	corpus	V	p-value
C	C. Greek	0.14	0.014
E	English	0.04	0.306
M	Mandarin	0.06	0.273
N	Nafsan	0.09	0.047
K	N. Kurdish	0.27	<0.001
S	S. Dargwa	0.12	0.017
B	Tabasaran	0.15	<0.001
T	Teop	0.19	<0.001
L	Tulil	0.15	0.001
V	Vera'a	0.10	<0.001

Table 6.19 | Cramér's V for lexicality of anaphoric subjects and sequence of mention.

Cramér's V is a measure of association between two categorical variables. It can be interpreted analogously to a correlation coefficient, but scales from $V = 0$ (no association) to $V = 1$ (perfect association).

	corpus	V	p-value
C	C. Greek	0.07	0.537
E	English	0.25	<0.001
M	Mandarin	0.16	0.069
N	Nafsan	0.23	0.002
K	N. Kurdish	0.15	0.060
S	S. Dargwa	0.09	0.572
B	Tabasaran	0.09	0.397
T	Teop	0.21	0.004
L	Tulil	0.25	0.001
V	Vera'a	0.13	0.005

Table 6.20 | Cramér's V for lexicality of anaphoric objects and sequence of mention.

Cramér's V is a measure of association between two categorical variables. It can be interpreted analogously to a correlation coefficient, but scales from $V = 0$ (no association) to $V = 1$ (perfect association).

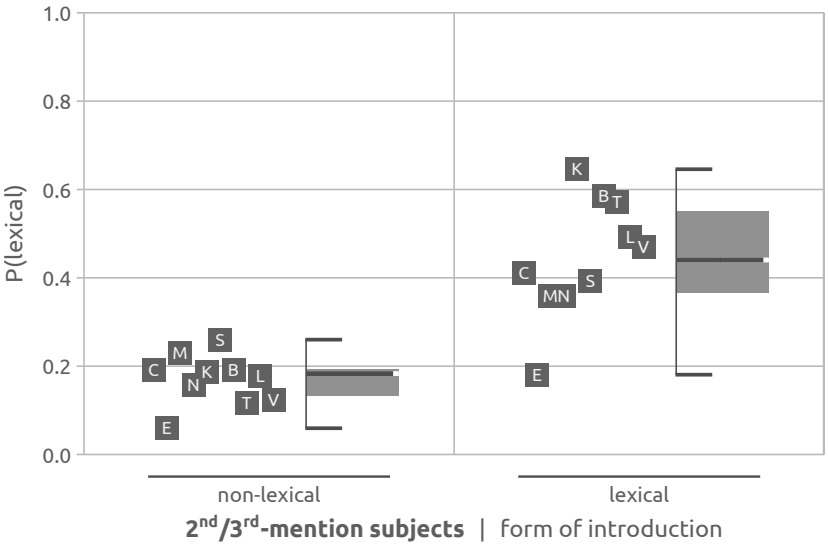
rates for first mentions from Figure 5.5 in Section 5.3 to this cline yields an almost logarithmic decrease in lexicality from introduction to eventual full establishment. Table 6.17 supports these impressions: On average, the odds of a third mention subject being lexical are $e^{\beta} = 0.56$ ($p < 0.001$) times lower than for a second mention; for subsequent mentions, the odds are only $e^{\beta} = 0.93$ times lower ($p < 0.386$) than for third mentions, but $e^{\beta} = 0.56 \times 0.93 = 0.52$ times lower than for second mentions.

As regards variation across corpora, all but one corpus show a monotonous decrease in lexicality rate from second and third mentions to subsequent mentions, but second and third mentions are not equally well differentiated in all corpora. In some corpora, second mentions are somewhat more likely to be realized lexically than third mentions, as in Teop ($r_{pb} = -0.25$, $p < 0.043$), Northern Kurdish ($r_{pb} = -0.19$, $p < 0.137$), and Mandarin ($r_{pb} = -0.12$, $p < 0.208$), while in others the difference between second and third mentions is minimal, as in Cypriot Greek ($r_{pb} = 0.02$, $p < 0.857$). The most egregious exception to the general trend is once again the English corpus, which exhibits practically no association between sequence of mention and the form of subjects (Cramér's $V = 0.04$, $p < 0.306$). Apart from these, lexicality rates are fairly homogenous in each group, with the regression model indicating fairly low variation among intercepts of the random effects of corpus ($\sigma = 0.255$) and speaker ($\sigma = 0.353$).

6.9.2.2 | Objects

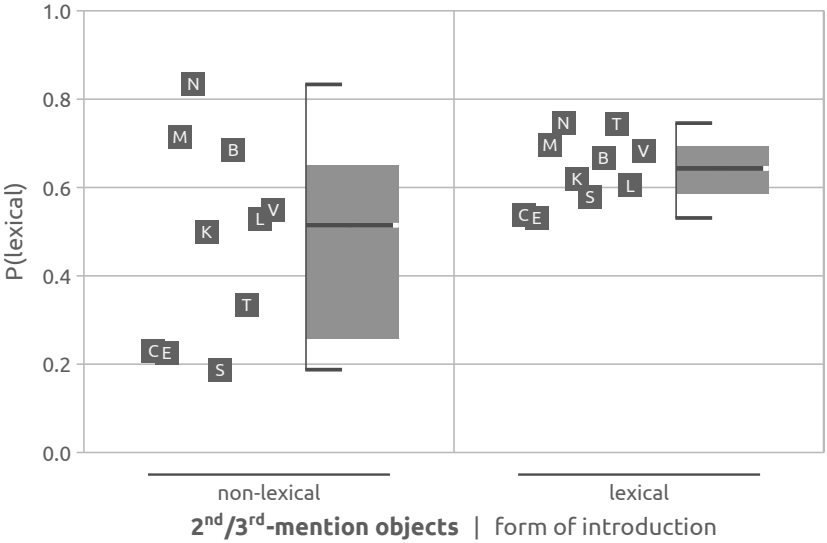
Figure 6.33 shows the corresponding picture for mentions in object position. While second and third mentions are more likely to be realized lexically than later mentions (cross-corpus mean $\varphi = 0.12$, $\sigma = 0.070$), the former two do not form a neat downwards slope ($\rho = -0.42$ with $p < 0.012$) like they do for subjects, likely due to high variance in the third mention group. Third mentions make up 13% of all object anaphors ($\sigma = 2\%$), but split across the ten corpora in the sample, each subsample unfortunately ends up fairly small, resulting in the unclear picture in Figure 6.33. The above-mentioned general association between sequence of mention and lexicality is nevertheless clearly identifiable, even if it is less pronounced than for subjects due to the higher baseline lexicality rate for objects.

Coincidentally, the English data traces an almost perfect linear decrease in lexicality from second to third to subsequent mentions. Counter to the



		non-lexical			lexical				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	4	21	0.19	16	39	0.41	0.22	0.085
E	English	11	185	0.06	43	238	0.18	0.18	<0.001
M	Mandarin	6	26	0.23	29	81	0.36	0.12	0.229
N	Nafsan	6	38	0.16	24	67	0.36	0.21	0.029
K	N. Kurdish	3	16	0.19	31	48	0.65	0.40	0.001
S	S. Dargwa	13	50	0.26	20	51	0.39	0.14	0.157
B	Tabasaran	5	26	0.19	31	53	0.58	0.37	0.001
T	Teop	2	17	0.12	28	49	0.57	0.40	0.001
L	Tulil	5	28	0.18	31	63	0.49	0.30	0.005
V	Vera'a	11	89	0.12	56	119	0.47	0.37	<0.001
	totals	66	496	—	309	808	—	—	—

Figure 6.34 | Lexicality of the second and third mentions of referents in subject position by the form of the introduction.



		non-lexical			lexical				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	6	26	0.23	42	78	0.54	0.27	0.006
E	English	23	102	0.23	120	226	0.53	0.29	<0.001
M	Mandarin	10	14	0.71	39	56	0.70	0.02	0.896
N	Nafsan	10	12	0.83	44	59	0.75	0.08	0.517
K	N. Kurdish	5	10	0.50	44	71	0.62	0.08	0.468
S	S. Dargwa	3	16	0.19	22	38	0.58	0.36	0.008
B	Tabasaran	13	19	0.68	46	69	0.67	0.02	0.885
T	Teop	5	15	0.33	26	35	0.74	0.39	0.006
L	Tulil	9	17	0.53	26	43	0.60	0.07	0.594
V	Vera'a	17	31	0.55	103	151	0.68	0.11	0.152
	totals	101	262	—	512	826	—	—	—

Figure 6.35 | Lexicality of the second and third mentions of referents in object position by the form of the introduction.

near absence of an effect of sequence of mention for subject anaphors, English objects hence show a reasonably strong association (Cramér's $V = 0.25$, $p < 0.001$); this suggests that the characteristics of English discourse structure that render subject anaphors virtually indifferent to most of the tested factors do not equally apply to anaphors in object position. This idea will be developed further in Chapter 9.

6.9.3 | Form of new referent mentions

As we have seen in the previous section, lexicality rates of anaphors in subject and object position gradually approach the baseline level following the introduction of the referent. And as mentioned above and seen earlier in Figure 5.5 in Section 5.3, said introductions are mostly, but not not exclusively, lexical (cross-corpus mean $P = 0.85$, $\sigma = 0.324$), resulting in a monotonic downwards trend of lexicality. But what effect does the form of an introduction have on the form of the next few mentions, specifically if the introduction was already non-lexical?

Figure 6.34 illustrates the association between the form of a referent's introduction and the form of the following two mentions in subject position, and Figure 6.35 for second and third mention in object position. Fourth and later mentions are not shown here, as they are too close to be baseline level to show much of a difference. As might be expected, if the introduction was lexical, then next few subject mentions are realized lexically at a rate much higher than the overall baseline level in all corpora (cross-corpus mean $P = 0.45$, $\sigma = 0.137$), with the sole exception being English. If, however, the introduction was non-lexical, then second and third mentions are lexical at a rate close to that of subsequent mentions in Figure 6.32 ($P = 0.17$, $\sigma = 0.058$), meaning they require fewer instantiations to become "established" in discourse.

For objects, since there are very few observations in the non-lexical group for most corpora, it is best to limit our observations to those with the largest group sizes, this being English, Cypriot Greek, and Vera'a, and examine even those with a modicum of reservation. For two of these three corpora, English and Cypriot Greek, objects pattern similarly to subjects in that non-lexical introductions yield appreciably lower lexicality rates for second and third mentions ($\varphi = 0.29$ with $p < 0.001$ in English; $\varphi = 0.27$ with $p < 0.006$ in Cypriot Greek), presumably for much the same reason. For the third, Vera'a, the dif-

ference is somewhat smaller ($\varphi = 0.11$ with $p < 0.152$), making the association still noticeable, but less obvious.

6.10 | Clause type

6.10.1 | Definition and methodological issues

Clause type and subordination as factors in referential choice and the role given to them in the literature were discussed in Section 2.4.3.3. In this section, we will examine the effect of different clause types, specifically relating to syntactic dependency, on the choice of lexical expressions. We test the broad distinction between

a. fully independent clauses and
b. syntactically subordinated and other types of clauses.

An independent clause is one that is not syntactically subordinate to another clause, and capable of expressing an utterance by itself. By this definition, about a third of all clauses in our sample of ten corpora are classified as not fully independent (cross-corpus mean $P = 0.31$, $\sigma = 0.167$). Unsurprisingly, a substantial degree of variation exists between languages in this regard, as some languages make either heavy use of subordination (e.g. Sanzhi Dargwa with $P = 0.49$) or comparatively little (e.g. Mandarin with $P = 0.07$, Vera'a with $P = 0.11$, and Teop with $P = 0.17$). Mandarin, for instance, tends to string clauses together with minimal marking of syntactic dependencies, resulting in long chains of clauses that are here classified as independent. While these discrepancies are by and large the result of typological variation in clause linking strategies, the specific implementation of syntactic dependency in the corpus annotations may also differ slightly from language to language, and is hence somewhat contentious; see the discussion in Sections 3.3.1 and 4.6.1.2 above.

The benefits of collapsing all non-independent clauses into a single category are twofold: For one, it aligns with the general design principles of this study, which strives for methodological simplicity, since it means only one of the two categories need to be defined. Secondly, and perhaps more importantly, it allows us to gloss over many of the aforementioned language-specific edge cases of syntax, and so focus on the broader, cross-linguistic perspective.

An example of an analytical issue of this kind can be found in Sanzhi Dargwa and Tabasaran, the two Nakh-Daghestanian languages in the sample,

though similar structures can also be found in other languages such as Japanese. In the Sanzhi and Tabasaran corpora, clauses which are not clearly subordinated to any other clause, but also do not meet all criteria for being fully independent clauses, are quite frequent; these tend to be non-finite clauses, usually converb clauses (Forker 2020: ch. 18). In the case of the latter, they commonly appear in chains, often with no discernible matrix clause, as in (78). In Sanzhi, 41% of subjects in the sample occur in clauses of this kind, 30% in Tabasaran.

(78) Sanzhi Dargwa

a. *Ca zamana erit:i sark'ul caw:* …

	ca	*zamana*		*er*	*=it:i*
	one	time		look	=after
#cv	ln_deti	np:other	0.h:a_cps	other:lvc	=lv_adp

sark'-ul	*ca-w*
inspect.IPFV-ICVB	be-M
v:pred	rv_aux

'Once he looked behind himself: …'

b. *duc' rik'ul,* …

	duc'	*r-ik'-ul*
	run	F-move.IPFV-ICVB
#cv	0.h:s_cps	other:lvc v:pred
	0002	

'Running, …'

c. *c'al purχ rik'ul,* …

	c'a-l	*purχ*	*r-ik'-ul*
	fire-ERG	spit	F-say.IPFV-ICVB
#cv	0.h:s_cps_cv	np:obl	other:lvc v:pred
	0002		

'spitting fire, …'

d. *hit:i ruqunne ruc:i, …*

	hit:i-r-uq-un-ne	*ruc:i*
	SUPER-F-go.PFV-PRET-CVB	sister
#cv	v:pred	np.h:s
		0002

‘[his] sister was coming, …’

e. *saˤq’uˤnne*

		saˤ-q’-uˤn-ne
		hither-go-PRET-CVB
#cv	0.h:s	v:pred
	0002	

‘coming [after him].’ [mc_sanzhi_dragon_0020]

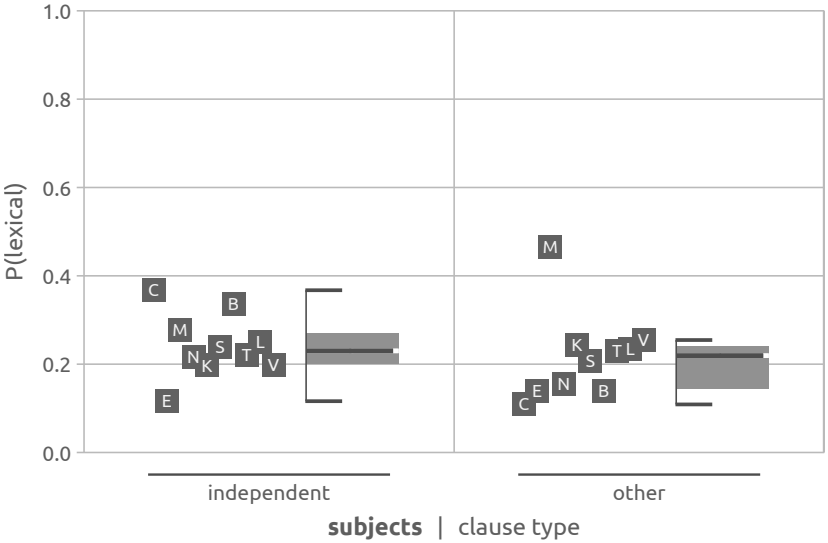
In the following, it is hence advisable to understand the values presented for these corpora in the light of these restrictions, and treat them with due care for the purposes of cross-corpus comparison.

6.10.2 | Lexicality by clause type

6.10.2.1 | Subjects

Figure 6.36 and Table 6.21 summarize the findings for subjects. Given the above-noted high frequency of independent clauses overall, it is no surprise that on average 70% of the subjects in the sample occur in independent clauses ($\sigma = 0.212$). Overall, subject anaphors show a slightly higher rate of lexical expression if they occur in independent clauses (cross-corpus mean $m = 0.24$; $\sigma = 0.072$) compared to other clauses ($m = 0.22$; $\sigma = 0.22$), but the difference is not particularly large or robust (mean $\varphi = 0.09$; $\sigma = 0.099$). As Table 6.21 indicates, subjects in non-independent clauses have $e^{\beta} = 0.82$ ($p = 0.004$) times lower odds of being realized lexically.

In sum, the data do not indicate a strong association between clause type and lexicality. Given the choice between an independent and a dependent clause, the former appears slightly more likely to be the locus of more informative and more explicitly referring subject anaphors, but this might in fact be



		independent			other				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	54	147	0.37	32	294	0.11	0.31	<0.001
E	English	130	1119	0.12	33	236	0.14	0.03	0.310
M	Mandarin	188	676	0.28	20	43	0.47	0.10	0.009
N	Nafsan	121	562	0.22	21	136	0.15	0.06	0.113
K	N. Kurdish	78	396	0.20	60	247	0.24	0.05	0.168
S	S. Dargwa	56	235	0.24	60	289	0.21	0.04	0.400
B	Tabasaran	161	479	0.34	43	307	0.14	0.22	<0.001
T	Teop	155	699	0.22	12	52	0.23	0.01	0.880
L	Tulil	93	372	0.25	58	248	0.23	0.02	0.647
V	Vera'a	437	2214	0.20	55	216	0.25	0.04	0.046
	totals	1473	6899	—	394	2068	—	—	—

Figure 6.36 | Lexicality of anaphoric subjects by type of clause.
Not fully independent clauses ('other') ...

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	clause type	(*independent*, other)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.181
speaker	37	0.351

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.28	−1.270	0.093	−13.71	<0.001
(A_1)	clause type	= other	0.82	−0.202	0.070	−2.88	0.004

c. | scaled residuals

min.	lower	median	upper	max.
−0.705	−0.546	−0.476	−0.324	3.411

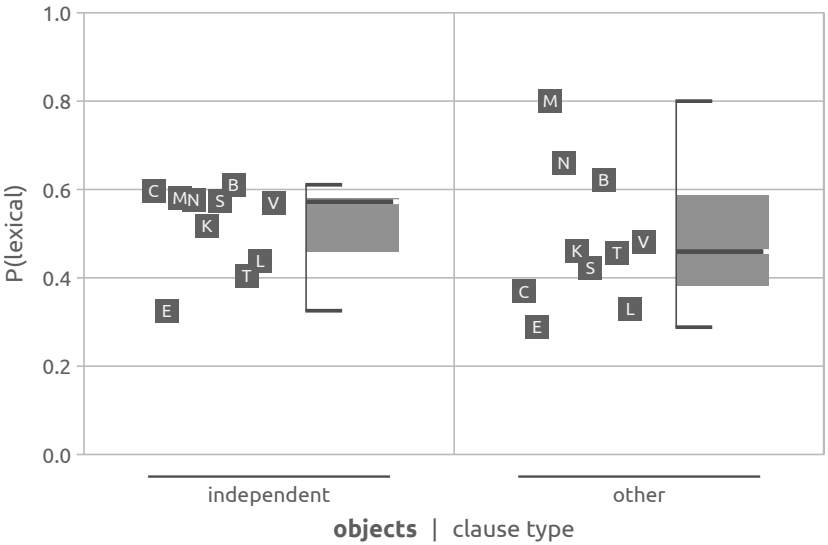
d. | correlation of fixed effects

	(intercept)
(A_1)	−0.183

e. | model evaluation

observations	8967	AIC	9038
model deviance	9030	log-likelihood	−4515
residual d.f.	8963	conditional R^2	0.047
		marginal R^2	0.002

Table 6.21 | Regression model results for the lexicality of anaphoric subjects by clause type, with corpus and speaker as random effects.



		independent			other				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	40	67	0.60	62	168	0.37	0.21	0.001
E	English	180	553	0.33	49	170	0.29	0.03	0.361
M	Mandarin	112	193	0.58	8	10	0.80	0.10	0.168
N	Nafsan	101	175	0.58	35	53	0.66	0.07	0.279
K	N. Kurdish	86	166	0.52	42	91	0.46	0.05	0.386
S	S. Dargwa	35	61	0.57	27	64	0.42	0.15	0.090
B	Tabasaran	91	149	0.61	59	95	0.62	0.01	0.872
T	Teop	89	220	0.40	16	35	0.46	0.04	0.557
L	Tulil	56	128	0.44	36	109	0.33	0.11	0.091
V	Vera'a	304	533	0.57	39	81	0.48	0.06	0.133
	totals	1094	2245	—	373	876	—	—	—

Figure 6.37 | Lexicality of anaphoric objects by type of clause.
Not fully independent clauses ('other') ...

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	clause type	(*independent*, other)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.254
speaker	37	0.573

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	1.09	0.087	0.141	0.62	0.537
(A_1)	clause type	= other	0.79	−0.238	0.091	−2.63	0.009

c. | scaled residuals

min.	lower	median	upper	max.
−1.800	−0.870	−0.620	0.966	1.675

d. | correlation of fixed effects

	(intercept)
(A_1)	−0.187

e. | model evaluation

observations	3121	AIC	4149
model deviance	4141	log-likelihood	−2070
residual d.f.	3117	conditional R^2	0.109
		marginal R^2	0.003

Table 6.22 | Regression model results for the lexicality of anaphoric objects by clause type, with corpus and speaker as random effects.

an indirect consequence of subordinate clauses commonly repeating referents from the matrix clause.

6.10.2.2 | Objects

The findings for object anaphors are summarized in Figure 6.37 and Table 6.22. From broad cross-linguistic perspective, independent clauses tend to have higher rate of lexical objects than other clause types in most corpora, same as with subjects above, with non-independent clauses having $e^{\beta} = 0.79$ ($p = 0.009$) times lower odds of having lexical objects than independent clauses. Overall, clause type makes either very little or no difference for the lexicality of object anaphors in half the corpora in the sample – Vera'a, Northern Kurdish, Teop, English, and especially Tabasaran (all $\varphi \leq 0.06$) – and not much more in the remainder.

6.11 | Clause length

6.11.1 | Definition and methodological issues

We discussed the effect of clause length on planning load and its concomitant effect on referential choice, as reported on in Arnold et al. (2009) and Arnold (2010), in Section 2.4.2.5. In this section, we will examine its effect on the selection of lexical expressions in particular.

The length of each clause in the sample is calculated as the total number of grammatical words contained in it, not counting subordinated clauses other than relative clauses. As defined in Section 4.6.1, clitics are counted as separate words, other bound forms are not. Zero anaphors naturally have a length of zero. Depending on which is being examined, the word length of the subject or object NP is subtracted from this word count.

The clause in the following example has total length of $l = 12$ grammatical words; as the subject NP itself is $l = 3$ words long and the object NP $l = 4$ words, the effective length of the clause is $l = 12 - 3 = 9$ if the focus is on the subject anaphor, and $l = 12 - 4 = 8$ words if it is on the object anaphor:

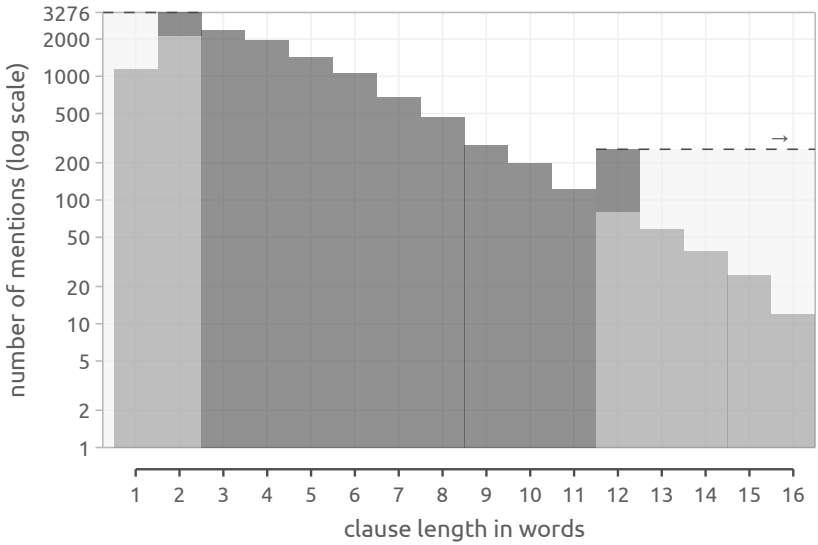


Figure 6.38 | Histogram of clause lengths measured for subject and objects anaphors.
The darker bars indicate the winsorized sample.

(79) Tulil

me abət o vənu laika bəvəppən ma məngəda abət.

	me	*a-bət*	*o*	*vənu*	*laik*	*=a*
	and	3SG.M-PROX2	TOP	sun	big	=SG.CL:MASC
##	other	dem_pro:dt_p	other	np:a	rn	=rn

bə=	*və-p pən-m*	*=a*	*məngəd*	*=a*	*a-bət*
ASP=	3SG.M.PST-RED~hit	=PAT	village	=SG.CL:MASC	3SG.M-PROX2
=other	v:pred	=ln	np:p	=rn	rn

'That (village), the strong sun was destroying the village.'

[mc_tulil_alrm_0009]

Like anaphoric distances (Section 6.3) and total mention frequencies (Section 6.2) discussed earlier, the distribution of clause lengths has a long tail. Shorter clause lengths are extremely frequent, and increasingly higher lengths less and less so, as Figure 6.38 illustrates. The single most common length value, excluding the length of the anaphor in question, is $l = 2$ words. Notably, clauses with a length of $l = 1$ word, that is, containing only a single other grammatical word besides the anaphor in question, as in (80), are only half as frequent as the next higher length group.

(80) Northern Kurdish
Dibe, tere, …

	di-b-e		*tere*
	IND-take.PRS-3SG		IND.go.PRS.3SG
`## 0:a 0:p`	`v:pred`	`## 0:s`	`v:pred`

'(She) takes (him) and (she) goes, …' [mc_nkurd_muserz03_0015]

The reason for this is that their incidence is necessarily limited to languages in which the verbal complex may consist of only a single word, that is, languages without any obligatory auxiliaries, cliticized person indices, and other such elements. Compare (80) to this example from Nafsan, where a minimal clause is composed of two grammatical words:

(81) Nafsan
isok, …

	i=	*sok*
	3S.RS=	jump
`## 0.d:s`	`=lv`	`v:pred`

'(The whale) jumped, …' [mc_nafsan_tafra_0020]

For reasons of cross-corpus comparability, lengths $l \leq 2$ have hence been merged into a single category.

In a similar vein, since very long clause lengths are uncommon individually, they have been winsorized at an (arbitrarily chosen) threshold of $l = 12$ words. In other words, lengths above $l > 12$ words are counted as occuring at that threshold instead.

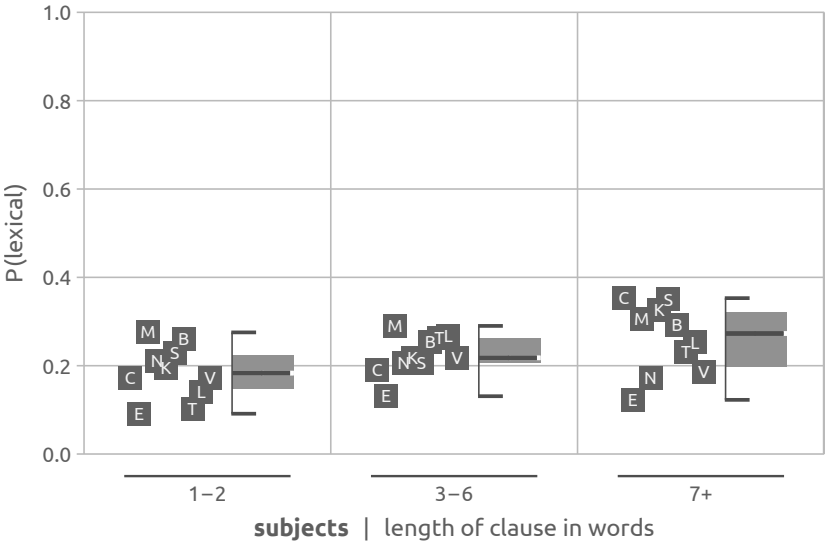
6.11.2 | Lexicality by clause length

For the purpose of visualization, the figures below group observations into three length categories: 1 to 2 words, 3 to 6 words, and 7 or more words. The intervals delineated by the latter two echo those selected in Arnold et al. (2009). As before, the regression model and any statistics reported in the text are not calculated on the basis of these groupings, but the actual (winsorized) values.

6.11.2.1 | Subjects

Figure 6.39 shows the distribution of lexical subject anaphors by clause length. From a very broad cross-corpus perspective, there is a very small association between lexicality and clause lengths (cross-corpus mean point-biserial correlation coefficient $r_{pb} = 0.04$; $\sigma = 0.044$), in that longer clauses are very slightly more likely to have lexical subjects. As the regression model summarized in Table 6.23 suggests, subjects have $e^{\beta} = 1.03$ times higher odds of being lexical for every word in the clause, a statistical significant ($p = 0.003$) but miniscule increase in odds. For a number of corpora, this association is particularly weak or even non-existent, as for instance in Mandarin ($r_{pb} = 0.01$), Vera'a ($r_{pb} = 0.01$), English ($r_{pb} = 0.02$), and Tabasaran ($r_{pb} = 0.02$); in fact, the correlation coefficient for Nafsan dips very slightly into the negative ($r_{pb} = -0.03$), counter to the general trend. Overall, there are no strongly pronounced patterns or notable outliers among the corpora; the distribution of corpora is relatively homogenous, reflecting only differences in lexical baseline levels.

In short, clause length appears to have only a marginal effect on the choice of form for subjects. Arnold et al. (2009: 11–12) observe roughly a 10% difference between clauses more and fewer than $l = 7$ words in length in the selection of pronominal and lexical forms in their English data, with longer clauses leading leading to a preference for reduced forms. In the Multi-CAST English corpus, the same distinction in clause length amounts to only about a 0.3% difference in lexicality ($P = 0.120$ for $l < 7$ words length; $P = 0.123$ for $l \geq 7$ words) when tested across all subject mentions, and furthermore points in the opposite direction.



		1–2 words			3–6 words			7+ words		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	30	175	0.17	44	232	0.19	12	34	0.35
E	English	27	296	0.09	97	741	0.13	39	318	0.12
M	Mandarin	57	207	0.28	105	362	0.29	46	150	0.31
N	Nafsan	29	137	0.21	102	497	0.21	11	64	0.17
K	N. Kurdish	54	278	0.19	70	322	0.22	14	43	0.33
S	S. Dargwa	52	228	0.23	57	276	0.21	7	20	0.35
B	Tabasaran	92	354	0.26	95	374	0.25	17	58	0.29
T	Teop	16	158	0.10	113	429	0.26	38	164	0.23
L	Tulil	13	93	0.14	90	337	0.27	48	190	0.25
V	Vera'a	83	482	0.17	319	1464	0.22	90	484	0.19
	totals	453	2408	—	1092	5034	—	322	1525	—

Figure 6.39 | Lexicality of anaphoric subjects by length of clause, measured in grammatical words.
Length values are minus the length of the subject NP itself.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	clause length	(*1;2*–12+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.171
speaker	37	0.360

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.25	–1.402	0.094	–14.90	<0.001
(A)	clause length	* [0, 10]	1.03	0.034	0.011	3.01	0.003

c. | scaled residuals

min.	lower	median	upper	max.
–0.797	–0.548	–0.477	–0.349	3.327

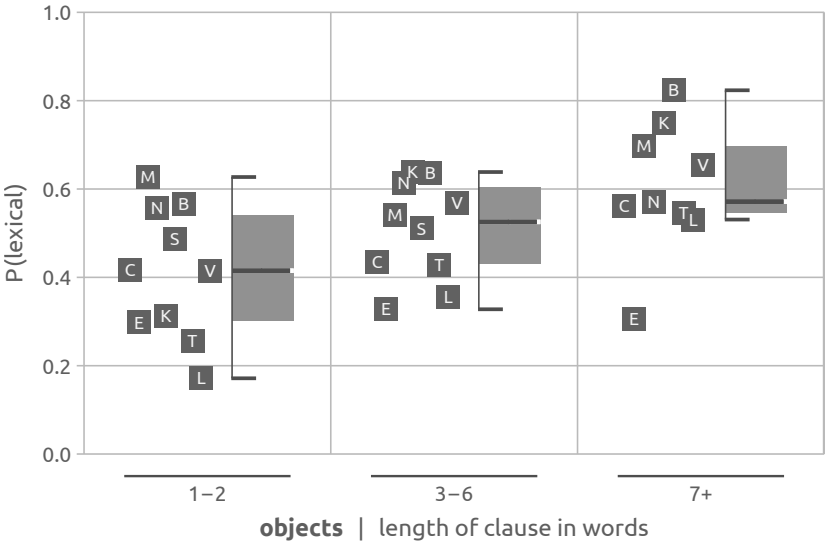
d. | correlation of fixed effects

	(intercept)
(A)	0.000

e. | model evaluation

observations	8967	AIC	9037
model deviance	9029	log-likelihood	–4515
residual d.f.	8963	conditional R^2	0.048
		marginal R^2	0.002

Table 6.23 | Regression model results for the lexicality of anaphoric subjects by the length of the clause in words, with corpus and speaker as random effects.



		1–2 words			3–6 words			7+ words		
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)
C	C. Greek	47	113	0.42	46	106	0.43	9	16	0.56
E	English	42	141	0.30	136	415	0.33	51	167	0.31
M	Mandarin	37	59	0.63	60	111	0.54	23	33	0.70
N	Nafsan	34	61	0.56	94	153	0.61	8	14	0.57
K	N. Kurdish	36	115	0.31	83	130	0.64	9	12	0.75
S	S. Dargwa	38	78	0.49	24	47	0.51	0	0	—
B	Tabasaran	68	120	0.57	68	107	0.64	14	17	0.82
T	Teop	12	47	0.26	75	175	0.43	18	33	0.55
L	Tulil	6	35	0.17	43	121	0.36	43	81	0.53
V	Vera'a	41	99	0.41	232	408	0.57	70	107	0.65
	totals	361	868	—	861	1773	—	245	480	—

Figure 6.40 | Lexicality of anaphoric objects by length of clause, measured in grammatical words.
Length values are minus the length of the objects NP itself.

objects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	clause length	(*1;2*–12+)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.332
speaker	37	0.576

b. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.81	–0.213	0.161	–1.32	0.185
(A)	clause length	* [0, 10]	1.11	0.107	0.017	6.10	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
–2.224	–0.883	–0.584	0.965	1.828

d. | correlation of fixed effects

	(intercept)
(A)	–0.001

e. | model evaluation

observations	3121	AIC	4118
model deviance	4110	log-likelihood	–2055
residual d.f.	3117	conditional R^2	0.133
		marginal R^2	0.016

Table 6.24 | Regression model results for the lexicality of anaphoric objects by the length of the clause in words, with corpus and speaker as random effects.

	corpus	r_{pb}	p-value
C	C. Greek	0.12	0.011
E	English	0.02	0.409
M	Mandarin	0.01	0.822
N	Nafsan	–0.03	0.501
K	N. Kurdish	0.07	0.077
S	S. Dargwa	0.04	0.319
B	Tabasaran	0.02	0.592
T	Teop	0.09	0.015
L	Tulil	0.02	0.617
V	Vera'a	0.01	0.686

Table 6.25 | Point-biserial correlation coefficients for lexicality of anaphoric subjects and clause length (ungrouped).

	corpus	r_{pb}	p-value
C	C. Greek	0.08	0.248
E	English	0.00	0.917
M	Mandarin	0.10	0.169
N	Nafsan	0.02	0.811
K	N. Kurdish	0.27	<0.001
S	S. Dargwa	0.05	0.595
B	Tabasaran	0.17	0.009
T	Teop	0.13	0.034
L	Tulil	0.28	<0.001
V	Vera'a	0.15	<0.001

Table 6.26 | Point-biserial correlation coefficients for lexicality of anaphoric objects and clause length (ungrouped).

6.11.2.2 | Objects

The data for objects are illustrated in Figure 6.40. Note, first of all, clauses with lengths of $l \geq 7$ and above words are considerably less frequent for objects than they are for subjects. In particular, there are no instances of objects in this length category in Sanzhi Dargwa, and very few observations in most other corpora. One explanation for this is that since objects are overall more likely to be lexical than subjects, and lexical expressions are longer than non-lexical expressions, object NPs tend to make up a larger fraction of a clause's length than subjects, and since the length of object phrase itself is subtracted from the overall clause length, the remainder tends to be quite short. Similarly, as objects may co-occur with an overt subject NP in the same clause, clause lengths below $l \leq 2$ are likewise rarer than they are for subjects. Note however that this sparsity of data is not a problem for the statistical analyses and regression models, as the data is grouped in this way only in Figure 6.40 for presentational purposes.

Taking into account the generally greater spread of lexicality rates among objects, longer clauses are overall more likely to host lexical objects. This association is stronger than for subjects, with Northern Kurdish ($r_{pb} = 0.27$) and Tulil ($r_{pb} = 0.28$) being particularly notable examples. That being said, in a number of corpora lexicality rates for objects are unaffected by increases in clause length; in particular, these are English ($r_{pb} = 0.00$) and Cypriot Greek ($r_{pb} = 0.08$), two corpora for which subjects likewise show little association. Table 6.24 indicates $e^{\beta} = 1.11$ ($p < 0.001$) times higher odds of a lexical object anaphor for every word increase in clause length, which is marginally higher than the increase in odds for subjects. As such, a clause containing $l = 7$ other words has $e^{\beta} = 1.11^{7} = 2.11$ times higher odds than a clause containing only $l = 1$ word; a clause with $l = 10$ words $e^{\beta} = 1.11^{10} = 2.91$ times higher odds.

6.12 | Transitivity

6.12.1 | Definition and methodological issues

This section examines what difference transitivity makes for the selection of lexical forms for subject anaphors. The criteria for defining transitivity and for distinguishing transitive from intransitive clauses were already given in Section 4.3. Specifically, we distinguish

a. transitive from
b. all other types of clauses.

The latter category is mostly composed of intransitive clauses, in addition to a small number of other clause types that do not fully meet the criteria for transitivity, such as those with quirky subject case.

In addition to the basic association with lexicality in Section 6.12.2, this section briefly discusses the interaction between lexicality, transitivity, and humanness (Section 6.12.3), anaphoric distance (Section 6.12.4), clause type (Section 6.12.5), and clause length (Section 6.12.6).

6.12.2 | Lexicality of subjects by transitivity

Note, first of all, that the relative proportions of transitive and intransitive clauses is remarkably stable across the ten corpora (cross-corpus mean $P = 0.38$, $\sigma = 0.065$), as visualized in Figure 6.41. This observation has some implications for the observed baseline rate of lexical subjects (Section 5.2.1), which we return to discuss in greater detail in Section 9.3.1. This cross-corpus stability in transitivity is not readily accountable for; the selection of transitive clauses is presumably based on fairly localized discourse criteria (i.e. largely content-dependent), rather than on global discourse planning processes, and there are no obvious candidates for local constraints that could yield such a stable profile. There is, however, a notable degree of variation among the texts within each corpus (with a total range of $P = 0.11$ to $P = 0.63$), which is expected under the assumption that transitive constructions express certain types of events, and that the individual texts in the corpora differ in content and hence in the types of events expressed. As such, as with the lexical baseline rate, this relatively stable discourse profile only emerges when averaging over multiple texts produced by different speakers.

Figure 6.42 shows the proportions of lexical subjects in transitive and intransitive clauses in the sample. Across corpora, intransitive clauses are more likely to have lexically realized subjects than transitive clauses; as already mentioned above, this pattern is both persistent and fairly consistent in all corpora.

The mixed-effects regression model with a transitivity as the sole fixed effect summarized in Table 6.27 indicates the odds of a lexical form for the subject of an intransitive clause are $e^{\beta} = 1.99$ ($p < 0.001$) times higher compared

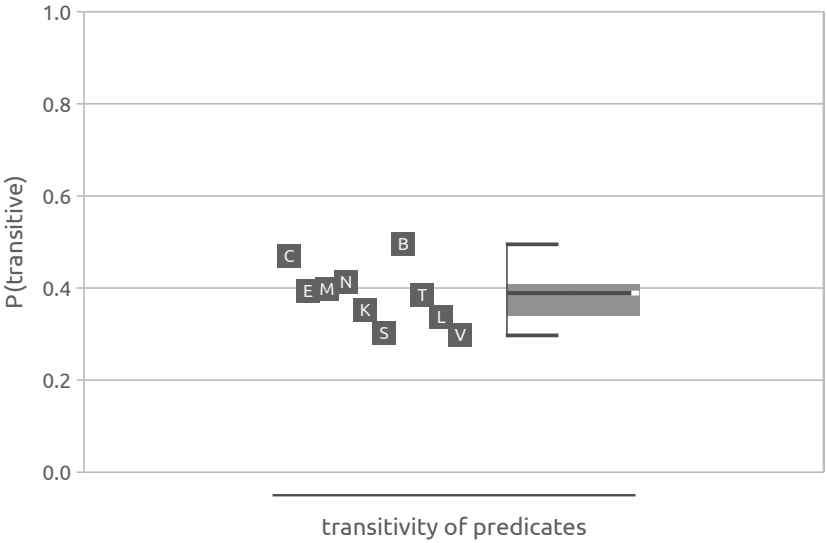
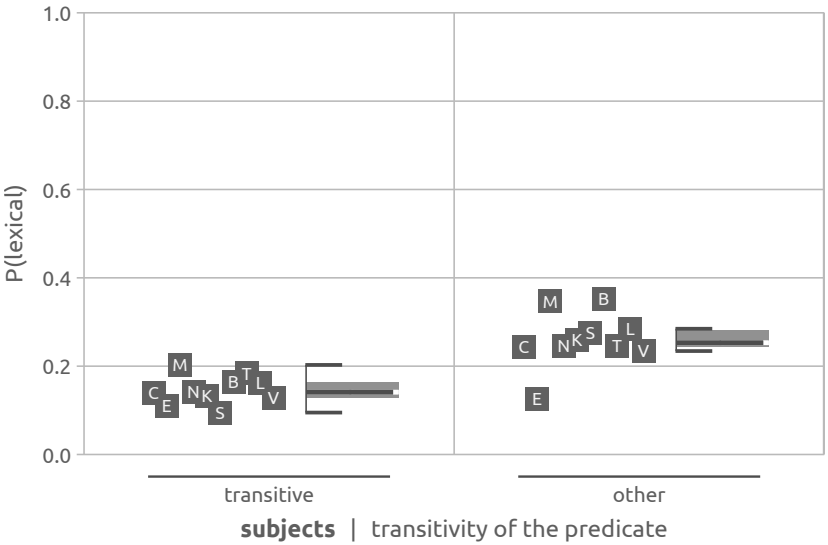


Figure 6.41 | Proportions of transitive clauses among all clauses in the ten corpora. Only clauses with subjects that meet the general selection criteria in Chapter 4 are counted.

to those of transitive clause. It also suggests a fair degree of homogeneity among the two random effects in the model: The association between transitivity and lexicality does not vary much by corpus (intercept $\sigma = 0.202$) or by speaker ($\sigma = 0.337$). In fact there is very little inter-corpus variation among transitives in particular (cross-corpus mean $P = 0.15$, $\sigma = 0.033$). As such, the well-known constraint on lexical subjects in transitive clauses is well represented in all corpora in the sample; see the discussion of Du Bois (1987b) in Section 2.2.5.1 above. Intransitives also show a comparatively narrow spread, but with a few outliers. Tabasaran ($\varphi = 0.21$, $p < 0.001$) and Mandarin ($\varphi = 0.16$, $p < 0.001$) have lexical subjects in intransitive clauses at higher rates than the cross-corpus mean, whereas subjects in the English ($\varphi = 0.02$, $p < 0.382$) and Teop ($\varphi = 0.07$, $p < 0.042$) corpora are largely indifferent towards transitivity. As we have seen numerous times in this chapter, the insensitivity of the Eng-



		transitive			other				
	corpus	N(lex)	N(all)	P(lex)	N(lex)	N(all)	P(lex)	φ	*p*-val.
C	C. Greek	29	207	0.14	57	234	0.24	0.13	0.006
E	English	59	533	0.11	104	822	0.13	0.02	0.382
M	Mandarin	58	286	0.20	150	433	0.35	0.16	<0.001
N	Nafsan	41	288	0.14	101	410	0.25	0.13	0.001
K	N. Kurdish	30	226	0.13	108	417	0.26	0.15	<0.001
S	S. Dargwa	15	158	0.09	101	366	0.28	0.20	<0.001
B	Tabasaran	64	389	0.16	140	397	0.35	0.21	<0.001
T	Teop	53	289	0.18	114	462	0.25	0.07	0.042
L	Tulil	34	209	0.16	117	411	0.28	0.13	0.001
V	Vera'a	92	722	0.13	400	1708	0.23	0.12	<0.001
	totals	475	3307	—	1392	5660	—	—	—

Figure 6.42 | Lexicality of anaphoric subjects by transitivity.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	lexicality	(*non-lexical*, lexical)
fixed effect	transitivity	(*transitive*, intransitive)
random effects	corpus	
	speaker	

a. | random effect intercepts

	groups	σ
corpus	10	0.202
speaker	37	0.337

b. | fixed effect coefficients

			e^{β}	β	SE	z-val.	p-val.
	(intercept)	—	0.17	−1.781	0.103	−17.22	<0.001
(A_1)	transitivity	= intransitive	1.99	0.688	0.060	11.54	<0.001

c. | scaled residuals

min.	lower	median	upper	max.
−0.787	−0.568	−0.460	−0.327	3.950

d. | correlation of fixed effects

	(intercept)
(A_1)	−0.003

e. | model evaluation

observations	8967	AIC	8905
model deviance	8897	log-likelihood	−4449
residual d.f.	8963	conditional R^2	0.074
		marginal R^2	0.031

Table 6.27 | Regression model results for the lexicality of anaphoric subjects by transitivity, with corpus and speaker as random effects.

lish corpus to many of the tested factors is a notable characteristic of the data, and one we will discuss more extensively in Chapter 9.

In sum, as objects are overall more likely to be lexical than not, these observations appear to point towards a sort of "one lexical expression per clause" tendency in the vein of Du Bois (1987b, 2003b, 2017; cf. also the "one new idea at a time" constraint in Chafe 1976). However, as we shall see in the next section, the magnitude of this effect reduces substantially when animacy is factored in, as transitive subjects are overwhelmingly human.

6.12.3 | Interaction with humanness

Haig & Schnell (2016) argue that the differences in lexicality rates between subjects of transitive and intransitive clauses can be explained as side-effect of the association with humanness (cf. also Everett 2009). In the Multi-CAST corpora, the vast majority of transitive subjects are human (cross-corpus mean $P = 0.93$, $\sigma = 0.071$); while intransitive subjects are likewise mostly human, the proportion is noticeably lower ($P = 0.81$, $\sigma = 0.133$). Given the association between humanness and lexicality we have noted earlier in Section 6.1, then, it does not come as a surprise that Figure 6.43 shows that across corpora, the subjects of both transitive and intransitive clauses are much less likely to be lexical if they are human, with very little cross-corpus variability. Conversely, there is a lot of variation among corpora for non-human referents. As such, the cross-linguistically stable differences in lexicality rates seen in Figure 6.32 are not so much the result of morphosyntactic differences between transitive and intransitive clauses (as claimed in Du Bois 1987b, 2003a, 2017), but rather a side-effect of the strong association between humanness and agentivity. The outliers among non-human subjects in intransitive clauses in Figure 6.43 are English ($P = 0.13$), Tabasaran ($P = 0.78$), and Northern Kurdish ($P = 0.82$), which showed likewise patterns in Section 6.1 above.

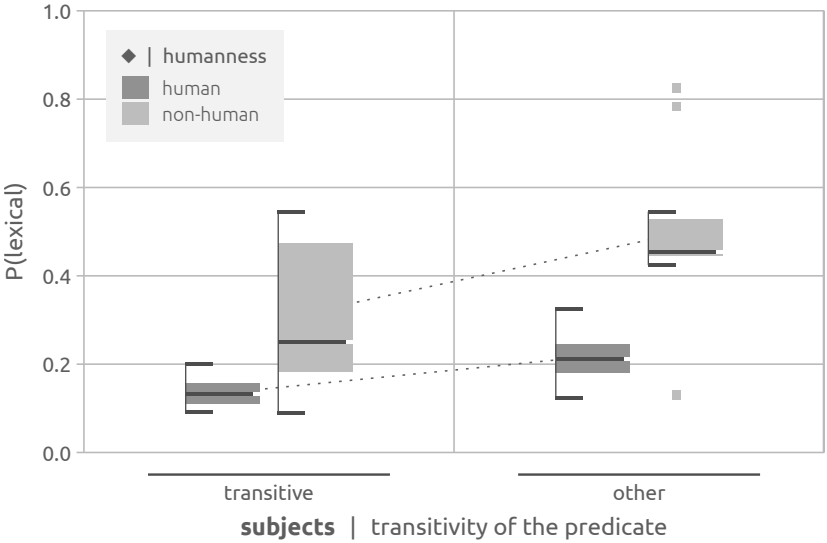


Figure 6.43 | Interaction of the effect of transitivity and humanness on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

6.12.4 | Interaction with anaphoric distance

The interaction between transitivity and anaphoric distance (Section 6.3) is shown in Figure 6.44. We find similar spread in lexicality rates across corpora for subjects of both transitive and intransitive clauses, especially at intermediate distances. As noted above, this gives further indication of transitivity not being highly determinative of lexical expression cross-linguistically. However, what is notable here is that the relative difference between the rates for transitive and intransitive clauses by and large remains the same irrespective of the distance to the antecedent. This suggests that while transitivity itself does not impact lexical choices as much as discourse-contextual factors such as anaphoric distance do, its influence is ubiquitous, applying to anaphors in any context, of which the ones delineated by anaphoric distance are among the most distinctive.

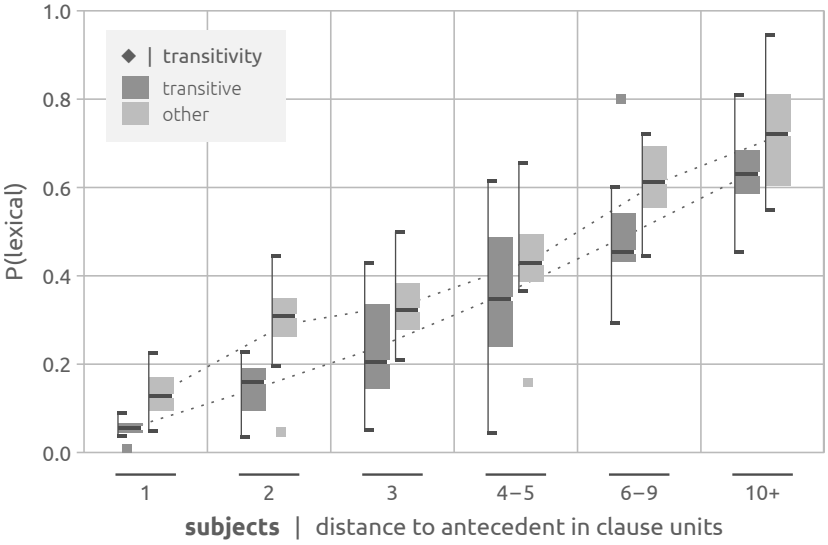


Figure 6.44 | Interaction of the effect of transitivity and anaphoric distance on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

6.12.5 | Interaction with clause type

Figure 6.45 shows the interaction between transitivity and clause type (Section 6.10) and their combined effect on the lexicality of subject anaphors. The two outliers are in both cases Mandarin, which has a characteristically low rate of syntactic subordination. As the figure indicates, no special association between the two factors is apparent: The subjects of transitive clauses are less likely to be realized lexically, as noted above, as are the subjects of subordinated and other not-fully-independent clauses. The effects are in essence additive, in that no combination of their levels converges on a notably higher or lower rate of lexical expression.

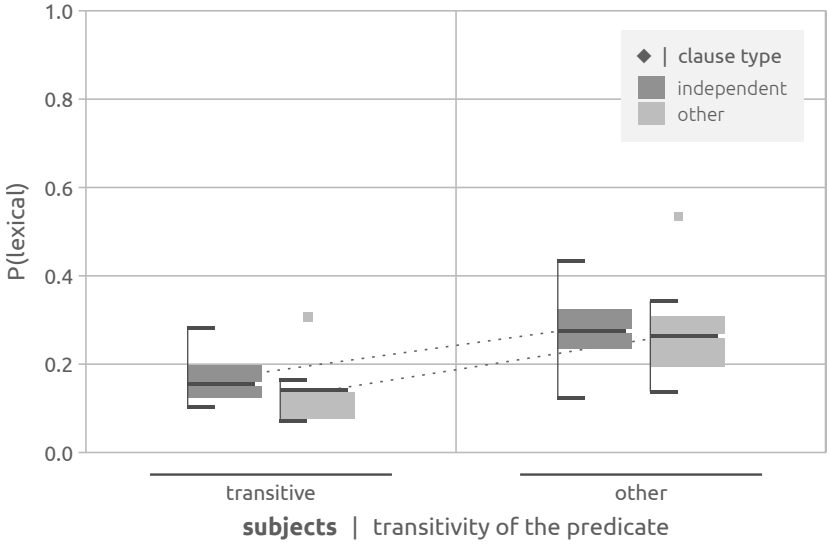


Figure 6.45 | Interaction of the effect of transitivity and clause type on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

6.12.6 | Interaction with clause length

Figure 6.46 illustrates another interaction with transitivity, this time with clause length (measured in words, Section 6.11). As observed earlier in Section 6.11, longer clauses are slightly correlated with higher rates of both lexical subjects and objects, although the effect is not particularly strong when examined across all mentions in the sample. The interaction shown in Figure 6.46 further adds to this picture, indicating that the two factors impact lexicality rates independently: Transitive clauses are not noticeably more or less likely to host lexical subjects whether they are longer (i.e. with lexical objects and additional constituents) or shorter (i.e. with non-lexical objects and no further elements).

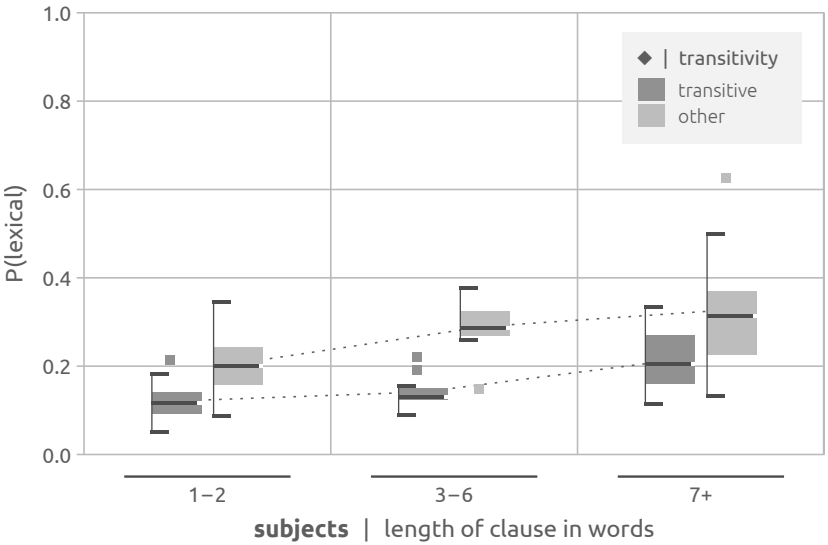


Figure 6.46 | Interaction of the effect of transitivity and clause length on the lexicality of anaphoric subjects, by corpus.
The small squares indicate outliers from the central distribution.

This contradicts findings in Arnold et al. (2009) and Arnold (2010), as noted above, as well as what would be expected based on the non-lexical-A constraint (Du Bois 1987b, see Section 2.2.5.1). Assuming that, all else being equal, transitive clauses are longer than intransitive ones due to the presence of another argument (which also happens to frequently be lexical), we would expect to find that this would make the subjects of longer clauses (such as transitive ones) less likely to be lexical overall, which is not an association we find represented in the data. As it stands, transitivity is not strongly associated with clause length in the sample data (cross-corpus mean $r = -0.22$, $\sigma = 0.077$) except in very short clauses of two or fewer words, see Figure 6.47.

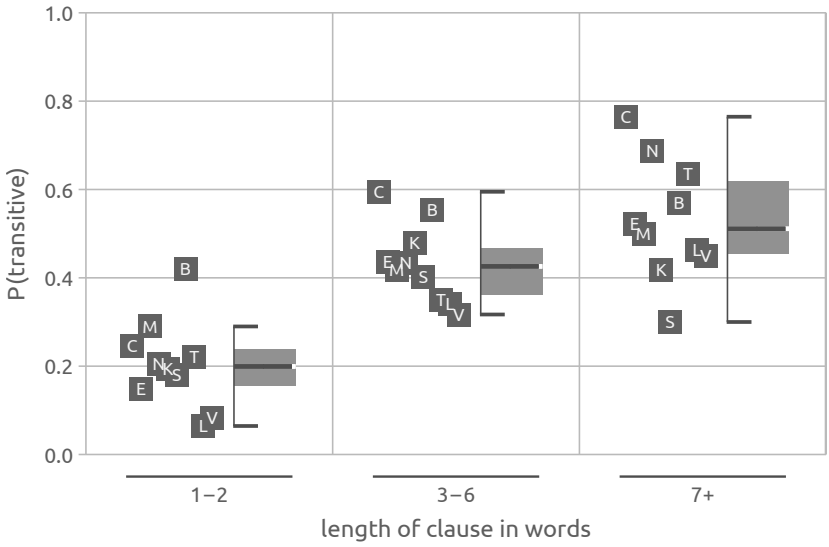


Figure 6.47 | Proportion of transitive clauses by length of the clause in words.

7 | Multifactorial analysis

The previous chapter has looked at each factor individually, in addition to a small selection of interactions between multiple factors. This chapter brings everything together into a pair of multifactorial models, primarily for the purposes of explaining the variation present in referential choices for subjects and objects, but secondarily also for predicting specific referential outcomes.

As already discussed in Chapter 5, subjects and objects show substantial differences in lexicality rates. There also appears to be considerably more variation in these rates both within and between corpora among objects, which suggests that language-specific (as well as content-dependent) factors affect anaphors in different positions differently. As we have seen in Sections 6.3, 6.7, and 6.8, subjects and objects behave quite differently in contexts of low anaphoric distance, especially with non-lexical antecedents in a prominent position, as these contexts are substantially more frequent for subjects than for objects. These differences appear to have an underlying cognitive basis (Wang et al. 2009; Kwon et al. 2010). For these reasons, we will in the following examine lexical choices in subject and object position separately, as essentially independent questions. The alternative would be to introduce role as a facetting factor into the models; this has the disadvantage that third-person anaphors in subject position outnumber those in object position by about a factor of two, which would lead to outcomes heavily skewed towards patterns specific to subjects. Running separate models conversely has the advantage of allowing us to compare role-specific associations in more detail, especially those that might otherwise be obscured by highly predictive patterns.

This chapter has two parts, each of which deals with subject anaphors and objects anaphors in turn: In the first (Section 7.1), we will examine single decision trees built from the entirety of the sample data. Tree models have in recent years become increasingly popular in quantitative linguistics, either as an alternative or as a complement to traditional regression models (Baayen & Arppe 2011; Baayen et al. 2013; Walker et al. 2015; Gries 2019). They provide a nicely illustrative and relatively intuitive perspective on complex relationships in the data, and are particularly helpful for identifying otherwise overlooked high-level interactions between factors.

The second part (Section 7.2) addresses the shortcomings of the first via the use of so-called ensemble methods, which build on the successive generation of thousands of decision trees to reduce the chance of misclassifying outcomes. Ensemble methods come from the toolbox of machine learning and related fields, and are still relatively uncommon in linguistics (see Gries 2019 for a recent review). They are less accessible than single-tree models but also substantially more powerful, providing useful insights into the relationships inherent in the data. With the aid of these models, this chapter evaluates the relative influence of factors on the selection of lexical expressions, both independently and across interactions (Sections 7.2.3.2–7.2.3.4 and 7.2.4.2–7.2.4.4). It also examines the predictions made by the models on the basis of fresh data sets, finding them to perform slightly better than the baseline rate (Sections 7.2.3.5 and 7.2.4.5). For multilingual spoken data as heavily imbalanced as this (i.e. since non-lexical subjects in particular are much more common than lexical ones), this is a good result, as it means that the factors selected for the model accurately capture some – though not all – of the motivators for lexical anaphora across languages.

7.1 | Single decision trees

7.1.1 | Introduction

Classification and regression trees (CART, based on the algorithm in Breiman et al. 1984) are a method of predictive modelling that results in a visual structure that can be traversed from root to leaf analogously to other decision graphs. Classification trees apply to discrete response variables, regression trees to continuous ones; as we are here modelling the binary choice between

lexical and non-lexical expressions, we are only concerned with the former. Classification trees successively subdivide data sets along values of specific predictors in a way that results in the greatest purity of one or the other response value in the terminal nodes. At each step, the algorithm selects the level of a predictor that most clearly divides up the data along the response variable: Each subdivision, called a "split" in CART parlance, is chosen to minimize the misclassification error, that is the proportion of cases that is assigned to the incorrect outcome group. This proceduce is repeated along each branch of the growing tree until the algorithm runs out of useful splits. Each path along the branches from root to leaf yields a high-order interaction of the specific predictors involved. How much a tree is allowed to grow is defined by a set of hyperparameters (given in Section 7.1.2.3 for the trees in this section), which place constraints on the tree's complexity.

The result of this procedure is a tree composed of binary splits, in which each of the leaf nodes represent a subset of the observations. Ideally, the leaves contain as many of one type of outcome (i.e. in this case, either lexical or non-lexical expressions) and as few of the other as possible: The higher the proportion of one outcome, the more strongly the particular combination of factor levels running from the root of the tree down to the leaf node is associated with that outcome. Conversely, leaves with outcome ratios closer to 1:1 are more ambigious, indicating a weak association between factor levels. The path from root to leaf represents a *k*-fold interaction between the predictors selected for the splits, where *k* is the depth of the tree. This makes decision trees very useful for identifying complex interactions that would otherwise have gone unnoticed.

But while tree-based models are relatively intuitive to understand and interpret, they also have a number of shortcomings. For one, they are liable to overfit the data (Kuhn & Johnson 2013: 182), and their structure is somewhat volatile, especially the further down along their branches one goes (Strobl et al. 2009: 332). Small changes in the underlying data or the hyperparameters can substantially affect the shape of the tree. For another, the tree-building algorithm is designed to select splits opportunistically, with maximum greed and no foresight: It always selects whichever split offers the best result at each step, meaning it ignores splits that are worse at first glance, but actually provide better overall results in combination with later splits (Strobl et al. 2009: 333; Gries 2019: 13). Lastly, note that the same predictor can be selected for splits multiple times with different levels in the same tree. For

this reason, tree-based models have a tendency to be slightly biased towards predictors with more levels, as they have more opportunities to be selected compared to binary variables, which can only ever appear once. The first three of these shortcomings are addressed in Section 7.2 via the use of more advanced models that aggregate more than one tree; as for the third, it is important to keep the bias towards scalar and categorical factors in mind when interpreting the models presented both in this section and the next.

In the following sections, we will discuss predictor and hyperparameter selection (Sections 7.1.2.1 and 7.1.2.3) as well as the issue of balancing common and rare events against each other (Section 7.1.2.2) before briefly examining two trees fit to the data in Section 7.1.3. Keep in mind that since the primary purpose of this section is to lay the groundwork for the analyses in Section 7.2, the trees shown below should be understood as little more than illustrative examples. For this reason, we will also pass on any steps that might further improve the explanatory power of these trees, such as tweaking the hyperparameter or pruning the trees.

7.1.2 | Model design

The trees below are calculated with the *rpart* package (Therneau & Atkinson 2019) for R; the visualizations are based on functions from the *partykit* package (Hothorn & Zeileis 2015).

7.1.2.1 | Factor selection

Unlike traditional regression models, where careful factor selection is paramount for building a robust model, approaches based on decision trees will autonomously disregard any factors that do not provide useful splits in the data. The trees and models in this section and the next therefore include all of the twelve factors listed in Section 4.5, excluding transitivity for objects anaphors:

1. humanness (*human*, non-human)
2. total mention frequency (*5% most*, 95% least frequent)
3. anaphoric distance in clauses (*1* – 10+)
4. recent co-referential mentions (*0* – 5+)
5. recent related mentions (*0* – 2+)
6. recent competing mentions (*0* – 6+)

7.	antecedent role	(*subject*, object, other)
8.	antecedent form	(*non-lexical*, lexical)
9.	sequence of mention	(*second* < third < subsequent)
10.	clause type	(*independent*, other)
11.	clause length in words	(*[1,2]* – 12+)
12.	transitivity	(*transitive*, intransitive)
13.	corpus	(10 corpora)

Where in Chapter 6 differences between corpora were checked for as a random effect in the mixed-effects regression models, corpus is here directly added to the model as a predictor. If major splits in the tree occur between different corpora, then the variation between those corpora supercedes the effects of later splits for other variables; conversely, splits occuring earlier in the tree are more likely to apply equally across all corpora. For sake of simplicity, we gloss over variation between speakers (and individual texts) from this point onwards, as the large number of factor levels (37 speakers, 54 texts) would not work well with tree-based models. As discussed earlier in Section 5.4 in the context of lexicality rates, intra-corpus variability for subjects in particular is no greater than inter-corpus variability; while object anaphors are subject to a greater degree of inter-speaker variation, they are nevertheless homogenous enough to simplify the models by excluding speaker identity as a variable.

7.1.2.2 | Balancing rare events

Lexical and non-lexical expressions are not equally frequent in the sample, or indeed in natural discourse in general. This is especially true for mentions in subject position, where non-lexical expressions greatly outnumber lexical ones: Lexical expressions account for only 21% of subject mentions. Object mentions are closer to an equal split, with 47% being lexical across all corpora.

Strong imbalances in the response variable pose a problem for the analysis, as outcomes are liable to be overwhelmingly influenced by the more common response, when it is usually the rarer response that is of interest (Menardi & Torelli 2012). This also affects the accuracy of predictions: A naïve model could technically achieve 79% accuracy for predicting the form of subject anaphors simply by always guessing a non-lexical outcome and disregarding the comparatively rarer possibility of a lexical form.

Dealing with rare events and unbalanced data is a well-known issue in statistical analysis, and there are numerous ways to address it, all with their own

advantages and disadvantages. For this study, we forgo options that alter the underlying data (such as resampling) in favour of assigning different weights to different cases at the modelling stage. In essence, case weights function as a penalty on misclassification, and are chosen so as to be proportional to the relative rarity of the case. They tell the modelling algorithm that the misclassification of rarer cases should be avoided in accordance with their heavier weight, making it harder for them to be glossed over in favour of the more common outcomes.

In addition to the response variable, that is lexicality, the data are also imbalanced in terms of the proportions of different corpora, with some, such as English and Vera'a, contributing a larger share than others. The motivation for rebalancing the corpora is the same as above: We would prefer not to change the data by synthesizing or cutting observations, but we also need to avoid giving larger corpora undue influence over the results.

All observations in the data are thus assigned weights equal to the relative proportions of the cross-section between lexicality and corpus, with smaller groups receiving higher weights. Weights are determined separately for the subject and object models. The most populous group always has a weight of one, and the number of cases in each group times their weights is the same for all groups. The exact values are given in Tables 7.1 and 7.2 for subjects and objects.

7.1.2.3 | Model hyperparameters

In addition to the choice of predictors, the form of classification trees is defined by a number of parameters that limit their maximum complexity. The following values were selected for the illustrational trees in the next section:

- learning rate $lr = 0.001$
- interaction depth $tc = 5$ (subjects)
  $tc = 4$ (objects)[1]
- min. observations in splits $N_{split} = 200$
- min. observations in nodes $N_{node} = 50$
- cross-validation folds $N_{cv} = 100$

1 A slightly lower value is used for the object tree for presentational reasons.

		non-lexical		lexical	
		N(cases)	weight	N(cases)	weight
C	C. Greek	355	5.46	86	22.53
E	English	1192	1.63	163	11.89
M	Mandarin	511	3.79	208	9.32
N	Nafsan	556	3.49	142	13.65
K	N. Kurdish	505	3.84	138	14.04
S	S. Dargwa	408	4.75	116	16.71
B	Tabasaran	582	3.33	204	9.50
T	Teop	584	3.32	167	11.60
L	Tulil	469	4.13	151	12.83
V	Vera'a	1938	1.00	492	3.94

Table 7.1 | Relative case weights assigned to subject anaphors in the tree models. The reference level has a weight of $w = 1.0$.

		non-lexical		lexical	
		N(cases)	weight	N(cases)	weight
C	C. Greek	133	3.71	102	4.84
E	English	494	1.00	229	2.16
M	Mandarin	83	5.95	120	4.12
N	Nafsan	92	5.37	136	3.63
K	N. Kurdish	129	3.83	128	3.86
S	S. Dargwa	63	7.84	62	7.97
B	Tabasaran	94	5.26	150	3.29
T	Teop	150	3.29	105	4.70
L	Tulil	145	3.41	92	5.37
V	Vera'a	271	1.82	343	1.44

Table 7.2 | Relative case weights assigned to objects anaphors in the tree models. The reference level has a weight of $w = 1.0$.

The Gini-coefficient is used as the splitting criterion for the trees. The learning rate *lr* determines the contribution of each generated tree as it is added to the model (specifically the minimum reduction in relative error), with lower learning rates resulting in better and more reliable estimates of the response variable, but requiring greater numbers of trees and longer computation times. Any split that does not improve the model fit by at least a factor of *lr* is not attempted by the algorithm. The maximum interaction depth *tc* constrains the generated trees' complexity by allowing them to grow at most *tc* interactions ("levels") deep: An interaction depth of $tc = 5$ allows for up to five-way interactions to be fitted, with the root node counting as interaction 0. The minimum observations in each split and terminal node likewise limit complexity, as splits that result in fewer than $N_{split} < 200$ observations in a split and $N_{node} < 50$ in a terminal node are not attempted. This also limits the influence of highly differentiated but extremely rare subgroups.

7.1.3 | Example decision trees

In the two trees shown below, the splits and terminal nodes are numbered sequentially in the order they were undertaken by the algorithm. Starting at the root node – the single most distinctive split – each branch is labelled with the respective classifier, that is the exact level (or levels) of the predictor at which the split occurs. The terminal nodes at the bottom of the trees show the total number of observations in each group (N) and the misclassification error (E) which indicates the relative homogeneity of the group: The closer the error is to 0, the greater is the purity of the node. Conversely, an error of 0.5 indicates an even distribution of the two outcomes, meaning they are in essence selected randomly. For the subject data in particular, however, it is important to keep in mind that two outcomes (non-lexical and lexical) are not balanced when interpreting misclassification errors. This is less of a concern for objects, where both outcomes are more closely balanced. The bar plots underneath the terminal nodes visualize the relative purity of the node by showing the proportion of lexical expressions in each node.

To reiterate what was said at the beginning of this section, single-tree models have limited utility due to their volatility – even small changes in data can alter the shape of the trees – as well as their sensitivity to the hyperparameters used to grow them and their tendency towards overfitting. As such, it is best to accept the following observations with a certain degree of scepticism. The

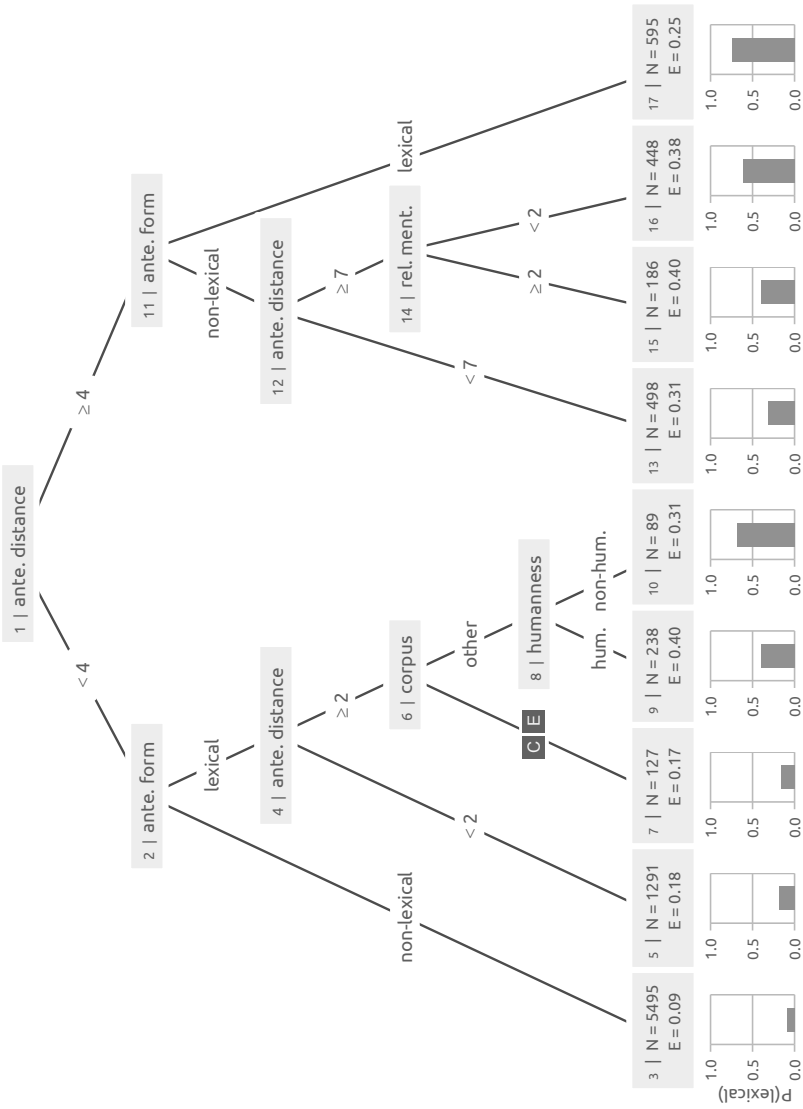


Figure 7.1 | An example form classification tree for subjects.
In the terminal nodes, 'N' is the size of the subclass and 'E' is the misclassification error. See the accompanying text for details on the model specifications.

ensemble methods used for the analysis in next section address the issues with single-tree models, offering more powerful and robust insights at the cost of higher interpretative complexity. With this disclaimer out of the way, let's look at some trees.

7.1.3.1 | Subjects

Figure 7.1 is the decision tree for subject anaphors. The first split is by anaphoric distance (node 1), with the left branch ($d < 4$ clauses) leading to the majority of non-lexical subjects, and the right ($d \geq 4$ clauses) to most lexical subjects. Along both branches, the next split is by the form of the antecedent (nodes 2 and 11). The most fundamental interaction is hence one between the distance to and the form of the antecedent, which fairly clearly delineate the contexts in which non-lexical and lexical forms are selected.

For low-distance anaphors, non-lexical antecedents lead to the lion's share of both non-lexical outcomes and mentions overall (node 3, accounting for 61% of all observations). The misclassification error in this group is only 9%, meaning only 9% of cases are lexical; even so, the sheer size of this node means that in absolute terms, this is the node with the highest number (though not the highest proportion) of lexical subjects. Low-distance lexical antecedents again split by anaphoric distance (node 4), this time into antecedents in the previous clause $d < 2$ clauses, leading to a terminal node with 82% purity dominated by non-lexical expressions (node 5), and antecedents three clauses away (i.e. between $d \geq 2$ and $d < 4$) clauses, at which point corpus-based differences in the effect of humanness (nodes 6 and 8) come into play: In the English and Cypriot Greek corpora, anaphors $d = 3$ clauses away from a lexical antecedent are 83% non-lexical (node 7); in all other corpora, these anaphors are 60% non-lexical if human (node 9), but only 31% non-lexical if non-human (node 10). This suggests that outside of English and Cypriot Greek, a mention is more likely than not to be realized lexically at the intermediate distance of $d = 3$ if it is non-human and its antecedent was lexical, while in these two corpora, humanness distinctions make little if any difference in this context. For English, this is not surprising – we have already noted the indifference of the English corpus towards animacy distinctions in Section 6.1 above. Cypriot Greek, however, has been noted to be one of the corpora in the sample most sensitive to humanness, but as Figure 7.1 shows, the strengtth of this sensitivity is not universal, but dependent on context. Do

note however that the terminal nodes at the end of this branch are quite small, with only a few hundred cases split across ten corpora.

On the right side of the tree, we see that across all corpora, anaphors at distances above $d \geq 4$ clauses with lexical antecedents come out as the group most likely to be lexical (node 17, with 75% purity). There are no further fruitful splits in this branch, as constrained by the hyperparameters. The other branch, for non-lexical antecedents, splits again by distance (node 12), mirroring the structure of the left half of the tree, but here the threshold is a distance of $d = 7$ clause units: Subjects anaphors with non-lexical antecedents between $d \geq 4$ and $d < 7$ clauses away are 69% non-lexical (node 13). This reflects a gradual transition from non-lexical to lexical expressions at intermediate distances from the antecedent. Notable is the branch for long-distance anaphors ($d \geq 7$ clauses) with non-lexical antecedents – here, the frequency of related mentions offers the best split (node 14), a factor that has only negligible influence when examined across the entire sample, but becomes moderately relevant in this specific context (nodes 15 and 16). This nicely illustrates the power of tree-based models for highlighting complex interactions that would have otherwise gone unnoticed, tenuous though they may be.

Overall, we find that the majority of subject anaphors fall into a rather pure node resulting from the interaction between low anaphoric distance ($d < 4$ clauses) and non-lexical antecedents, but that conversely, there are no nodes of this purity containing mostly lexical expressions. Given that non-lexical subjects greatly outnumber lexical subjects in the data, this is not too surprising, but it does suggest that is not only generally easier to identify specific contexts in which reduced forms are selected, but also that reduced forms are a viable (though not always the preferred) option in essentially all contexts. From the inverse perspective, the majority of the overall much rarer lexical subjects are largely restricted to certain contexts – mostly on the right side of the tree – and there are environments, such as the above-mentioned, in which they are exceedingly rare.

It is furthermore quite notable that the entirety of the early, major splits are dominated by two variables, anaphoric distance and the form of the antecedent, on both sides of the tree. As mentioned above, this suggests that broadly speaking, referential choices are delineated into four configurations, with low-distance non-lexical antecedents and high-distance lexical antecedents respectively leading to the clearest cases of non-lexical and lexical subject anaphors, respectively, and the others to more mixed outcomes.

We will discuss these notions in greater detail (and with more robust evidence) in Section 7.2.3 below.

Other predictors in the model – humanness, frequency of related mentions, and corpus – only show up further down on the tree, and the tree stops growing before any of the remaining predictors could be chosen for splits. Notably, differences between corpora are selected only for one split in this tree, and there in the context of humanness; this suggests that cross-corpus variability exerts no fundamental influential on the selection of lexical expressions, and that the predictors selected higher up in the tree affect outcomes more or less equally in all corpora, or at least with a limited degree of variability.

As we have noted above in Section 6.8, the form and role of antecedents are strongly associated in low-distance contexts, most strongly in same-role anaphoric chains. It is likely for this reason that antecedent role is not chosen for any splits in the tree: After the early and significant splits by antecedent form, there is little more to be gained from further splitting by antecedent role (vis-à-vis other options). Given our observations in Sections 6.7 and 6.8, however, it is surprising that form is picked over role here rather than the other way around.

7.1.3.2 | Objects

Figure 7.2 is the decision tree for object anaphors. Here as with subjects above, the root node is anaphoric distance (node 1), but the classification threshold is slightly lower: Distances below $d < 3$ clauses lead to the majority of non-lexical objects, and distance above $d \geq 3$ clauses to the majority of lexical objects. On the left side of the tree, most splits are shared with those in the subject tree, beginning with a split by the form of the antecedent (node 2). Non-lexical antecedents at short distances lead to the biggest subgroup of anaphors (node 3), mostly composed of non-lexical objects, with a relatively high purity of 82%.

For lexical antecedents, the next interaction is again with highly specific values of anaphoric distance (node 4), where antecedents in the previous clause ($d < 2$ clauses), end in a terminal node containing 61% non-lexical forms (node 5), while among antecedents two clauses over (i.e. between $d \geq 2$ and $d < 3$ clauses) inter-corpus differences become distinctive, split along the lines of English versus the rest – a pattern we have repeatedly remarked on throughout Chapter 6. In most corpora, object anaphors three clauses from

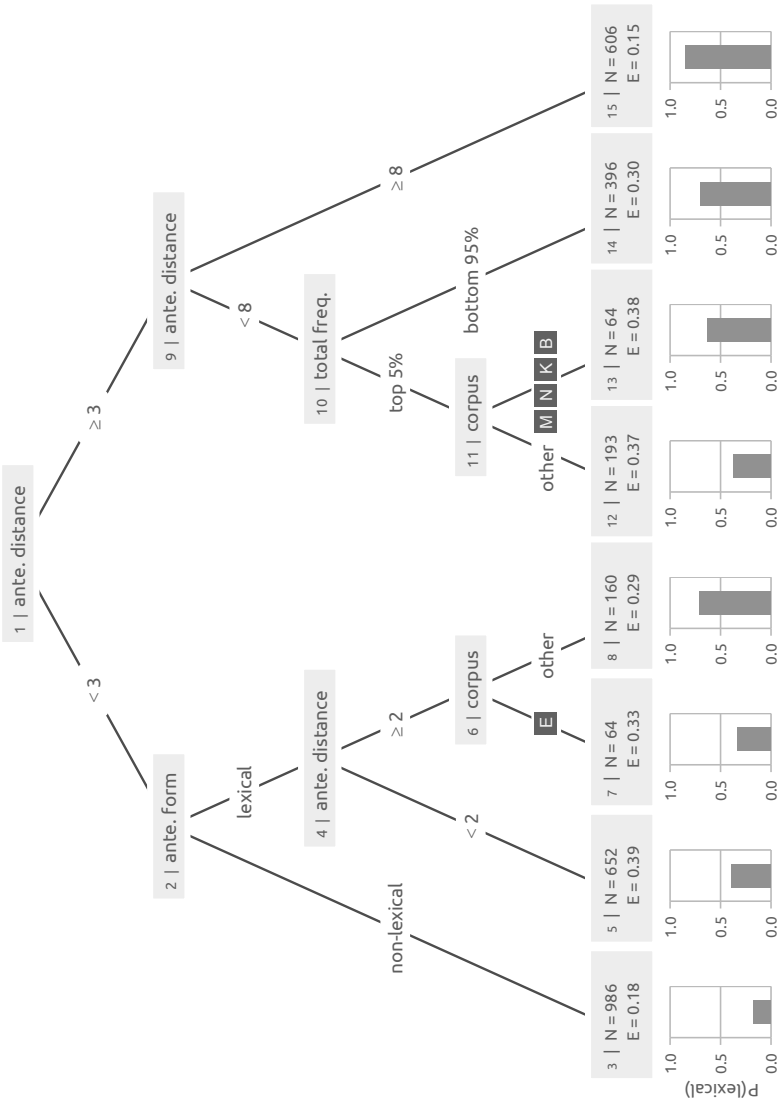


Figure 7.2 | An example form classification tree for objects.

Antecedent distances are given in clause units, other scalar factors in number of mentions. In the leaf nodes, 'N' is the size of the subclass and 'E' is the misclassification error. See the accompanying text for details on the model specifications.

a lexical antecedent are largely lexical themselves (node 8 with 71% purity); English instead tends towards more non-lexical realizations in this context (node 7 with 67% purity). This reflects the tendency (shared with subject anaphors) for the shift towards lexical expression to be delayed to higher distances in English compared to the other corpora in the sample (see Section 6.3), where the transition happens earlier and more gradually.

On the right-hand side of the tree, the second split following the root (node 9) is again with anaphoric distance, in essence resulting in two distance intervals. One of these involves the upper extreme of the distance scale ($d \geq 8$ clauses), which leads into a large terminal node where lexical expressions are decidedly dominant (node 15, with 85% purity). The other branch captures distances between $d \geq 3$ and $d < 8$ clause units distance from the antecedent, which then splits by total mention frequency (node 10): At these distances, object mentions of the 95% least common referents – which among objects are in the majority – are largely lexical (node 14, with 70% purity), while lexicality rates among the 5% most common referents are differentiated by corpus (node 11), with Mandarin, Nafsan, Northern Kurdish, and Tabasaran leading to a small and rather mixed terminal node, slightly favouring lexical expressions (node 13, 62% pure) and the rest of the corpora slightly favouring non-lexical expressions instead (node 12, 63% pure).

According to this tree – keeping in mind the shortcomings of single-tree models – as with subjects, the lexicality of object anaphors is primarily determined by distance to the antecedent. Mentions at long distances ($d \geq 8$ clauses) in particular are overwhelmingly, though not exclusively, lexical, and mentions at very short distances ($d < 3$ clauses) tending strongly towards non-lexical expression, especially of the antecedent is located in the previous clause and happens to be non-lexical as well. This is largely the same picture as seen above for subjects, albeit with shifted thresholds and an overall slightly less ambigious distribution of forms. Compared to subjects, the largest nodes (such as 3, 14, and especially 15) in particular show somewhat smaller misclassification errors, suggesting that in these common contexts object realization is reasonably straightforward to predict, which is possibly influenced by lexical objects not being quite as rare as lexical subjects.

The structure of this tree also reflects the somewhat greater degree of cross-corpus variation among object realizations seen already throughout Chapter 6. While differences between corpora only show up twice in the tree and fairly low down at that, in both cases they delineate a marked disparity in the com-

position of the resulting groups. Given the primacy of anaphoric distance, only three other predictors find room in the tree apart from corpus. Noticeably absent are humanness – presumably since most objects are non-human – and any measures of local discourse structure (frequency of recent co-referential, related, or competing mentions, etc.).

7.2 | Boosted decision trees

7.2.1 | Introduction

While the individual decision trees presented in the previous section are nicely illustrative and fairly intuitive to understand, they are also prone to overfitting unless rigorously validated (Kuhn & Johnson 2013: 182), and relatively volatile, especially the further one goes down along their branches (Strobl et al. 2009: 332). If the underlying data or model parameters are altered even slightly or if new data is added, single decision trees are liable to change in unpredictable ways. On their own, this makes them inherently less robust than many traditional statistical methods in terms of their predictive and explanatory power.

One means of getting around these limitations is the use of so-called ensemble methods (Elith et al. 2008; Strobl et al. 2009; Ridgeway 2020), which have enjoyed considerable popularity in recent decades, particularly in the field of natural language processing, but increasingly also in corpus linguistics (Baayen & Arppe 2011; Baayen et al. 2013; Walker et al. 2015); see Gries (2019) for a critical evaluation of CART and ensemble methods (specifically random forests) as used in corpus linguistics. Rather than fitting a single 'best' model, ensemble methods improve accuracy by learning as they go, either by fitting and merging multiple decision trees (e.g. via bagging, stacking, or model averaging) or by sequentially growing an improving model (via boosting). In this way, they generally offer predictions that are superior to those of single trees and many traditional modelling approaches (Elith et al. 2008), since, as Strobl et al. (2009: 336) note, "an ensemble of trees has the advantage that it gives each variable the chance to appear in different contexts with different covariates, and can thus better reflect its potentially complex effect on the response."

But while a chief advantage of single decision trees is their relative simplicity, ensemble methods tend to be highly complex beasts. This can make gaining an understanding of the path taken by the model to its final results quite challenging; in particular, there is no single "final" tree that can be calculated from them, as the output of an ensemble model follows from the aggregation of the input of numerous trees, and is as such not a tree itself. Ensemble methods instead require other means of interpretation and evaluation of their predictions.[2] That being said, despite their somewhat daunting complexity, ensemble methods do not necessarily have to be treated as black boxes. There are ways to summarize them that offer critical insights into the relationships between the model parameters, as will hopefully become apparent later in this chapter.

The ensemble method we will be employing here is called gradient boosting (Friedman et al. 2000; Friedman 2001, 2002; see also the summaries in Elith et al. 2008 and Ridgeway 2020). Gradient boosting machines (GBM) sequentially fit multiple decision trees to the data, building on previously fitted trees. They gradually increase predictive accuracy by emphasizing observations modelled poorly by earlier steps in the sequence.[3] In this way, GBMs attempt to minimize the variance in the final model, lowering the rate at which observations are misclassified. As each tree uses only a random subset of the data, the final model inherently has a degree of randomness (Friedman 2002), which improves predictive performance (Elith et al. 2008: 804).

2 Some studies use the results of single-tree models to summarize ensemble models; this practice is problematic (see Gries 2019: 15).

3 Another popular ensemble method is random forests (Ho 1995), which instead fit multiple trees in parallel, each with a randomly selected subset of predictors.

7.2.2 | Model design

The analysis presented in this section relies on the *gbm* package (Greenwell et al. 2020) for R, which is an implementation of Friedman's (2001) algorithm.

7.2.2.1 | Factor selection and balancing

Since GBM are based principles shared with individual decision trees, many of the same design decision apply to them. Like single trees, GBMs automatically disregard irrelevant predictors, and also account for interaction effects between predictors. We thus again include all of the thirteen factors listed above in Section 7.1.2.3 (and defined in Chapter 6) in the model, even those for which only a small effect on lexical choice has been observed. The factors are the same for both models, with the exception of transitivity, which is applicable only to the subject model. The response variable in all cases is as before the binary distinction between lexical and non-lexical forms.

To account for imbalances between the number of observations of lexical and non-lexical forms as well as between the ten corpora, the same case weights given in Section 7.1.2.2 above are applied here as well for the subject and object models. Higher case weights penalize misclassification by the model, making the influence of rarer cases (i.e. lexical forms, smaller corpora) less likely to be drowned out by the more common ones.

For the sake of simplicity, intra-corpus variation between speakers and texts is also ignored again in the following. As we have noted in Section 5.4, the magnitude of the differences between corpora, speakers, and individual texts as regards overall lexicality rates is not statistically significant, meaning that as far as lexical anaphors are concerned, the data are reasonably homogenous. Do note, however, that the examination of individual factors in Chapter 6 has suggested that object anaphors tend to be subject to greater degrees of inter-speaker (and hence inter-textual) variation, which is likely attributable to differences in the content and subject matter of the texts.

7.2.2.2 | Model hyperparameters

The following hyperparameters were used to generate the boosted models discussed in this section:

- number of trees $N_{trees} = 10000$
- learning rate $lr = 0.001$
- interaction depth $tc = 7$
- min. observations in nodes $N_{node} = 25$
- cross-validation folds $N_{cv} = 10$

Most of these parameters have already been explained in Section 7.1.2.3 above. The higher the number of trees in the model, the better its fit will be in most cases, though there is an optimal number of trees determined by the number of observations in the data as well as the other model parameters, beyond which the addition of more trees has diminishing returns vis-à-vis computation time. The number chosen here, $N_{trees} = 10000$, exceeds that threshold for both the subject and object data. The models use a Bernoulli error distribution function, which is tailored towards binary response variables.

Lastly, it should be noted that since GBMs have a stochastic component, the models will yield slightly different results every time they are generated, even if the data and hyperparameters are kept identical. This random element can be removed by priming the random number generator in R with a 'seed', ensuring consistent results.[4] The seed used in this study is '6'.

7.2.2.3 | Training and validation data

In addition to the cross-validation performed by the `gbm` function itself (see Section 7.2.2.2), we deliberately retain a fraction of the data for manual accuracy testing. The sample data have been separated into training (80%) and test sets (20%) for subject and object anaphors. As the name suggests, the former is used to calculate the actual model, the quality of whose predictions is then tested against the observations in the latter. The split between the two data sets is performed randomly (but fixable using the seed given above), but done

4 Random number generation in computers is in truth only pseudo-random; the exact sequence a supposedly random process will take can be fixed by having it start from the a particular state, as defined by the seed, every time.

in such a way that the proportion of the number of observations of each level of the response variable in each corpus is the same between training and test data, that is, it maintains the balance represented by the assigned case weights given in Section 7.1.2.2.

7.2.3 | Subjects

7.2.3.1 | Correlations and associations between predictors

Tree-based methods are overall less senstitive to collinearity than regression models due to their non-parametric nature, but strongly associated predictors can nevertheless inflate the importance of variables (Gries 2019: 2; 16). It is hence useful to first detour slightly by examining the correlations and associations between the predictors, which are shown for the subject data in Table 7.3.[5]

The predictors are a mix of continuous variables (e.g. anaphoric distance) and categorical variables (either with two levels, e.g. humanness, or more, e.g. antecedent role); while the calculation of correlation coefficients between two continuous variables is trivial, it is not possible to calculate correlations with or between non-continuous variables. As such, different measures are required to evaluate the associations between continuous and categorical variables and between two categorical variables (cf. Khamis 2008):

- continuous × continuous: Spearman's rank correlation coef. ρ
- continuous × binary: point-biserial correlation coef. r_{pb}
- continuous × nominal: correlation ratio η
- binary × binary: φ-coefficient
- binary × nominal: Cramér's V
- nominal × nominal: Cramér's V

Among the predictors examined in this study, humanness, total frequency (via protagonisthood), antecedent form, clause type, and transitivity are binary categorical variables; antecedent role and sequence of mention are categorical with more than two levels (labelled 'nominal' here); and the remainder of predictors are continuous. For associations with nominal variables with an

5 These use the full data set, rather than the training data used for the GBM model further below.

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)
(B)	0.32										
(C)	0.07	−0.12									
(D)	−0.18	0.30	−0.54								
(E)	−0.09	−0.06	0.07	−0.10							
(F)	0.11	−0.04	0.04	−0.10	−0.37						
(G)	0.21	0.08	0.09	0.13	0.07	0.10					
(H)	0.26	0.23	0.11	−0.37	−0.05	0.12	0.25				
(I)	0.20	0.32	0.05	0.38	0.00	0.11	0.13	0.27			
(J)	0.04	0.02	0.01	0.00	−0.08	0.05	0.07	0.05	0.05		
(K)	0.05	0.00	0.02	−0.03	0.05	0.04	0.02	0.01	0.02	−0.10	
(L)	0.16	0.06	0.05	−0.06	0.00	−0.05	0.12	0.08	0.04	0.03	−0.20

(A)	humanness	(E)	related mentions	(I)	sequence
(B)	total freq.	(F)	comp. mentions	(J)	clause type
(C)	ante. distance	(G)	ante. role	(K)	clause length
(D)	co-ref. mentions	(H)	ante. form	(L)	transitivity

Table 7.3 | Correlations and associations between each pair of predictors for mentions in subject position.
See the accompanying text for an explanation of which correlation or association measure is used for which combination of predictor types (continuous, binary, nominal).

internal order (i.e. sequence of mention), the sign of the coefficient is determined by the internal category ordering as listed in Section 7.2.2.1, from left to right. Comparison is only valid between the same type of measure, meaning different measures should generally not be juxtaposed directly. Instead, Table 7.3 should be used only for identifying which predictor pairs are most strongly correlated/associated. The φ-coefficient (for 2×2 tables), Cramér's V, and the correlation ratio η range from 0 to 1, all other measures are mapped a scale ranging from −1 to 1; in all cases, values at or around 0 indicate a lack of correlation or association.

A few explanatory notes on the less commonly used measures: The point-biserial correlation coefficient r_{pb} is equivalent to the Pearson correlation coef-

ficient r for a continuous and a binary categorical variable. The correlation ratio η is the weighted variance of the mean of each category divided by the variance of all samples. In essence, it indicates how well the value of a continuous variable can be identified with a specific value of a categorical variable. Lastly, Cramér's V is a measure of association between categorical variables; it is equivalent to the Pearson χ^2 statistic rescaled to values between 0 and 1, which for two binary variables is in turn equivalent to the φ-coefficient.

As Table 7.3 shows, there are no very strong correlations or associations between the predictors included in the model for subjects, meaning there is limited collinearity between most combinations of predictors. The most notable correlations and associations are as follows, some of which have already been touched upon in Chapter 6. The single strongest case of collinearity is between anaphoric distance and the frequency of recent co-referential mentions; as mentioned before, this is largely a consequence of how these factors are defined – given that recent discourse is defined as the previous six clause units, then anaphors more than $d > 6$ clauses away from their antecedent necessarily have no co-referential mentions in that stretch of discourse. Inversely, those that are closer than $d \leq 6$ clauses have at least one, and the shorter the distance to the most recent antecedent, the greater the opportunity for additional co-referential mentions in the six clause interval. For the subject data, this is the only correlation that approaches a substantial effect, and so needs to be kept in mind for the following analysis. Notably, even though anaphoric distance exerts the greatest influence on lexical choice, as evidenced by the single-tree model above as well as by the subsequent GBM model, its correlation with recent mention frequency is the only one that is in any way notable.

The same cannot be said for recent mention frequency, which has numerous associations of middling strength: with total mention frequency (i.e. protagonisthood), with sequence of mention, and with antecedent form (but not role). The first two come as no surprise, as all three factors are essentially measures of the same aspect of discourse, just applied to different scopes. In the case of sequence of mention, the first few anaphoric mentions of a newly introduced referent naturally cannot have a high recent mention frequency. The association with antecedent form is caused by the prevalence of reduced forms in anaphoric chains, of which high recent mention frequencies are strongly indicative of. Since the majority of anaphoric chains involve mentions in subject position, antecedents in other position likely reduce the strength of the corresponding association with antecedent role.

The association between humanness and total frequency is simply a consequence of most highly frequent referents being human, as already noted in Section 6.2.3. As regards total frequency and sequence of mention, given the Zipfian distribution of mention frequencies across referents (see Figure 6.3 in Section 6.2.1 above), low-frequency referents are overrepresented among second and third-in-sequence mentions compared to high-frequency referents, which make up a larger proportion of later mentions. And lastly, assuming a more or less even density of non-co-referential mentions in a given stretch of discourse, related and competing mentions essentially draw from the same "capacity", and hence come out as negatively correlated – the higher the rate of one, the lower the rate of the other.

In sum, there a few instances of collinearity between predictors that need to be given special attention in the following analyses, with the probable exception of anaphoric distance and recent mention frequency. While there are some structural associations among the remainder, none come out as strong enough to expect a substantial effect from.

7.2.3.2 | Relative importance of predictors

We begin our analysis of the gradient boosting model proper with an examination of the relative influence of the predictors used to train it. The GBM algorithm largely ignores non-informative predictors when fitting trees; as such, measures of relative influence work reasonably well to quantify the importance of predictors, since irrelevant ones have minimal effect of the outcome. Estimates of how much a predictor contributes to the overall model relative to the other predictors are calculated by comparing the number of times it is selected for splitting the data (Elith et al. 2008: 808; Friedman 2001; Friedman & Meulman 2003). This selection frequency is then weighted by the improvement of the model as a result of the split and averaged over all trees in the model. The result is an estimate of the relative importance of each predictor, scaled so that their sum adds up to 100. Higher values indicate greater importance, and hence stronger influence on the response. Conversely, predictors with lower values were selected more rarely for splits, and so end up having less of an effect on the predicted outcomes. It is important to keep in mind that these values are not an indication of absolute, but rather of relative importance: By themselves, relative importance says nothing about how large of an influence is actually statistically significant, only how much of a difference

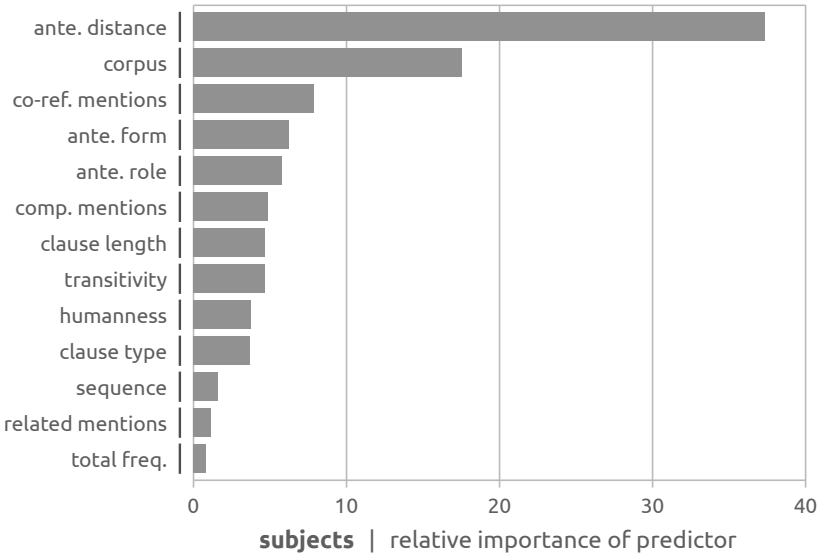


Figure 7.3 | Relative importance of the predictors in the model for subjects.

each predictor made during calculation of the model in relation to all other predictors.

Figure 7.3 shows the relative importance of the predictors in the model for lexical and non-lexical subject anaphors. Here, anaphoric distance has inarguably the by far greatest influence on the lexicality of subject anaphors, eclipsing all other predictors by a factor of at least two. This confirms our initial observations of the single tree model in Section 7.1.3.1, where distance was likewise the earliest split. The next most influential variable, at less than half the importance of distance, are differences between corpora. This does not necessarily mean that cross-corpus variation is drowning out the influence of the other predictors (other than anaphoric distance), but rather, as suggested by our analyses throughout Chapter 6, this is more likely than not the result of the accumulation of small, sporadic deviations, that is where one or two corpora (or texts therein) deviate from the general trend, in part because of their typological profile, in part because of differences in content. This effect

is then further amplified by the small inherent bias that GBMs have towards predictors with multiple levels: While a binary predictor can only ever offer a single split in a given tree, a predictor with multiple levels can be selected multiple times, as seen for instance with anaphoric distance in the single trees shown above in Section 7.1.3. This can in turn inflate its importance, which, as mentioned above, is based on the number of times it is selected for fruitful splits by the algorithm. That being said, a substantial contributor to the variability between corpora is the English data: Since a number of the tested factors make little difference for the form of subjects in English, the model can be expected to split the data along the line of English versus the rest fairly often. In fact, in a variant model (not shown here) that excludes the English data but is otherwise equivalent to the one presented in this chapter, the corpus variable still comes out as the second-most influential, but with only marginally higher relative importance than the third-ranked factor.

The remainder of the tested factors forms a more or less steady cline of decreasing relative importance for lexical choice, from the frequency of co-referential mentions at the top, down to clause type at the bottom, followed by three factors with extremely low influence – sequence of mention, frequency of related mentions, and total mention frequency. The frequency of co-referential mentions, as noted above, is somewhat correlated with anaphoric distance (see the previous Section 7.2.3.1), especially at low distances from the antecedent. Beyond that it confirms the direction indicated by distance: The presence or absence of co-referential mentions in recent discourse bears strong influence on the selection of lexical expressions. The form and role of antecedents play further into this, either by tightening or loosening local discourse coherence.

The frequency of competing mentions as well as clause length show little effect when examined individually, but here they outrank other predictors that do make appreciable difference individually, as seen earlier in Chapter 6, which come out as not particularly important in the multifactorial analysis. This is most notable for humanness, transitivity, and to a lesser degree also the role and form of the antecedent. Compared to anaphoric distance, these factors are only rarely selected for splits during model generation, despite their individually strong associations with lexicality, a pattern that only becomes recognizable when bringing everything together. Since the algorithm always selects the most promising splitting variable first, much of the variance explained by these variables is in essence already captured by anaphoric

distance, that is, once distance has been accounted for, these other factors have less to add to the model than the examination of their individual patterns would suggest. But as Gries (2019) argues, this can potentially obscure highly predictive interactions between variables, which is why it is essential to combine and augment our examination of relative importance with examinations of the individual and combined effects of each factor, which is done in the next two sections.

7.2.3.3 | Individual marginal effects

The marginal effect plots (also called partial dependence plots) in Figure 7.4 show the effect each predictor has on the response variable after accounting for the effect of all other predictors in the model, with scalar variables held at the mean and categorical variables at their most frequently occurring level. The closer the shown effect for a category is to 1, the more strongly predictive it is of a lexical subject anaphor, and accordingly for effects closer to 0 and non-lexical subjects; effects around 0.5 conversely make little difference either way. As Elith et al. (2008: 809) hasten to caution, however, "these graphs are not a perfect representation of the effects of each variable, particularly if there are strong interactions in the data or [if] predictors are strongly correlated": We will examine selected second order interactions in the next section, and correlations have been checked for in Section 7.2.3.1 above. Even with this caveat, however, marginal effects plots provide a useful basis for the interpretation of modeling outcomes (Friedman 2001; Friedman & Meulman 2003).

Figure 7.4 shows the marginal effects of each of the thirteen factors in the model for subjects. Here, we again find that the effect of anaphoric distance is the most pronounced and most clearly differentiated by far. Only antecedents in the previous clause (i.e. $d = 1$), that is those in tight anaphoric chains, push unmistakably towards non-lexical subject realization. Antecedents two clauses over ($d = 2$) are not predictive of either outcome, hovering around the 0.5 mark, and any antecedents at greater distances ($d \geq 3$) are strongly indicative of a lexical outcome, with the likelihood of such increasing steadily with distance. This in essence echoes our earlier observations in Section 6.3 above, which is hardly surprising given the evident primacy of anaphoric distance in the multifactorial models.

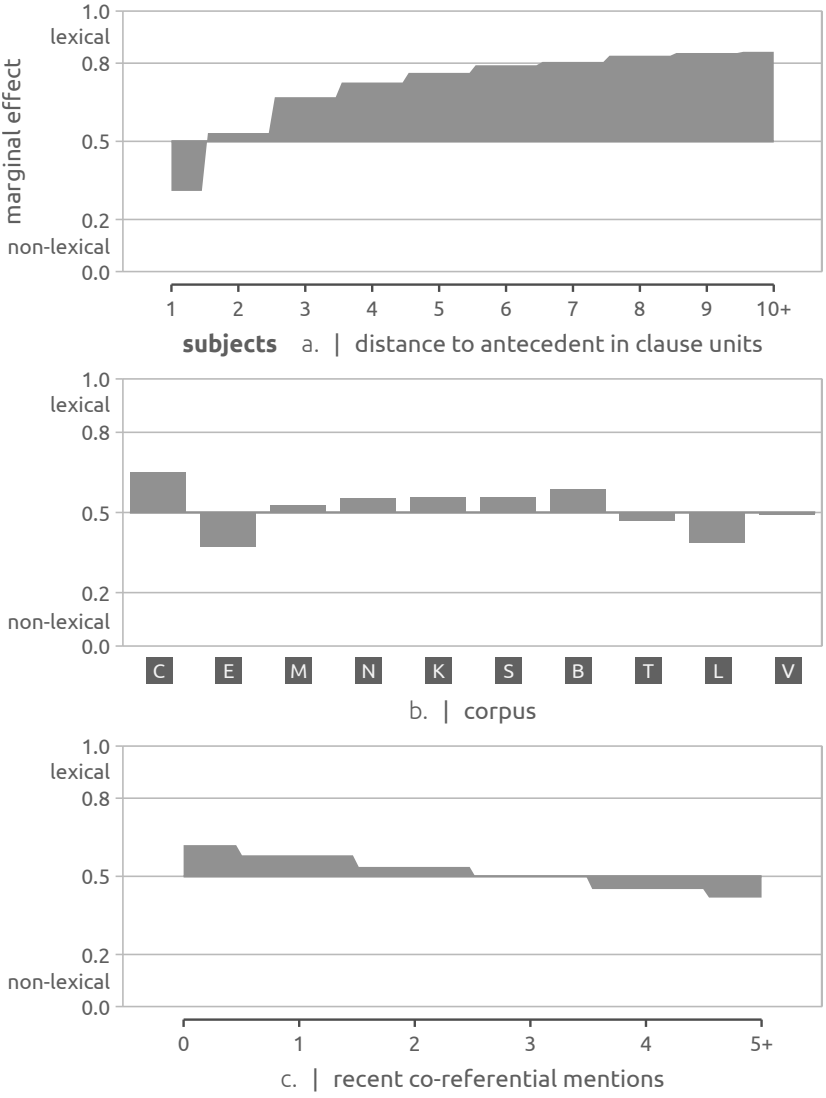
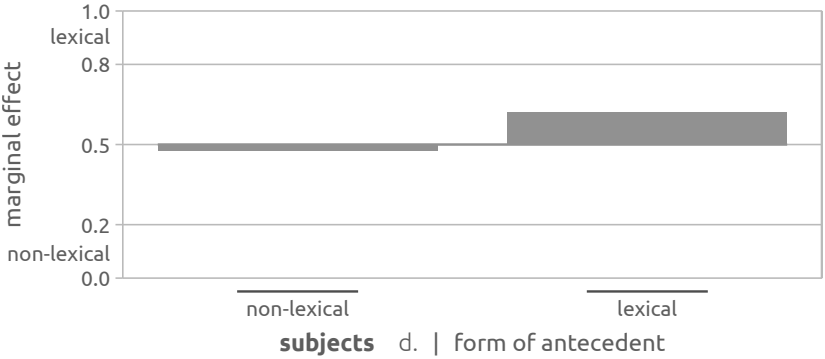
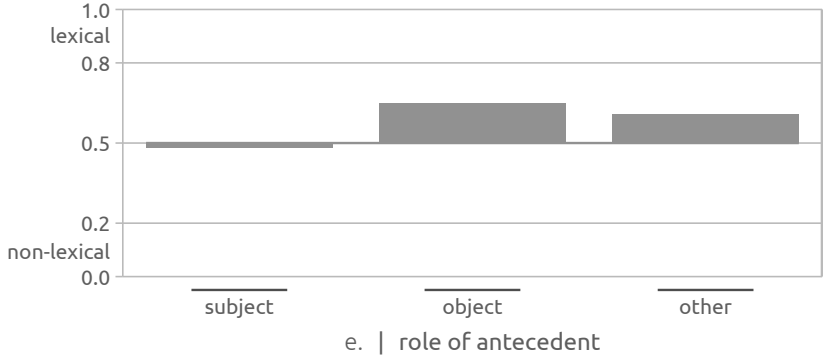


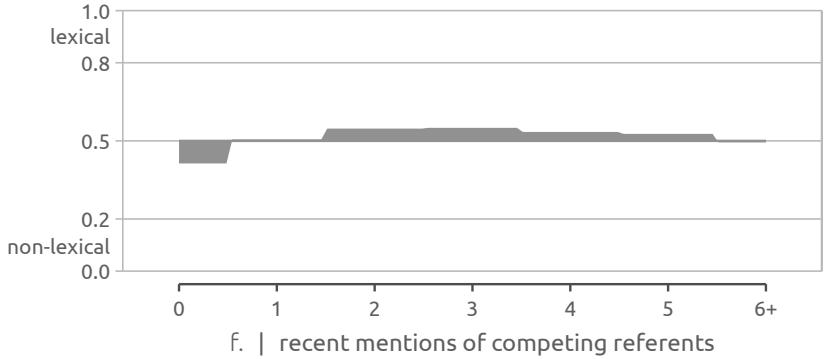
Figure 7.4 | Marginal effects of the predictors in the subject model.



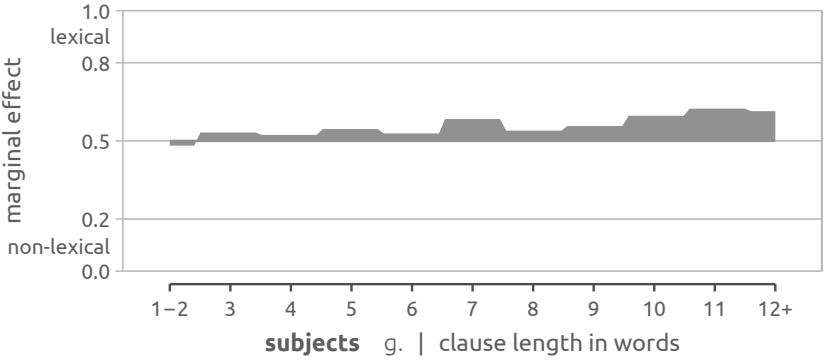
subjects d. | form of antecedent



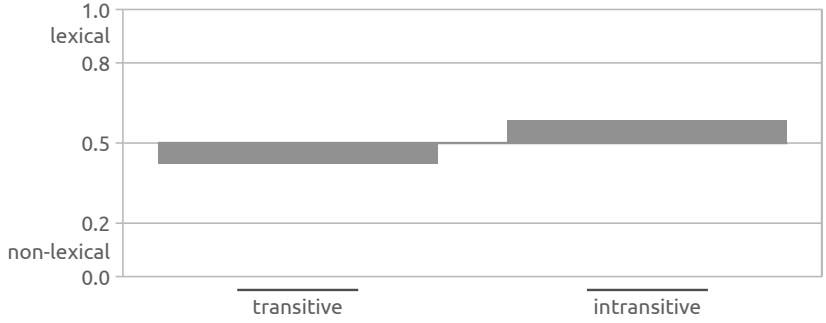
e. | role of antecedent



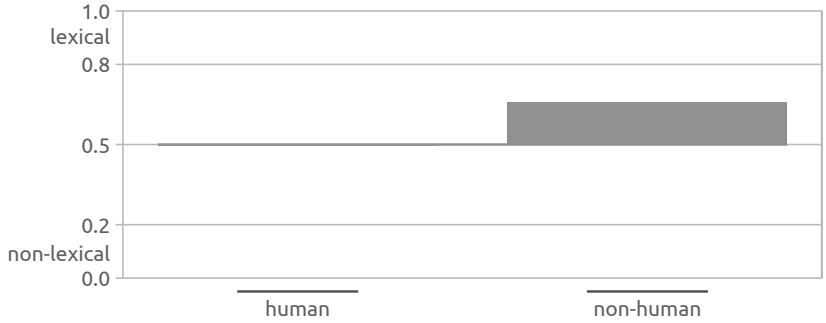
f. | recent mentions of competing referents



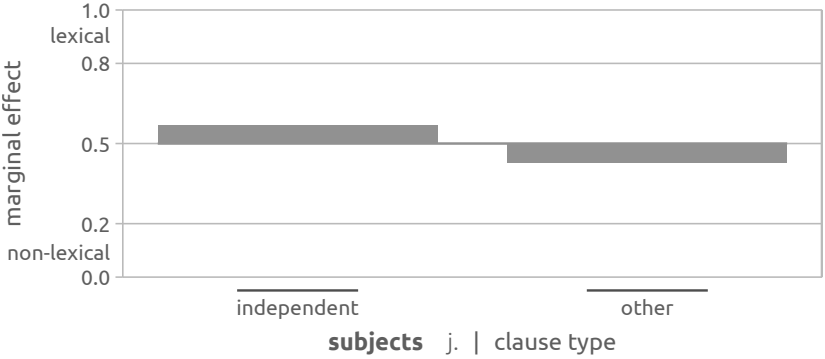
g. | clause length in words



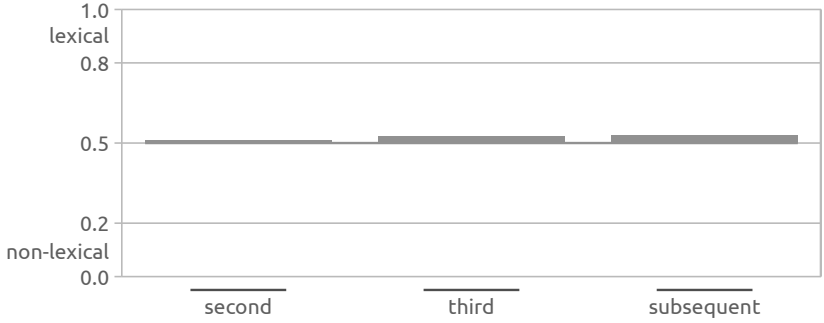
h. | transitivity



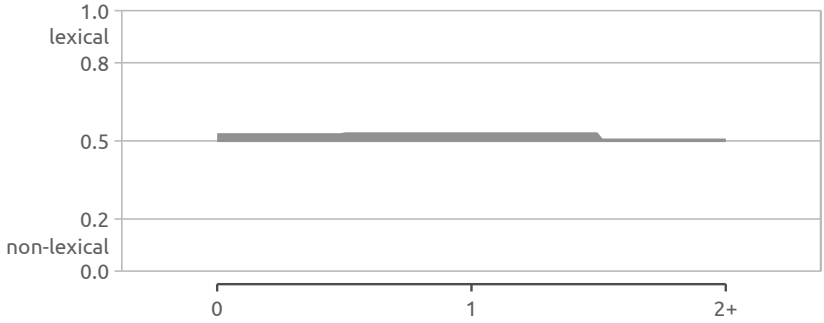
i. | humanness



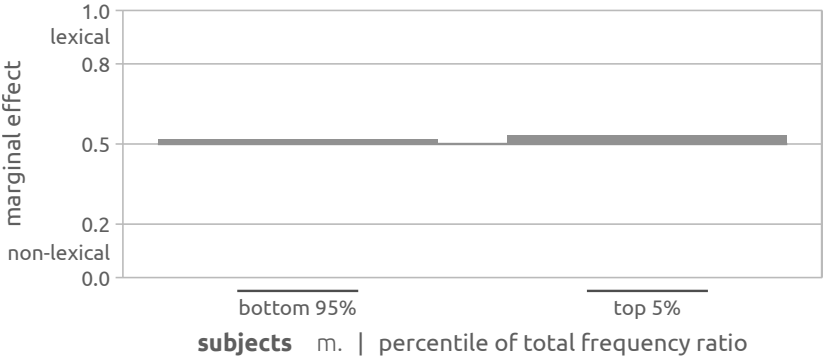
subjects j. | clause type



k. | sequence of mention



l. | recent mentions of related referents



As regards variability between the ten corpora, we find that the large relative influence seen above in Figure 7.3 is, as surmised there, in fact the result of relatively small effects spread out over the corpora. Cumulatively, these have some influence by appearing here and there in the trees, but never in a decisive fashion, unlike distance and certain other predictors. Another component of the variation between corpora are the small differences in baseline rate of lexical expression (originally seen in Figure 5.2 in Section 5.2.1), which we have remarked on numerous times already.

The marginal effects of none of the other predictors in the model are as extreme as those of anaphoric distance, but certain patterns nevertheless emerge. The frequency of co-referential mentions in recent discourse patterns as one would expect, with higher frequencies leading to higher probabilities of non-lexical expressions, and vice versa. The highest predicted lexicality rate results from the absence of co-referential mentions in the preceding $N = 6$ clauses, but as mentioned earlier, this is equivalent to distances of $d > 6$ clauses.

All else being equal, a lexical antecedent will likely lead to a lexical anaphor, but anaphors with non-lexical antecedents are largely ambivalent. A similar pattern shows for the role of the antecedent: Subject antecedents are not strongly predictive of either form, but object and other antecedents clearly result in higher rates of lexical expression. That is, there is a split between subject and non-subject roles. We have already touched on the notion of high local discourse coherence caused by a combination of low-distance anaphors and high recent mention frequencies, and anaphoric chains in which the antecedent

has a reduced form and shares the same role with the anaphor constitute close to the extreme case of that. But here we see that the model actually points to certain characteristics of low-coherence contexts (i.e. antecedent is lexical and not a subject) being a stronger indicator of lexical subject anaphora than high-coherence contexts (i.e. antecedent is non-lexical and a subject) are of non-lexical subject anaphora. In other words, though most non-lexical subjects do indeed occur in high-coherence contexts such as anaphoric chains, these contexts are not perfectly indicative of the selection of non-lexical forms, which instead may occur in essentially any context. Low-coherence contexts, however, appear to be more clearly indicative of lexical forms. Also relevant here are the interactions between the role and form of antecedent with anaphoric distance, which will be touched on further below, in the next section.

The role of humanness, which we remarked on as having comparatively little importance in Section 7.2.3.2, plays into these observations as well. First, recall that the majority of subject anaphors in the sample are human (cross-corpus mean $P = 0.85$; $\sigma = 0.109$), and that animacy is an inherent property of referents that applies irrespective of other contextually dependent factors. As Figure 7.4 shows, human referents are essentially undifferentiated as regards predictions of lexicality; it is the rarer non-human category that ends up being predictive of lexical outcomes. Even so, the ambivalence of the human category is much more pronounced than would be expected from the distribution seen in Section 6.1, which shows a clear distinction between human and non-human referents in terms of lexicality rates. This, together with the comparatively low relative importance of humanness, indicates that differences in humanness are concomitant with differences among other variables which the model deems more strongly predictive. Humanness could hence be argued to serve only as an amplifier of the aforementioned topical and highly coherent contexts defined by short anaphoric distances, high mention frequency, and same-role anaphoric chains. Each of these contexts is strongly associated with humanness, so that the individual effect of humanness is subsumed by them. Non-human referents fall out of this pattern of convergence, and hence come out as more strongly predictive in the model.

The remainder of the predictors in the models have only marginal effects on the model outcomes. Among these, transitivity and clause type show perhaps the most interesting associations: Though the actual effect is not particularly strong, the subjects of transitive clauses are predicted to be more likely to be non-lexical, and the subjects of intransitive clauses more likely to be

lexical. This is well in line with research on the interface between syntax and discourse structure (e.g. Du Bois 1987b; 2003b; Haig & Schnell 2016), as we have already remarked on earlier in Section 6.12, though the effect is perhaps not as strongly predictive of lexicality as might be expected, for instance, from the clearer distinction claimed to exist in Du Bois (1987b). Clause type likewise exhibits a clearly discernible if weak association between lexical subjects and independent clauses on the one hand, and non-lexical subjects and dependent and other clauses on the other. As noted earlier, these effects are not associated with one another, but instead additive; we should hence find slightly more non-lexical subjects in transitive subordinate clauses, slightly more lexical expressions in intransitive independent clauses, and intermediate rates in the remaining combinations. It is unclear whether these effects fall together with the coherence patterns described above, or whether they constitute a separate layer.

Lastly, the length of the clause shows a more clearly differentiated effect on lexicality than was identifiable in Section 6.11 above. As before, however, it is the opposite of what is expected based on the findings in Arnold et al. (2009) and Arnold (2010), where clause length has been found to be inversely proportional with lexicality rates. Here we instead find a weak but appreciably monotonous association between longer clauses and lexical subjects (to reiterate, the length of the subject NP itself is not counted as part of this measure). The other predictors in the model – frequency of competing and related mentions, sequence of mention, and total frequency – exert little discernible effect.

In sum, it is by and large the properties and relative position of the antecedent that matter the most, attenuated by humanness. Other tested factors, such as the transitivity of the predicate and the dependency of the clause, have a modulating effect, but as the multifactorial model indicates, the driving force behind the selection of lexical expressions are the those that relate to local discourse coherence. Next, we will further refine these observations by looking at selected second-order interactions between predictors.

rank	predictor A	predictor B	size	
1	ante. distance	corpus	1.7389	
2	humanness	corpus	1.1762	
3	ante. distance	ante. role	1.0556	◆
4	clause type	corpus	0.9996	
5	ante. form	corpus	0.8212	
6	comp. ment.	corpus	0.7681	
7	ante. role	transitivity	0.5178	◆
8	clause length	corpus	0.4960	
9	ante. distance	ante. form	0.4488	◆
10	humanness	ante. distance	0.4399	◆
11	co-ref. ment.	corpus	0.3516	
12	humanness	ante. role	0.3439	◆
13	ante. distance	transitivity	0.2904	◆
14	ante. role	corpus	0.2392	
15	humanness	clause type	0.2268	
16	humanness	ante. form	0.2088	
17	rel. ment.	corpus	0.1953	
18	comp. ment.	transitivity	0.1929	
19	comp. ment.	ante. form	0.1620	
20	co-ref. ment.	ante. form	0.1480	

Table 7.4 | Interactions between predictors in the gradient boosting model for subjects, sorted by size.
Interactions with marks beside them are examined in more detail in the following.

7.2.3.4 | Second-order interactions

Table 7.4 lists interactions between pairs of the predictors in the model for subjects, ranked by size; of the $13 \times (13 - 1) \times 0.5 = 78$ possible interactions, only the twenty largest are listed. As with the individual marginal effects above, the interaction effects are calculated based on having all other predictors held at their respective mean value (if scalar), or else their most common level. Note that a number of the interactions with the largest size are with individual corpora: In particular, the effects of distance and humanness on lexicality vary the most across corpora. All interactions with the corpus variable

share a similar pattern, in that it is one or two corpora, usually including English, that diverge from the general pattern, not uncommonly because of their slightly different baseline lexicality rates; see the discussion on the comparatively high relative importance of the corpus variable in Section 7.2.3.2. As these interaction patterns have in essence already been examined in Chapter 6 above (albeit not within a multifactorial framework), they will be disregarded in the following, and we will instead focus on interactions between other predictors.

The six most notable interactions – excluding those involving the corpus variable – are visualized in Figure 7.5. First of all, note that four of these interactions are with anaphoric distance. With distance being the by far most important predictor in the model, this suggests that a number major predictors in the model have a contextual effect on lexicality contingent on distance; as we will see in the following, the differentiations captured by these predictors become more indicative of certain referential choices, and hence indirectly relevant for referent identification, as per accessibility theory, at certain distances from the antecedent, but less so at others.

The third largest interaction overall is between anaphoric distance and the role of the antecedent. As already noted in Section 7.2.3.3 above, antecedents in object and other positions fall together here, essentially reducing this factor to a binary split by subjecthood. At high distances, all roles fall together to yield roughly the same marginal effect, being strongly indicative of lexical realizations. Where the roles diverge and the influence of antecedent properties becomes noticable is at the lower end of the distance scale: Compared to antecedents in non-subject roles, the model predicts subject anaphors with subject antecedents to be substantially less likely to be realized lexically if said antecedents are located in the previous clause (i.e. $d = 1$ clause). The same effect is also discernible for distances of $d = 2$ clauses, albeit weaker, and from there it attenuates as distance increases. Given our earlier observations, this pattern is not surprising: Chains of same-role mentions, especially in subject position, result in very high discourse cohesion, which appears to be a driving force in the selection of reduced forms.

The interactions between distance and antecedent form on the one hand and distance and humanness on the other pattern almost exactly the same. The former is also apparent in single tree model discussed above in Section 7.1.3.1, though the here much more pronounced interaction between distance and antecedent role is curiously absent there. Unlike that interaction, the different

antecedent forms as well as human and non-human referents diverge most notably at low-to-intermediate distances, specifically between $d \geq 2$ and $d \leq 4$ clauses from the antecedent. Presumably other effects take over at very short distances, or else non-lexical forms are simply almost universally preferred there, irrespective of the properties of the antecedent. At longer distances, the properties of antecedent become less and less important, and animacy-related differences in salience begin to level out. The model hence suggests that it is chiefly at intermediate distances that these factors are at their most distinctive for the identification of the intended antecedent. Notably, for these three interactions, the more coherent level – human, non-lexical, subject antecedent – shows almost exactly the same trajectory across distances.

The last interaction involving distance is with transitivity, where the greatest difference is also at intermediate distances, and gets narrower at either extreme end. It is not apparent how this association would come to pass or what it might entail. It is possible that it is simply the result of noise in the data; as noted numerous times before, the tentative nature of these findings – due to the limited size of the corpus data and variability in content – should be kept in mind for all analyses.

Moving beyond anaphoric distance, in the interaction between antecedent role and transitivity, it is antecedents in the object role that stand out, so that the small, but structured differences between transitive and intransitive clauses noted in the previous section is nullified if the antecedent is an object. This means that, according to the model, referents promoted from object to subject are realized lexically with roughly the same likelihood irrespective of the transitivity of the clause. As we have noted in Section 6.7 above, the move from object to subject happens only comparatively rarely, as role persistence from mention to mention appears to be the norm. Even though as a whole, the subjects of transitive clauses tend to be lexical at a much lower rate than those of intransitive clauses, the promotion from object incurs a substantial penalty in identifiability, resulting in an elevated rate of lexical expression that negates the usual transitivity distinctions.

And lastly, antecedent role and humanness do not interact strongly, but we nevertheless see here that for the selection of lexical expressions in subject role, humanness distinctions are at their most influential if the antecedent was itself in subject position. This aligns with the propensity for subjects in anaphoric chains to be humans, as noted above.

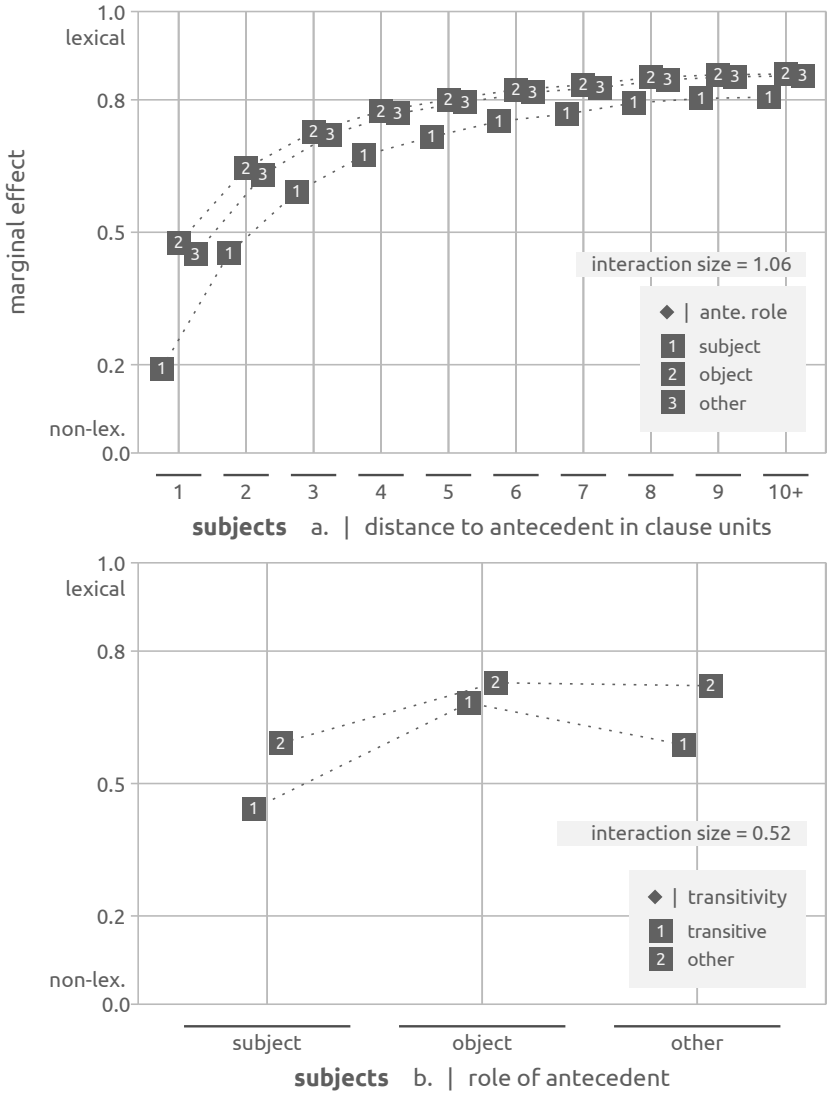
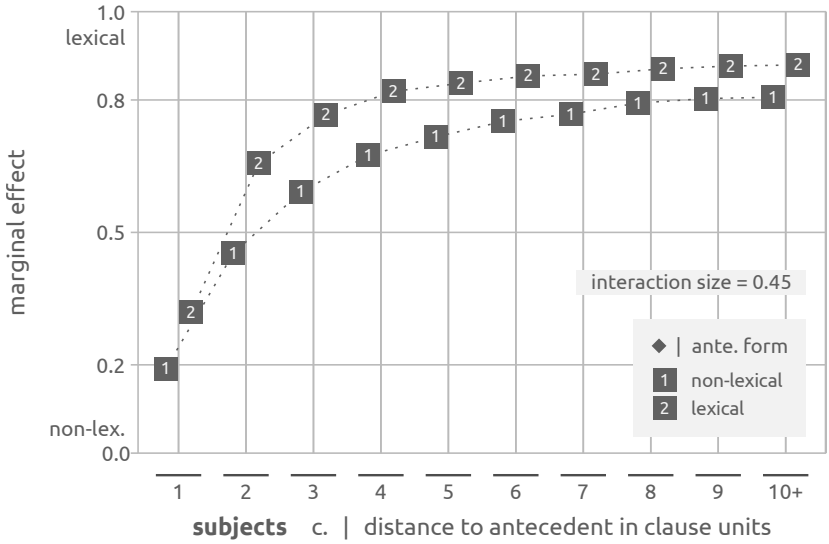
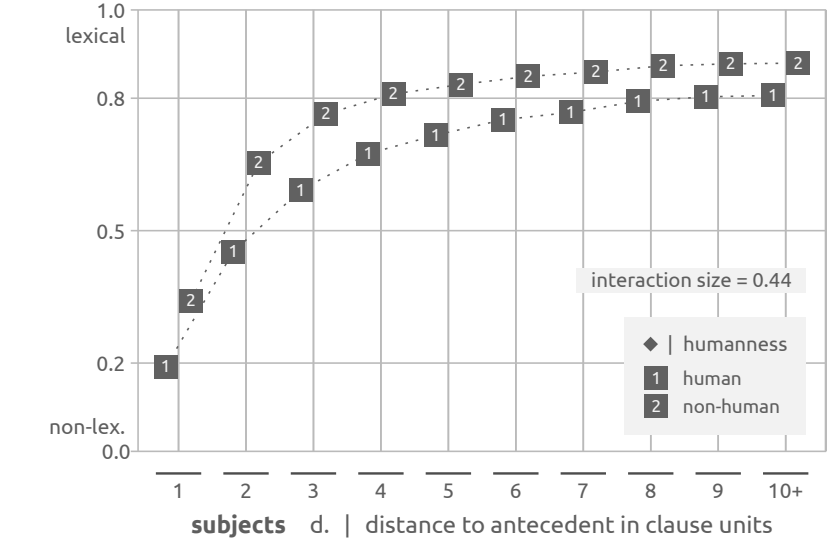
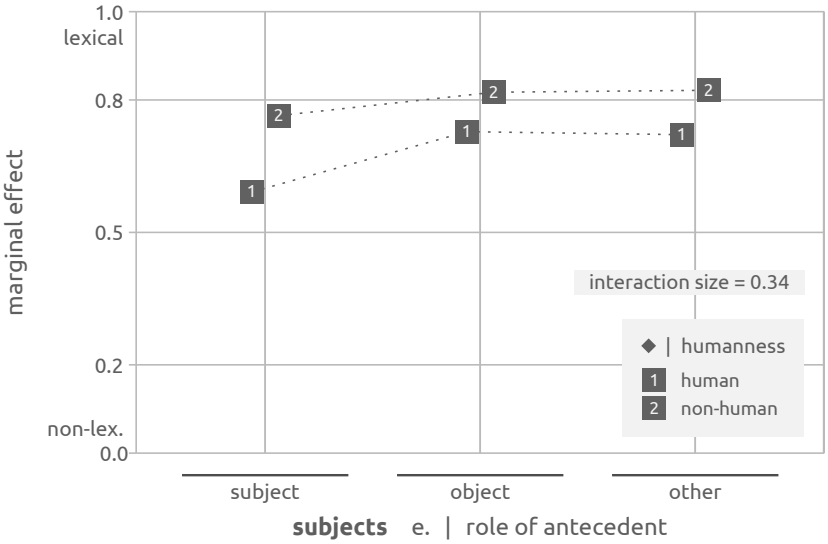
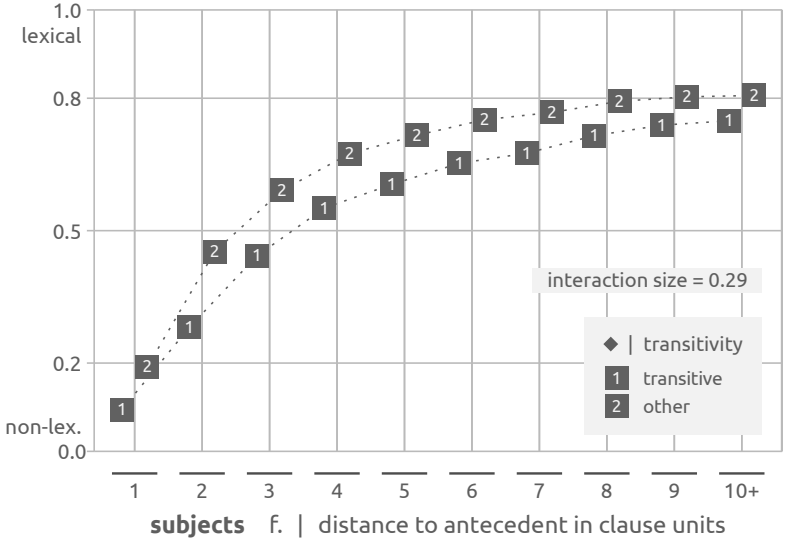


Figure 7.5 | Highest second-order interactions of predictors (excluding corpus) in the subject model.


marginal effect
1.0
lexical
0.8
0.5
0.2
non-lex.
0.0
interaction size = 0.45
◆ | ante. form
1 non-lexical
2 lexical
1 2 3 4 5 6 7 8 9 10+
subjects c. | distance to antecedent in clause units


1.0
lexical
0.8
0.5
0.2
non-lex.
0.0
interaction size = 0.44
◆ | humanness
1 human
2 non-human
1 2 3 4 5 6 7 8 9 10+
subjects d. | distance to antecedent in clause units


1.0
lexical
0.8
0.5
0.2
non-lex.
0.0
marginal effect
interaction size = 0.34
◆ | humanness
1 human
2 non-human
subject
object
other
subjects e. | role of antecedent


1.0
lexical
0.8
0.5
0.2
non-lex.
0.0
interaction size = 0.29
◆ | transitivity
1 transitive
2 other
1 2 3 4 5 6 7 8 9 10+
subjects f. | distance to antecedent in clause units

To summarize, none of the interactions in the model for subject come out as particularly strong, as even the largest discussed above do not exhibit substantial association effects. This is mostly due to anaphoric distance overpowering most other factors in the model, of which its involvement in many of the most notable interactions is also evidence. As such, the observations in previous section require little more qualification beyond what has been said in this section. In short, if the referent in question is human, the distance to the antecedent is low (especially in the previous clause), and said antecedent is non-lexical, then an anaphor in subject position can be expected to be almost categorically non-lexical. Lexical subjects are conversely more difficult to predict, as they are most common at high antecedent distances, where other factors have only an attentuating effect. All this being said, it should again be noted that the robustness of the models should not be overestimated, given the limitations in sample size, which is further divided into smaller subsamples in these interactions. As a final step in the exploration of the model for subject anaphors, the next section will evaluate its overall predictive power on the basis of the previously withheld test data.

7.2.3.5 | Evaluating predictive accuracy

While this is for the most part an exploratory study whose aim is to explain the mechanisms of referential choice, the methods used to do so also lend themselves to predicting choices on unseen data. For the evaluation of predictive performance, we use the test data that we had withheld from training the model, which accounts for 20% of the full sample; see Section 7.2.2.3 above. Generally, it is best to understand the output of exploratory models such as this one as relative probabilities falling between 0 to 1 (for non-lexical and lexical forms respectively), as seen in the figures above, rather than as a problem of (in this case binary) classification. However, the standard of comparison used by many approaches of this kind is predictive accuracy, which is the rate at which the models correctly anticipates outcomes:

$$\text{(82)} \quad \text{accuracy } acc = \frac{\text{true positives} + \text{true negatives}}{\text{total number of cases}}$$

Consider, for example, Kibrik et al. (2013) and Kibrik et al. (2016), which base their evaluations of various methods for predicting referential choices largely on comparisons of accuracy.

The calculation of predictive accuracy requires the transformation of the aforementioned probabilities into specific outcomes along a clear cutoff point. In order to evaluate whether an outcome predicted by the model should count as a lexical or non-lexical form, we first need to decide on a suitable classification threshold: In binary classification, the naïve approach would be set this threshold at a value of 0.5, but that is appropriate only for normally distributed data in which both outcomes are equally frequent (Menardi & Torelli 2012: 102; He & Garcia 2009; Weiss & Provost 2001; Weiss 2004). Neither condition applies to the corpus data used here, where among subject anaphors non-lexical realizations outnumber lexical ones by about a factor of four. The appropriate approach here is to shift the classification threshold so that the minority class – that is lexical expressions, which we happen to be most interested in, as per the title of this study – is easier to predict correctly (He & Ma 2013: 72).

We will be utilizing two approaches to selecting classification thresholds in this study, one for the subject model in this section and the other for objects in Section 7.2.4.5, each with their own interpretive focus. The one used for the object model seeks to maximize overall accuracy, but is suitable only for more or less balanced data, while the one used for the subject model prioritizes the correct classification of positive (i.e. lexical) outcomes in order to account for class imbalances.

Figure 7.6 shows the precision-recall (PR) curve for the subject data, which plots model recall against precision. Precision is the rate at which positive cases in the test data are correctly identified by the model as positive:

$$\text{(83)} \quad \text{precision } prc = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall is another name for sensitivity, the true positive rate:

$$\text{(84)} \quad \text{sensitivity } sns = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

In geometrical terms, the procedure for identifying the optimal classification threshold involves finding the point on the PR-graph that optimally balances recall and precision, this being the point closest to top-right corner of the graph, which indicates a model with perfect “skill”, that is one that always predicts outcomes correctly. The dashed line here indicates the minimum pos-

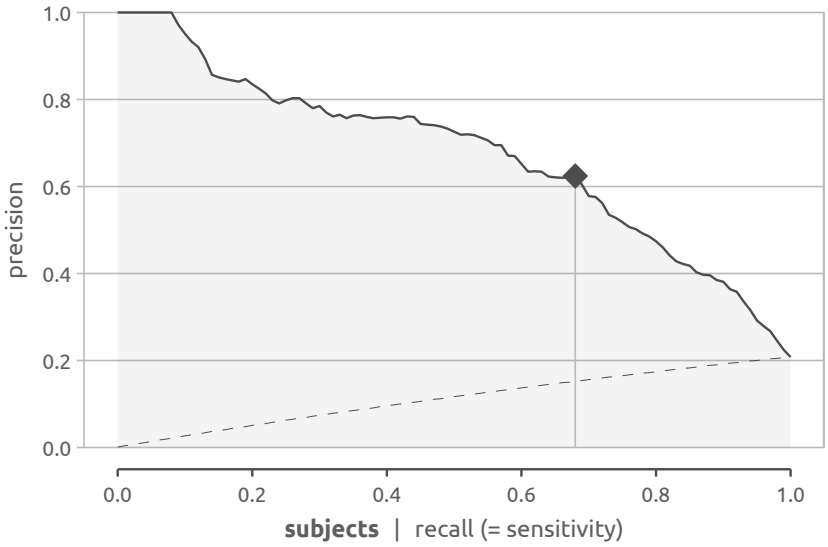
sible precision value, a no-skill model that always predicts the majority class. As the PR-graph indicates, the optimal classification threshold for the subject model is equal to a sensitivity of $sns_{prc} = 0.681$: Predicted values below this threshold will be classified as negative (i.e. non-lexical), values above as positive (i.e. lexical). For the sake of simplicity, we will use a single classification threshold for all corpora in the sample, but given the inter-corpus variation we have noted in earlier sections, it is quite likely that slightly better results could be obtained by calculating different thresholds for each corpus.

The bottom half of Figure 7.6 provides a confusion matrix for the model predictions on the test data based on the classification threshold, in addition to a number of associated performance statistics. Here, we find that in spite of our choosing a threshold that seeks to minimize misclassification of positive outcomes, over a third of lexical expressions remain incorrectly classified as non-lexical by the model (effective sensitivity $sns = 0.62$, $\sigma = 0.122$), while less than a tenth of non-lexical expressions are misclassified (specificity $spc = 0.91$, $\sigma = 0.029$).

$$\text{(85)} \quad \text{specificity } spc = \frac{\text{true negatives}}{\text{true negatives + false poistives}}$$

This aligns with what we have noted earlier: The selection of lexical expressions is conditioned by comparatively more restricted conditions such as in low-coherence contexts, whereas reduced expressions are a viable choice in essentially any context. As such, partly also due to their higher frequency, non-lexical outcomes are substantially easier to predict than lexical ones.

With mean accuracy across corpora of $acc = 0.84$ ($\sigma = 0.027$), the model for subjects outperforms the baseline rate of $p = 0.78$ ($\sigma = 0.045$) – that is, of randomly guessing outcomes – if only slightly. Of course, since non-lexical subjects are five times more likely than lexical subjects, much of this accuracy derives from being able to correctly guess the more common outcome. What is crucial is how much the accuracy exceeds the baseline rate, and as such actually manages to capture some of the mechanisms involved the selection of referring expressions, especially of lexical forms, the rarer outcome. The difference is lower than what is reported for comparable approaches in Kibrik et al. (2013: 3, tab. 1) and Kibrik et al. (2016: 8, tab. 5), but it is nevertheless a remarkable result considering that we are dealing here with spoken data from ten different languages, compared to the written English texts examined by



using PR-based threshold $t = 0.681$

		predicted	
		lexical	non-lex.
observed	lexical	237	136
	non-lex.	137	1283

	cross-corpus mean	σ
sensitivity (= recall)	0.62	0.122
specificity	0.91	0.029
precision	0.66	0.083
baseline rate	0.78	0.045
accuracy	0.84	0.027
F_1-score	0.63	0.057
MCC	0.54	0.066

Figure 7.6 | The precision-recall (PR) curve for the subject model, and the confusion matrix and selected performance statistics based on the optimal PR-based classification threshold.

Kibrik and colleagues. As such, when taking into account the inherent variability of cross-linguistic data that are uncontrolled for content, this level of accuracy is highly unexpected, and strongly indicative of the universal influence of discourse-contextual factors on reference form.

In this vein, it is worth noting that even though we are only testing for a small selection of factors from two linguistic domains – discourse structure and referent semantics – these already account for the aforementioned 84% of correctly classified cases. Any additional factors would be expected to add only marginally to the accuracy of the predictions, specifically those of lexical outcomes. But as Kibrik et al. (2016: 3) note, perfect prediction is likely to be impossible in practice, as there are contexts in which referential choices are inextricably ambivalent. This resonates with our own interpretation, as given above, of certain reduced expressions almost always being available for selection, irrespective of circumstances, unlike the more contextually restricted lexical expressions that alternate with them. We will be looking at selected cases where the model is off target in Section 7.2.5 below.

To close out, a few notes on the usefulness of raw accuracy as a measure of performance. The disadvantage of accuracy is that it considers positive (i.e. lexical) and negative (i.e. non-lexical) outcomes with equal weight; since, as mentioned above, the two outcomes are not balanced for subject anaphors, this can lead to potentially misleading results (Menardi & Torelli 2012). As such, for natural data such as those taken from speech corpora, with all their inherent variabilities, looking at base accuracies by themselves, as done here, has limited utility. It is better to combine accuracy-based evaluations with measures that take class distributions into account, such as the F_1-score or Matthews correlation coefficient (MCC), both of which are also provided in Figure 7.6. The F_1-score is the harmonic mean of the precision and recall, and the MCC is mathematically identical to Pearson's φ-coefficient (but usually not referred to such in the machine learning literature) and can serve as a measure of the quality of binary classifications. For the purpose of comparing the performance of the subject model with that of the corresponding model for objects in Section 7.2.4.5 below, it is hence best to rely on these measures instead of accuracy alone.

7.2.4 | Objects

7.2.4.1 | Correlations and associations between predictors

Table 7.5 summarizes the correlation and association between each pair of predictors in the object model. Refer to Section 7.2.3.1 above for an explanation of which measure is used for which combination of predictors.

Since the subject and object model make use of the same set of predictors, most of the associations that derive from the structural properties of the predictors are present among the stronger ones here as well, chief among them those between recent mention frequency, anaphoric distance, sequence of mention, and total mention frequency. Two associations that are markedly stronger in the object data are those between total mention frequency and humanness, and total mention frequency and antecedent form. The former is due to the majority of highly frequent referents being human (cross-corpus mean $P = 0.67$, $\sigma = 0.246$). The latter implies that antecedent forms – and by extension, the forms in anaphoric chains – are not equally distributed across more and less frequent referents. In fact, only 28% (cross-corpus mean, $\sigma = 9\%$) of object mentions of highly frequent referents have lexical antecedents, compared to 65% ($\sigma = 8\%$) for less frequent referents.

As before for subjects, the only notable instance of collinearity to look out for in the following is the case of recent mention frequency and anaphoric distance. The remainder are unlikely to exert a substantial effect on the model outcomes.

7.2.4.2 | Relative importance of predictors

Figure 7.7 shows the relative importance of the predictors in the model for anaphoric mentions in object position. Here we again see anaphoric distance top the list, but "only" with a relative importance over three times as great as all other predictors except corpus, which is a less extreme difference compared to the subject model above. Instead, a greater influence is exerted by inter-corpus differences, which account for a substantial proportion of the splits in model. As explained above, the influence of the corpus variable is by and large the result of the accumulation of small splits, where a handful of corpora diverge from the rest. Even so, these data suggest that differences between corpora are quite pronounced among object mentions. This is an observation we have

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
(B)	0.46									
(C)	0.06	−0.11								
(D)	−0.25	0.38	−0.64							
(E)	−0.13	0.08	−0.01	0.04						
(F)	0.11	−0.09	0.08	−0.20	−0.32					
(G)	0.20	0.14	0.06	0.12	0.11	0.07				
(H)	0.25	0.38	0.12	−0.40	−0.06	0.10	0.27			
(I)	0.20	0.33	0.06	0.41	0.05	0.17	0.11	0.29		
(J)	0.03	0.05	0.04	−0.05	−0.02	0.02	0.01	0.03	0.02	
(K)	0.04	0.03	0.05	−0.01	0.07	0.06	0.07	−0.01	0.01	−0.12

(A)	humanness	(E)	related mentions	(I)	sequence
(B)	total freq.	(F)	comp. mentions	(J)	clause type
(C)	ante. distance	(G)	ante. role	(K)	clause length
(D)	co-ref. mentions	(H)	ante. form		

Table 7.5 | Correlations and associations between each pair of predictors for mentions in object position.

See the accompanying text for an explanation of which correlation or association measure is used for which combination of predictor types (continuous, binary, nominal).

repeatedly made throughout Chapter 6, and before that in the initial discussion of the baseline lexicality rate of discourse in Chapter 5, where we noted that the overall proportion of lexical object realizations across corpora has noticeably greater range than the much more cross-linguistically stable subjects. Part of this variation is due to certain language-specific constraints on object form, as for example with human objects in Tulil (cf. Section 6.1.2.2) which only seldomly realized lexically. And lastly, the influence of both distance and the differences between corpora are likely to be slightly inflated due to the aforementioned bias of the modeling algorithm towards predictors with more levels, a shortcoming that shows in these relative importance measures especially.

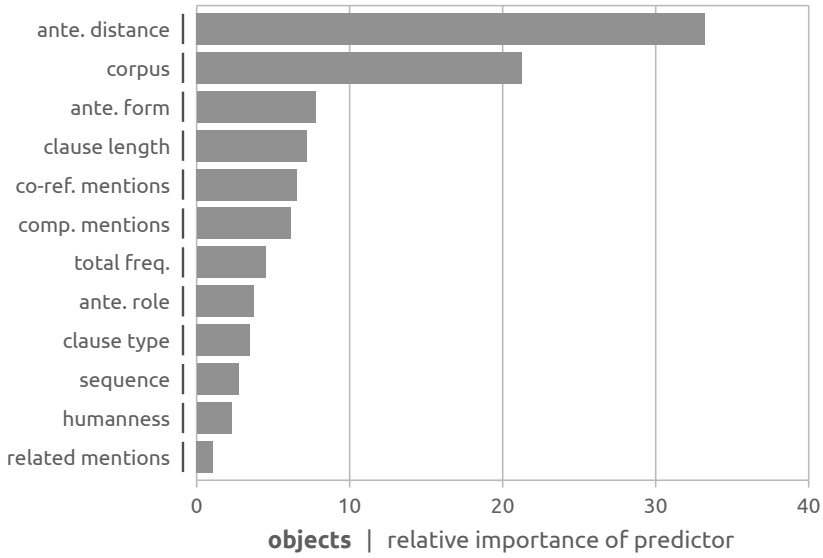


Figure 7.7 | Relative importance of the predictors in the boosted model for objects.

Following distance and the corpus variable in third and fifth place are the form of the antecedent and the frequency of recent co-referential mentions, which also played a notable role in the subject model. This suggests that both roles are affected similarly by the high-coherence contexts that these factors delineate together with anaphoric distance. The centrality of this pattern for referential choice has also shown in the single tree models above, and will again become evident in the interactions between these factors discussed below in Section 7.2.4.4.

Notably, clause length comes up as the fourth most important predictor in the object model. As seen in Section 6.11.2.2, longer clauses are more likely to host lexical objects, which themselves tend to be longer than non-lexical ones, in all corpora except English and Cypriot Greek, to bring up another example of cross-corpus differences among objects. As mentioned there, this runs counter to the expectation that an overall longer clause (and this is specifically a transitive clause, which mostly have short, non-lexical subjects)

would have an accordingly shorter object, so as to moderate the overall information content of clause (Arnold et al. 2009; Clark & Wasow 1998; Watson & Gibson 2004).[6]

Lastly, humanness is even less influential here than in the subject model above. Instead, we find that total mention frequency (i.e. protagonisthood), which, as noted above, is associated with humanness, bears comparatively greater influence. As such, the overall frequency of a referent in a text is more distinctive for the selection of lexical expressions than its humanness, since a large part of the variation in the former is already being captured by the former. In the subject model, the relative order of these two factors is reversed, with humanness more important than total frequency.

7.2.4.3 | Individual marginal effects

The marginal effects of each predictor in the object model are visualized in Figure 7.8. The most influential predictors show the pattern expected from the observations above: Higher anaphoric distances lead to higher rates of lexical expression, with only adjacent-clause references decidedly pushing towards non-lexical realizations, and distances of $d = 2$ clauses being largely ambivalent. There are no strong individual outliers among the ten corpora, only the accumulation of frequent minor splits that also inflate the relative importance of this predictor in Figure 7.7 above.

As with subjects, the most substantial combination of factors is of those capturing high (and by their absence, low) coherence contexts: Short anaphoric distance, high recent mention frequency, and non-lexical antecedents converge on low rates of lexicality; high distance, low frequency, and lexical antecedents on high rates. Notably, non-lexical antecedents are more strongly indicative of non-lexical object anaphors than is the case in the subject model, as there are more lexical forms in object position overall, and form persistence is less pronounced (see Section 6.8). The role of antecedent makes conversely very little difference, with only antecedents in non-core positions (e.g. obliques, adjuncts, NP modifiers, etc.) being slightly indicative of lexical objects. As such, promotions from subject to object position, though quite rare

6 Keep in mind that the measurements of clause length exclude the phrase in question, i.e. the object or subject NP.

(see Section 6.7.2), are expected to be realized lexically at essentially the same rate as object mentions in same-role chains, and only the shift from non-core roles incurs a penalty in recoverability that is reflected in the rate of lexical expression.

As touched on already above, only in very short clauses (1–2 words, excluding the length of object NP itself) do model predictions tend noticeably towards non-lexical objects; conversely, all clauses longer than $l \geq 3$ words push towards lexical forms, especially clauses that are very long ($l \geq 8$ words).

While the marginal effects of total mention frequency are not substantial, they nevertheless suggest that highly frequent referents are indicative of non-lexical objects, whereas less frequent referents tend towards lexical realizations. As highly frequent referents tend to be human, this might explain the lack of a corresponding effect of humanness on object form. Indeed, as mentioned above, almost all highly frequent objects are human (cross-corpus mean $P = 0.67$, $\sigma = 0.246$), leaving little for humanness to explain that total frequency did not already. These data hence imply that in object position, referents might be more readily identifiable on the basis of frequency than humanness.

In sum, while there are a number of notable differences, the most important predictors in the object model pattern similarly to those in the subject model. Specifically, objects are also strongly affected by local discourse coherence, of which anaphoric distance, recent mention frequency, and form of the antecedent are the most indicative. The role of the antecedent plays less of a role in determining object forms, however.

7.2.4.4 | Second order interactions

Table 7.6 lists the strongest interactions between the predictors in the object model, of which the four most notable are visualized in Figure 7.9. Note that as before, interactions with the corpus variable are not shown for presentational reasons; here, the same applies to a number of notable interactions with clause length, which are however touched upon below. It is also important to keep in mind that the sample of object anaphors is not particularly large, and that especially in interactions of predictors with many levels, subsample sizes can get quite small, which can impinge on the robustness of these observations.

As with subjects above, many of the strongest interactions are with anaphoric distance, including the second strongest interaction between distance and antecedent form. At all distances, non-lexical antecedents lead to non-lexical object anaphors and and vice versa for lexical antecedents and lexical anaphors, but the difference narrows noticeably at longer distances ($d \geq 8$ clauses), and slightly for antecedents in the previous clause ($d = 1$ clause). As such, the form of the antecedent – which, as noted above, is one of the more influential factors in the model – is most distinctive for determining object forms at intermediate distances. At high distances from the antecedent, its influence wanes as lexical expressions become the preferred form.

A similar pattern is found in the interaction between distance and humanness. As noted above, humanness comes out as largely irrelevant in the object model, likely to its association with total mention frequency (i.e. protagonisthood), but largest effect can be noted at intermediate distances, while at long distances ($d \geq 8$ clauses), the model expects no difference between human and non-human referents in terms of lexicality. The same is also true, if to a lesser extent, at very short distances ($d = 1$ clause). This pattern is also found in the subject model, though it is less pronounced there.

The last notable interaction involving anaphoric distance shown in Figure 7.9 is with clause type. Here, independent and other clauses only have an effect on lexicality at very low distances ($d \leq 2$ clauses), in particular in anaphoric chains (i.e. at $d = 1$). Most likely, this association points towards objects in subordinated clauses tending to be realized non-lexically if already mentioned in the corresponding matrix clause, though its strength relative to other interactions is somewhat surprising.

The interaction between recent mention frequency (i.e. in the previous $d = 6$ clauses) and total mention frequency (i.e. across the entire text) indicates that the elevated recoverability of more common referents is not limited to low coherence contexts ($N \leq 2$ recent mentions), but also affects the form of anaphors with a notable presence in recent discourse ($N > 2$ mentions).

There are two notable interactions with clause length that are not shown in Figure 7.9, one with anaphoric distance and one with total mention frequency. The first reveals that, like the aforementioned interactions with antecedent form and humanness, clause length is most distinctive at intermediate distances, with a widening of the difference between very short ($l \leq 3$ words) and very long clauses ($l \geq 9$ words) at these distances, and a narrowing at the lower and higher end of the spectrum. In the second, there is a greater dispar-

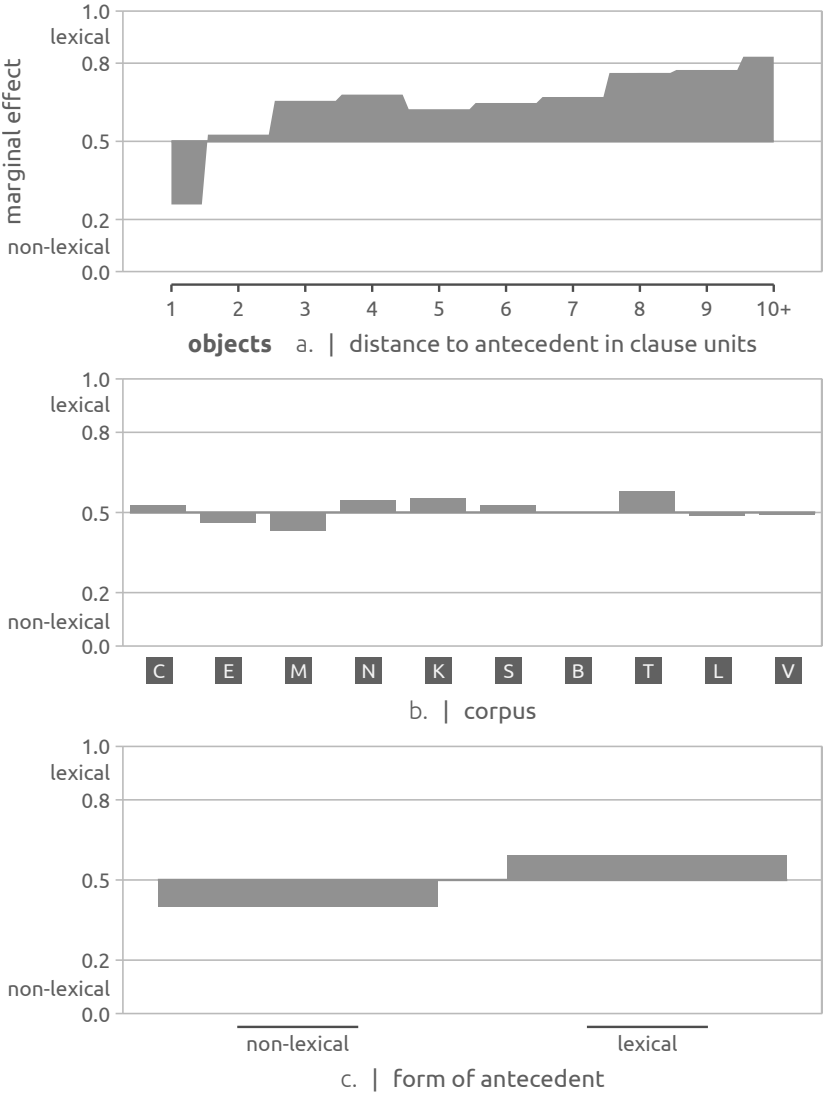
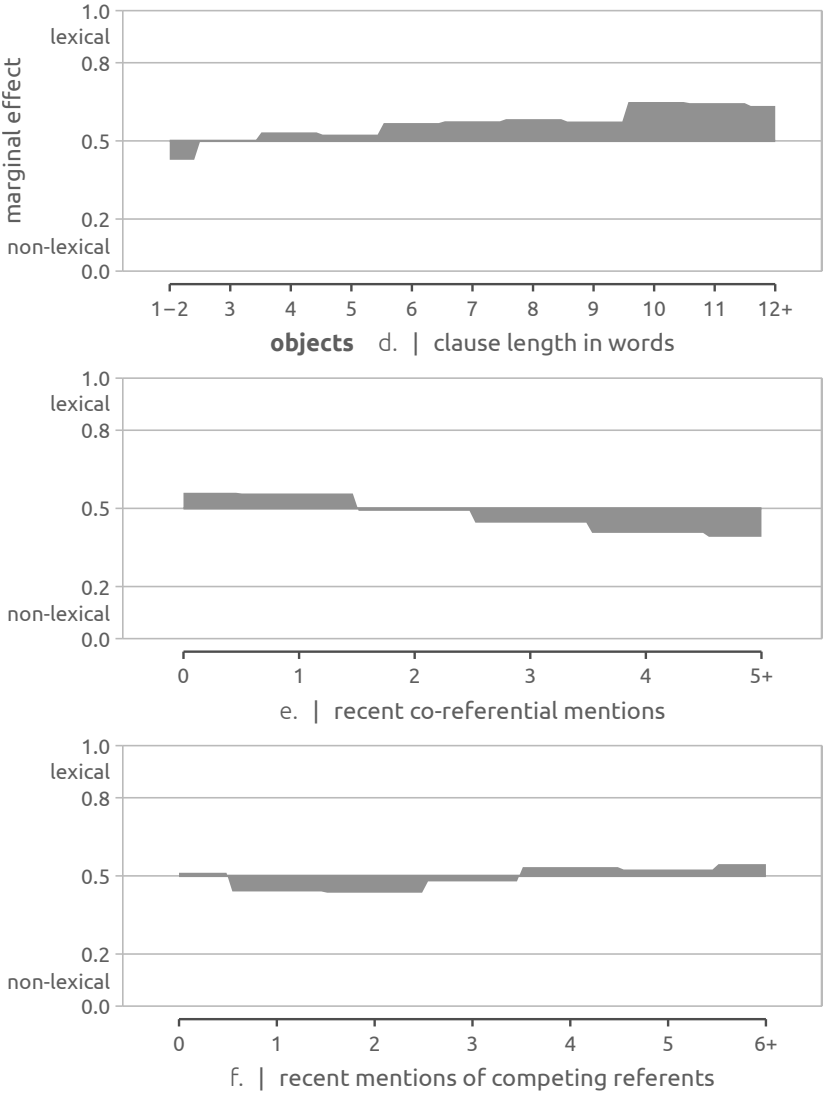
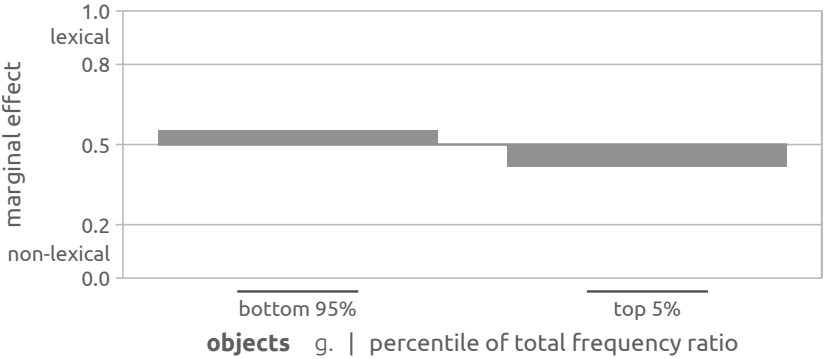
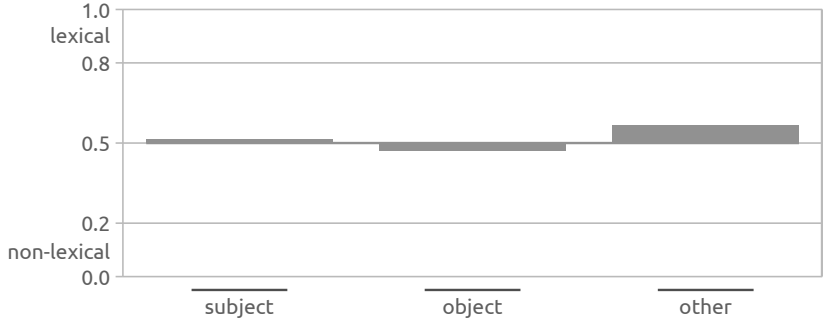


Figure 7.8 | Marginal effects of the predictors in the object model.

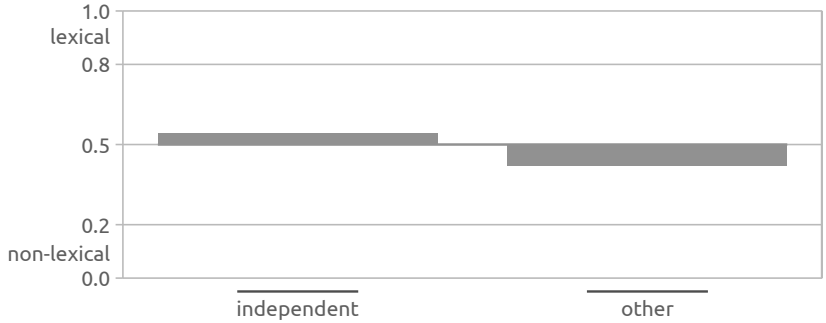

marginal effect
1.0
lexical
0.8
0.5
0.2
non-lexical
0.0
1–2 3 4 5 6 7 8 9 10 11 12+
objects
d. | clause length in words
0 1 2 3 4 5+
e. | recent co-referential mentions
0 1 2 3 4 5 6+
f. | recent mentions of competing referents



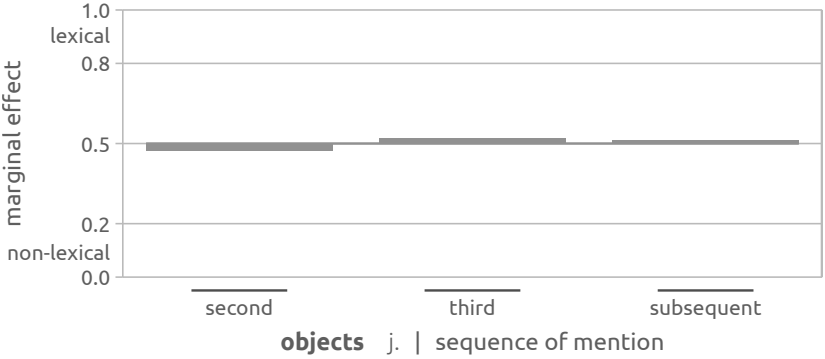
objects g. | percentile of total frequency ratio



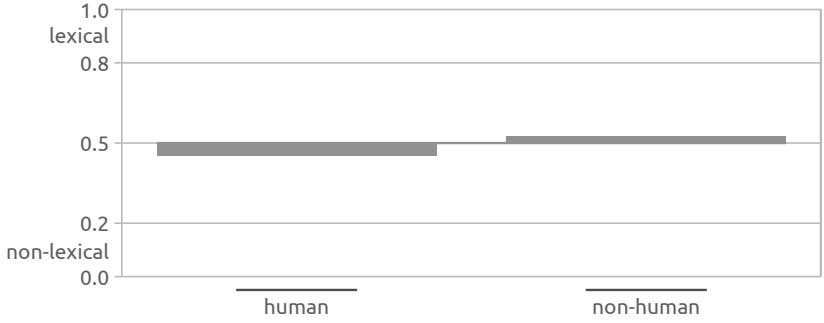
h. | role of antecedent



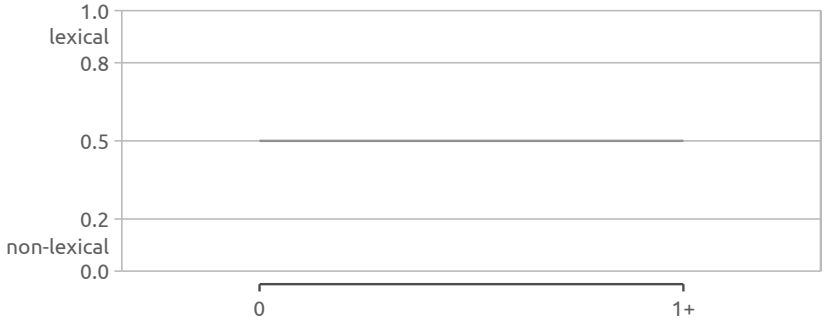
i. | clause type



objects j. | sequence of mention



k. | humanness



l. | recent mentions of related referents

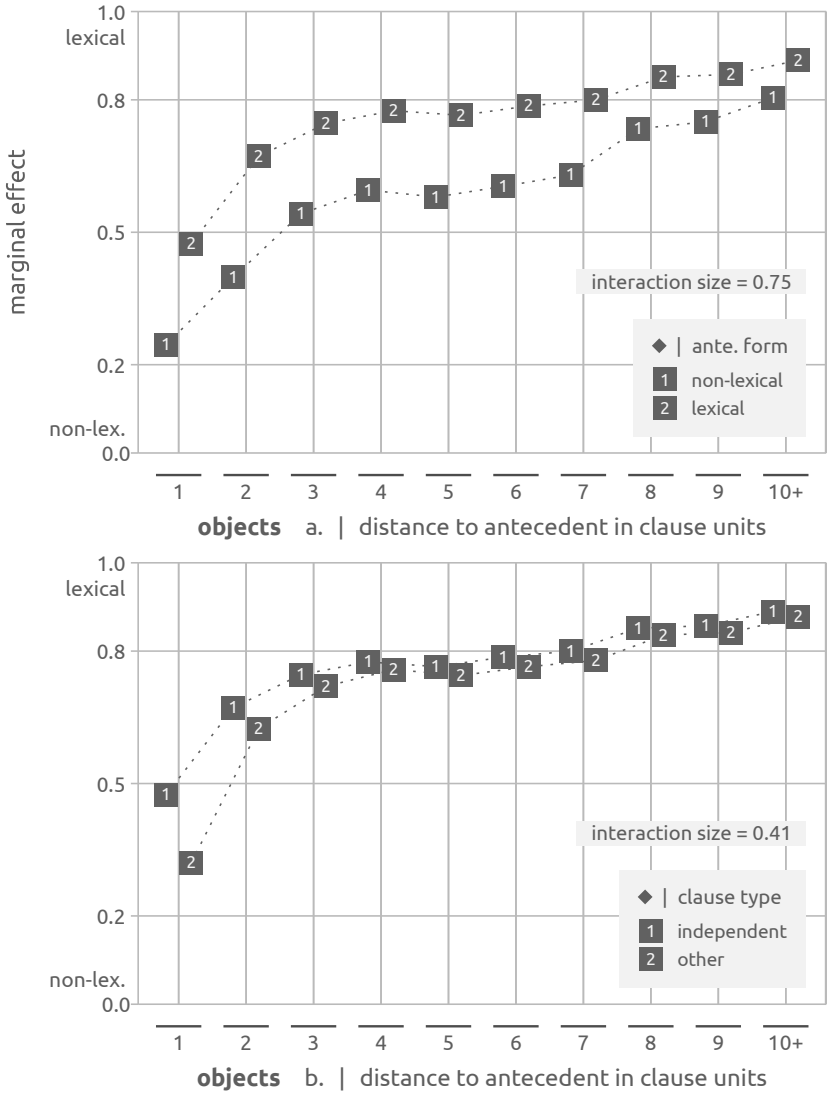
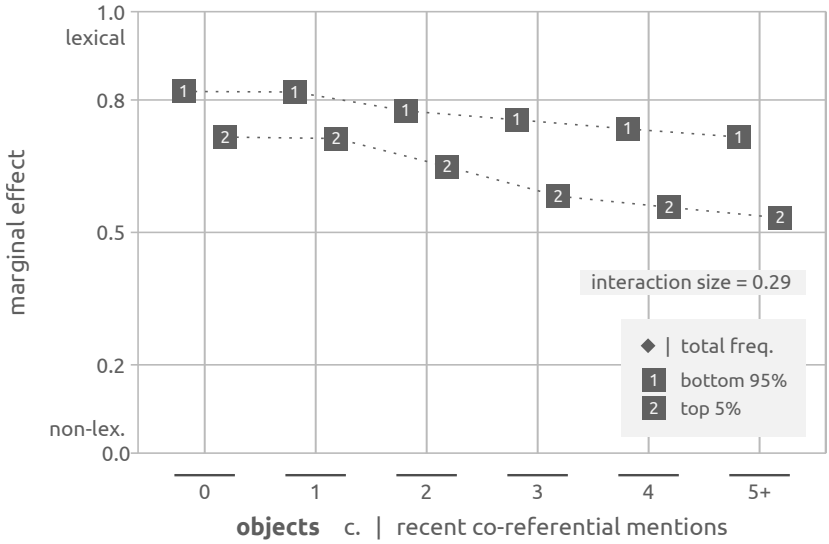
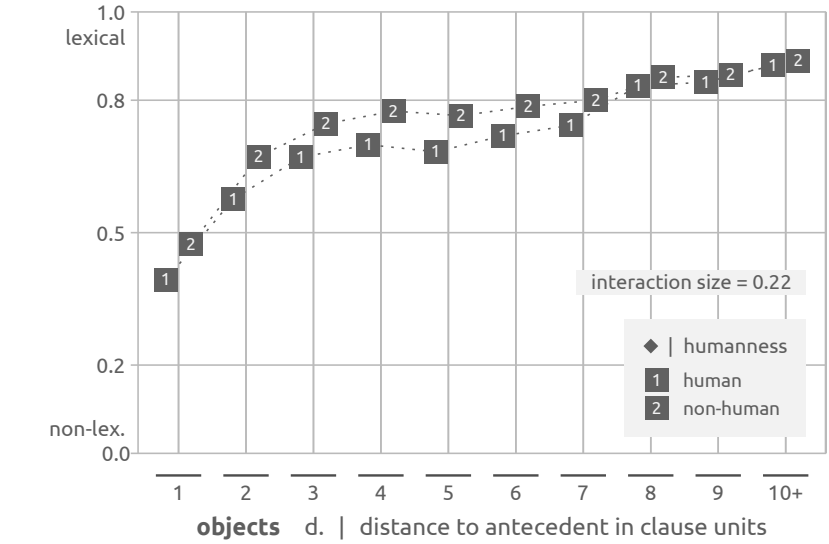


Figure 7.9 | Highest second-order interactions of predictors (excluding corpus) in the object model.


1.0
lexical
0.8
0.5
0.2
non-lex.
0.0
marginal effect
interaction size = 0.29
◆ | total freq.
1 bottom 95%
2 top 5%
0
1
2
3
4
5+
objects c. | recent co-referential mentions


1.0
lexical
0.8
0.5
0.2
non-lex.
0.0
interaction size = 0.22
◆ | humanness
1 human
2 non-human
1
2
3
4
5
6
7
8
9
10+
objects d. | distance to antecedent in clause units

rank	predictor A	predictor B	size	
1	ante. distance	corpus	0.9269	
2	ante. distance	ante. form	0.7459	◆
3	ante. form	corpus	0.7185	
4	ante. distance	clause type	0.4136	◆
5	comp. ment.	corpus	0.3394	
6	clause length	corpus	0.3246	
7	co-ref. ment.	total freq.	0.2898	◆
8	ante. role	corpus	0.2864	
9	ante. distance	clause length	0.2472	
10	humanness	ante. distance	0.2178	◆
11	comp. ment.	total freq.	0.2022	
12	co-ref. ment.	corpus	0.1746	
13	total freq.	clause length	0.1620	
14	rel. ment.	comp. ment.	0.1603	
15	co-ref. ment.	clause length	0.1567	
16	sequence	corpus	0.1488	
17	co-ref. ment.	ante. form	0.1442	
18	total freq.	corpus	0.1409	
19	ante. distance	co-ref. ment.	0.1391	
20	ante. distance	total freq.	0.1316	

Table 7.6 | The strongest interactions between predictors in the boosted model for objects.
Interactions with marks beside them are examined in more detail in the following.

ity between less and more frequent referents in short clauses, which gradually decreases as the length of the clause increases.

In sum, as with subjects, the effects of many factors is stratified across various anaphoric distances, exerting their greatest influence at intermediate, low, or high distances from the antecedent. Many associations are also shared between roles. Of course, with the sample size for objects being still smaller than that for subjects, some of these associations are tenuous at best, but even so, they point towards similar mechanisms underpinning referential choices across roles.

7.2.4.5 | Evaluating predictive accuracy

Unlike the subject data, the object data is almost balanced by lexicality, if not perfectly, with 47% of observations being lexical. As such, for classification, we are expected to get better results from selecting a classification threshold on the basis of an approach that maximizes accuracy, rather than one that aims to minimize misclassification of positive (i.e. lexical) outcomes specifically, as used above in Section 7.2.3.5 for subjects.

The approach used here, then, relies on the receiver operating characteristic (ROC), which is shown at the top of Figure 7.10. Unlike the similar PR-curve above, the ROC curve is a diagnostic tool that places sensitivity (i.e. the rate at which lexical forms are correctly predicted as lexical) against specificity (i.e. the rate at which non-lexical forms are incorrectly predicted as lexical).[7] Here, the dashed diagonal line indicates the curve of a so-called no-skill classifier, which always predicts the majority class. The optimal classification threshold is again the point on the ROC curve that is closest to the top-right corner of the graph, which indicates a model with perfect skill, that is one that always predicts outcomes correctly. For the object model, this point and hence the optimal classification threshold are located at a sensitivity of $sns_{spc} = 0.711$; to maximize accuracy, then, predicted values above this threshold should be classified as lexical, values below as non-lexical.

Using this accuracy-optimized threshold on the test data for objects yields the confusion matrix and associated statistics at the bottom of Figure 7.10. While the object model performs manages to outperform random guesses by a comfortable margin in terms of raw accuracy (cross-corpus mean $acc = 0.69$, $\sigma = 0.081$ vs. the baseline rate of $p = 0.51$, $\sigma = 0.102$), in absolute terms the predictive performance of the model is far from being perfect. This is perhaps indicative less of the difficulty of predicting the forms of object anaphors, and more a consequence of the inter-corpus variation in the data: Given the fairly large standard deviation ($\sigma = 0.081$) of the cross-corpus mean accuracy, the model appears to perform better in some corpora than in others; this is noticeable especially in comparison with the subject model above, which showed a greater degree of homogeneity across corpora. This difference between the

7 See Section 7.2.3.5 for a definition of these terms.

two roles (and data sets) has been remarked on a number of times before, and will return to it again for the final discussion in Chapter 9.

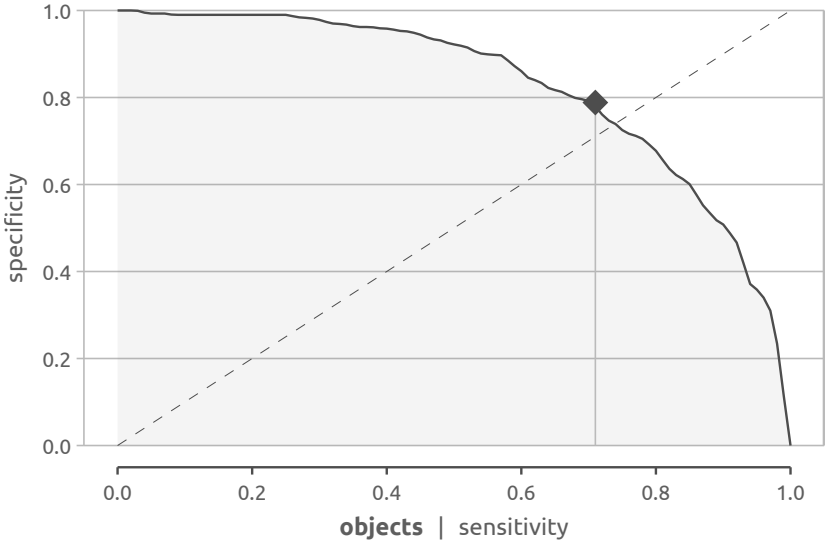
7.2.5 | Examining incorrectly predicted forms

As noted above, though these models do not perform perfectly in terms of predictive accuracy, they do work reasonably well for the most common referential patterns, especially in light of the fact that they track choices across data from ten languages. They do so because of strong cross-linguistic commonalities in the primary drivers of referential choices, though there is also some variation in predictivity of certain factors across languages (e.g. humanness). Of course, no statistical model of usage can offer perfect predictions, as it is impossible to take into account all possible eventualities. However, as Kibrik et al. (2016) argue, any prediction of referential choices is inherently flawed because said choices are not categorical: Even in contexts that are strongly predictive (e.g. at low anaphoric distances), there is a degree of randomness and idiosyncratic variation that no model can capture adequately. As such, while all of the factors examined in this and the previous chapter contribute to the outcome of form choices, none are determinative, and there are always exceptions. Moreover, multiple less influential factors can outweigh an opposing factors that is more influential.

In this section, we briefly look at a number of specific instances where the model predictions are incorrect, and hence where the generalizations drawn from them do not yield the expected outcomes. Given the central role of anaphoric distance in determining lexicality, these cases mostly fall into one of two categories:

- lexical expressions at short distances (i.e. false negatives), or
- non-lexical expressions at long distances (i.e. false positives).

As seen above, anaphoric distance is at its most distinctive at the lower and upper ends of the scale, and less so inbetween. Many of the examples shown in this section exemplify specific pragmatic patterns involving forms choices that diverge from the general tendency. These kinds of contexts are not well captured by interactions in the models, in part because of their comparatively low frequency (relative to the basic cases), in part because their occurrence is simply not predictable from the properties of the surrounding discourse. This is the case with shifts from one narrative episode to another, whose effect on



using ROC-based threshold $t = 0.711$

		predicted	
		lexical	non-lex.
observed	lexical	151	142
	non-lex.	37	295

	cross-corpus mean	σ
sensitivity (= recall)	0.50	0.101
specificity	0.88	0.095
precision	0.80	0.165
baseline rate	0.51	0.102
accuracy	0.69	0.081
F_1-score	0.61	0.109
MCC	0.41	0.155

Figure 7.10 | The receiver operating characteristic (ROC) curve for the object model, and the confusion matrix and selected performance statistics based on the optimal ROC-based classification threshold.

referential choices is not determined by discourse factors, but by the demands of narrative structure.

In this example of a false negative from English, the subject of (86c), *my father* (⟨0027⟩), is lexical despite being human and the subject of its clause, and occurring at a distance of only $d = 1$ clause from its immediate antecedent, which itself was the subject of its clause. One possible explanation is that (86a) is the conclusion of one narrative episode about barley and breweries, and (86) starts the next about mowing; episodic boundary have been identified as a borderline after which speakers tend to use lexical expressions even if the referent was recently mentioned and is highly accessible (Marslen-Wilson et al. 1982; Anderson et al. 1983; Tomlin 1987a; Fox 1986, 1987a).

(86) English

a. *Father did sell the breweries barley once;*

	Father	*did*	*sell*	*the*	*breweries*	*barley*	*once*
	Father	do.PST	sell.INF	the	brewery.PL	barley	once
##	pn_np.h:a	lv_aux	v:pred	ln_det	np:p	np:p2	other
	0027				0610	0611	

‘Father did sell the breweries barley once;’

b. *he’d grow a bit of barley.*

	he	*=’d*	*grow*	*a*	*bit*	*of*	*barley*
	3SG.M	=would	grow.INF	a	bit	of	barley
##	pro.h:a	=lv_aux	v:pred	ln	ln_detq	ln	np:p
	0027						0612

‘he’d grow a bit of barley.’

c. *My father could mow, you know.*

	my	*father*	*could*	*mow*	*you*	*know*
	1SG.POSS	father	could	mow.INF	2SG	know.PRS
##	ln_pro.1:poss	np.h:s	lv_aux	v:pred	other	other
	0000	0027				

‘My father could mow, you know.’

[mc_english_kent02_0840-0841]

This example from Tabasaran is another case of a false negative. Here, the water buffalo (index ⟨0101⟩) had been repeatedly mentioned in the preceding clauses, each time as a pronoun in object position as in (87a); however, in (87b) it is then mentioned as a lexical NP in subject position in the next clause. As noted earlier, promotions from object to subject are predominantly lexical, and this effect appears to outweigh any other that would push towards a reduced form here, including recency. At the same time, the buffalo contrast with the oxen (⟨0039⟩) in the same clause, which precede them in linear order.

(87) Tabasaran

a. *jada murar ǯirʁuz abxundazxa k'ur.*

	jada	*mu-rar*	*ǯi<r>ʁu-z*
	PRT	PROX-PL(ABS)	<PL>stop-INF
##ds.neg	other	0.1:a dem_pro:p	v:pred
		0006 0101	

a-b-x-un-da	*=z-x-a*	*k'ur*
PFV-NSG-become-AOR-NEG	=1SG-APUD-ELAT	CIT
rv_aux	=rv-pro_1_a	other

'[Selim said,] Hey, I couldn't stop them [= the buffalo].'
[mc_tabasaran_work_0206]

b. *ʁušnu k'ur jicar ʁušušlamina gamšara.*

	ʁ-uš-nu	*k'ur*	*jic-ar*	*ʁ-uš-u*
	<PL>go-AOR	CIT	OX-PL(ABS)	PFV-<PL>go-PTCP
##ds	v:pred	other	ln_np	ln
			0039	

šˊ-l-a-mi-na	*gamš-ar*	*=a*
place-SUPER-ELAT-PROX-LAT	buffalo-PL(ABS)	=ADD
np:g	np:s	other
0093	0101	

'The buffalo also went down the place [i.e. off a cliff] where the oxen went.' [mc_tabasaran_work_0207]

Example (88) is also from Tabasaran, and similar to the previous example. Here, there are long same-role chains in both subject and object position. The subject *güla^{ʕ}li* 'Gulali' (⟨0005⟩) behaves as expected: After the initial mention as a lexical NP in (88a), subsequent mentions are consistently non-lexical. The object *murtir* 'eggs' (⟨0040⟩) however is mentioned lexically in two adjacent clauses, first in (88b) (which, it should be noted, is not the introduction of that referent into discourse), and then again in (88c); this pattern was already commented on in Section 6.8.4 above. After that, the referent is expressed non-lexically.

(88) Tabasaran

a. *a^{ʕ}ru güla^{ʕ}li,*

	a^{ʕ}-ru	*güla^{ʕ}li*
	go-FUT	Gulali
##	v:pred	pn_np.h:s
		0005

'Gulali went,'

b. *χuru murtir,*

	χ-uru	*murt-ir*
	bring-FUT	egg-PL(ABS)
##	0.h:a	v:pred	np:p
	0005		0040

'(he) brought eggs,'

c. *durxnu murtir,*

	d-u<r>x-nu	*murt-ir*
	PFV-<PL>boil-PST	egg-PL(ABS)
##	0.h:a	v:pred	np:p
	0005		0040

'(he) boiled the eggs,'

d. *živru,*

	živ-ru	
	put-FUT	
##	0.h:a	v:pred 0:p
	0005	0040

‘(he) served (them),’

e. *ip’uru murari.*

	ip’-uru	*mu-rari*
	<NSG>eat-FUT	PROX-PL(ERG)
##	v:pred	dem_pro.h:a 0:p
		0036 0040

‘and they ate (them).’ [mc_tabasaran_naz_0053]

The third and final example from the Tabasaran corpus showcases a false positive prediction: an extremely long-distance pronominal anaphor. Here, the narrative context primes expectations for what is about to happen (i.e. the sudden arrival of the magical horse, ⟨0016⟩), and so makes it unambigious what the pronoun refers to, despite the very long distance ($d = 20$ clauses) to its immediate antecedent. The next mention of the horse in the following clause is expectedly a zero.

(89) Tabasaran

a. [A vengeful magic horse appears night after night to destroy the grave of the three protagonists’ father. Each time, one of the brothers tries to guard the grave by himself, but becomes frightened and runs away, keeping the reason a secret. After a long discussion without any mention of the horse, the youngest brother convinces his siblings that he will go in their stead on the fifth night. Unafraid, he waits in the graveyard.]

b. *qa ʁuˤru hamu, kakraz hadmu vaχtna, qana jišnu ʁuˤru.*

	qa	*ʁ-uˤru*	*ha-mu*		*kakraz*	*ha-dmu*
	then	come-FUT	EMPH-PROX(ABS)		right_then	EMPH-3.P(ATTR)
##	other	v:pred	dem_pro.d:s 0016	##	other	ln_dem

vaχtna	*qana*	*jišnu*	*ʁ-uˤru*
time(ERG)	PRT	night(ERG)	come-FUT
np:other	other	np:other 0.d:s 0016	v:pred

‘And then it [= the horse] came, it came right then in the night.’
[mc_tabasaran_horse_0053]

In example (90) from Vera’a, the object *ni’igi* ‘her child’ (⟨0005⟩) of the clause in (90a) is repeated as a lexical NP in the next clause, which echoes the first. Repetitions of this kind are not uncommon in traditional narratives, and usually emphasize a key juncture in the story, or else serve to give the speaker an opportunity to plan the subsequent discourse without falling silent. Notably, the repetition here is not verbatim, as the subject of the second clause is reduced to zero, and the object is elaborated on.

(90) Vera’a

a. *Dine len ni’igi.*

	di	*ne*	*le*	*=n*	*ni’i*	*-gi*
	3SG	TAM2:3SG	transfer	=ART	small	-3SG
##	pro.h:a 0002	lv-pro_h_a	v:pred	=ln	np.h:p 0005	-rn_pro.h:poss 0002

‘And she delivered her child.’

b. *Len ni'igin ni'i 'aa̱man.*

	le	*=n*	*ni'i*	*-gi*	*=n*
	transfer	=ART	small	-3SG	=ART
##	0.h:a	v:pred	=ln	np.h:p -rn_pro.h:poss	=ln
	0002			0005 0002	

ni'i	*'aa̱man*
small	man
np:appos	rn
0005	

'(She) delivered her child, a boy.' [mc_veraa_mvbw_0011-0012]

The final example, from Mandarin, is not an incorrectly predicted one, but rather illustrates the often inexorable effect of low anaphoric distance: Here, *sùniǎo* '(nesting) bird' is realized as a zero in the second clause of (91), despite only just having been introduced into discourse in previous clause, being promoted from object to subject role, and being non-human, because the distance between the two mentions is essentially null.

(91) Mandarin

a. *Tā hūrán jiù kàndào le nà gè sùniǎo,*

	tā	*hūrán*	*jiù*	*kàndào*	*le*	*nà*	*gè*	*sùniǎo*
	3SG	suddenly	ADV	see	ASP	that	CL	bird
##	pro.h:a	other	other	v:pred	rv	ln_dem	ln	np:p
	0001							0115

'Suddenly she saw this bird,'

b. *jīngfēi zì běi ér nán.*

	jīngfēi	*zì*	*běi*	*ér*	*nán*
	fly	from	north	to	south
##	0:s	v:pred	np:obl	adp	np:g
	0115		0116		0117

'flying from north to south.' [mc_mandarin_hml_0092]

8 | Types of lexical expression

Where in last two chapters we have tested some of the mechanisms driving the broad choice between lexical and non-lexical expressions, in the last part of the analysis we focus our attention on the selection of specific types of lexical expressions over others. The status of these expressions in various approaches to referential choice have been discussed in Section 2.5. In particular, tested here are speakers' choices between

1. lexical NPs headed by proper names and common nouns (e.g. *Jane* vs. *the woman*, Section 8.1),
2. heavy and light lexical NPs (e.g. *the very big dog with shaggy fur* vs. *the dog*, Section 8.2), and
3. lexical NPs with and without demonstrative determiners (e.g. *this book* vs. *the book*, Section 8.3).

Each is tested individually in the following, though of course there is some degree of overlap between them, as for instance proper names tend to also be fairly light expressions and not have demonstrative modifiers. We also restrict our investigation to subject anaphors only, and leave objects and other roles and the complex interplay between expression types for future investigations.

As the subsamples used for each of these tests are comparatively small – lexical subjects are, after all, the exception – we here do not employ the advanced modelling strategies used in previous chapter, but instead rely on

logistic mixed-effects regression models, as used throughout Chapter 6 above for individual predictors, though as noted there, logistic regression is generally not ideal for heavily imbalanced data. We do make use of the full range of factors tested for general lexical anaphora, and again include corpus and speaker as random effects in the models. As before, the analysis are performed using the `glmer` function from the *lme4* R package (Bates et al. 2020).

8.1 | Proper names

8.1.1 | Definition and methodological issues

The status of proper names within systems of referential choice was discussed in in Section 2.5.1, and which NPs are classified as proper names has been defined in Section 4.2.1.1 above. As noted earlier, narrative texts are perhaps perhaps not the ideal text type for investigations into the selection of proper names, as their occurrence is highly dependent on text content, especially in traditional narratives. In fact, a number of corpora in the sample lack a sufficient number of proper name mentions in subject position (or any position) to yield statistically useful results: The Nafsan, Tulil, and Northern Kurdish corpora have very few cases, each with less than $N \leq 10$ instances across both roles ($P \leq 0.04$ of lexical mentions), and Cypriot Greek has no proper names in subject position, only as possessors and vocative expressions. In the other corpora, proper names make up between 7% and 34% of lexical subject mentions. Those corpora with too few cases are excluded from the following analyses, which leaves an actual sample composed of the English, Mandarin, Sanzhi Dargwa, Tabasaran, Teop, and Vera'a corpora. The resulting subsample is summarized in Table 8.1.

Specifically, there is a noticeably higher proportion of proper name mentions in the Tabasaran corpus ($P = 0.32$ proper), where there are two texts with characters (in both cases, three brothers) that are mainly distinguished by their names – though in a third text that similarly has three brothers as its protagonists, these are instead consistently differentiated by age rank (i.e. 'the eldest brother', 'the middle brother', 'the youngest brother'). A comparatively high rate is also found in Mandarin ($P = 0.34$ proper), which has a tendency to refer to individuals by their names rather than pronouns, especially in direct speech

	total number of			referents		lexical mentions		
corpus	speakers	texts	clauses	all	smpl.	proper	all	P(x)
English	3	4	5649	1934	93	56	163	0.34
Mandarin	3	3	1194	466	57	70	208	0.34
S. Dargwa	4	8	1066	359	51	18	116	0.16
Tabasaran	2	5	1383	398	56	66	204	0.32
Teop	4	4	1303	267	45	12	167	0.07
Vera'a	10	10	3608	656	128	87	492	0.18
totals	26	34	14203	4080	430	309	1350	—

Table 8.1 | Overview of the subsample for proper names and common lexical expressions.
The 'all referents' column contains the total number of discourse entities referred to in the corpus data, while 'sampled referents' are those matching the sampling criteria outlined in this section (i.e. types); 'mentions' are the anaphoric instantiations (i.e. tokens) of the sampled referents.

(similar tendencies are also found in other East Asian languages, e.g. Japanese).

(92) Mandarin
Zhè gè shíhòu, hè yuánshuài cái huǎngrán dàwù.

	zhè	*gè*	*shíhòu*	*hè*	*yuánshuài*	*cái*	*huǎngrán*
	this	CL	moment	He	Marshal	only	suddenly
##	ln	ln	np:other 0173	pn_np.h:s 0065	rn	other	other

dà	*wu*
big	comprehend
ln	np:pred
	0174

'In that moment, Marshal He suddenly understood.'

[mc_mandarin_hml_0150]

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	proper names	(*common noun*, proper name)
fixed effects	humanness	(*human*, non-human)
	total freq.	(*bottom 95%*, top 5%)
	ante. distance	(*1*–10+)
	co-ref. ment.	(*0*–5+)
	related ment.	(*0*–2+)
	comp. ment.	(*0*–6+)
	ante. role	(*subject*, object, other)
	ante. form	(*non-lexical*, lexical)
	sequence	(*second* < third < subsequent)
	clause type	(*independent*, other)
	clause length	(*1;2*–12+)
	transitivity	(*transitive*, other)
random effects	corpus	
	speaker	

a. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.03	–3.681	0.803	–4.58	<0.001
(A_1)	humanness	= non-human	0.07	–2.701	0.395	–6.83	<0.001
(B_1)	total freq.	= top 5%	2.76	1.015	0.236	4.31	<0.001
(C)	ante. distance	* [0, 9]	1.01	0.011	0.038	0.28	0.780
(D)	co-ref. ment.	* [0, 5]	1.21	0.194	0.089	2.17	0.030
(E)	related ment.	* [0, 2]	1.17	0.155	0.117	1.32	0.185
(F)	comp. ment.	* [0, 6]	1.03	0.032	0.057	0.57	0.571
(G_1)	ante. role	= object	0.75	–0.285	0.293	–0.97	0.332
(G_2)		= other	1.72	0.539	0.203	2.65	0.008
(H_1)	ante. form	= lexical	1.08	0.076	0.198	0.38	0.701
(I_1)	sequence	= third	1.75	0.560	0.286	1.96	0.050
(I_2)		= subsequent	1.41	0.341	0.376	0.90	0.365
(J_1)	clause type	= other	1.07	0.069	0.257	0.27	0.789
(K)	clause length	* [0, 10]	1.04	0.042	0.034	1.26	0.207
(L_1)	transitivity	= other	0.85	–0.164	0.193	–0.85	0.397

Table 8.2 | Regression model results for lexical expressions headed by proper names and common nouns, with corpus and speaker as random effects.

b. | random effect intercepts

	groups	σ
corpus	6	0.992
speaker	26	2.124

c. | scaled residuals

min.	lower	median	upper	max.
−2.288	−0.366	−0.120	−0.015	9.863

d. | correlation of fixed effects
(none above $r = |0.05|$)

e. | model evaluation

observations	1350	AIC	998
model deviance	964	log-likelihood	−482
residual d.f.	1333	conditional R^2	0.716
		marginal R^2	0.242

8.1.2 | Selection of proper names

Though we are chiefly focused on the selection of subject forms in this section, it is worthwhile to make a couple of preliminary observations about the overall distribution of proper names across roles. While in absolute terms, proper names are more common in subject position than in object position ($P = 0.36$ occuring in subject position vs. $P = 0.06$ in object position), the majority of proper name mentions occur in other positions ($P = 0.58$). In relative terms, proper names make up about a fifth of the lexical subjects and oblique arguments in the six corpora in this subsample overall ($P = 0.23$ of subjects, $P = 0.22$ of obliques), but only a fraction of the object mentions ($P = 0.05$). The low proportion of proper names in the latter comes as a surprise, since it is object anaphors that are generally expected to be more explicit than subject anaphors.

This pattern can be explained by the fact that the majority of proper name mentions are human ($P = 0.76$), a proportion that approaches categoricity for

subjects in particular ($P = 0.97$ for subjects), but is noticeably lower for other roles ($P = 0.80$ for objects and $P = 0.62$ for other positions). This mirrors the association of specific roles with humanness overall. This observation is particularly noteworthy, as it suggests that in narrative discourse, it is primarily narratively central and human referents that are deemed noteworthy enough to receive proper names, rather than those lacking in salience. Other types of referents are only identified by name if they refer to real-world entities or locations, which is most notable in the biographical narratives, but also happens occasionally in traditional narratives set in a particular location:

(93) Veraʼa
Ba duru ga ʼōg ʼa Lēmērig.

	ba	*duru*	*ga*	*ʼōg*	*a*	*Lēmērig*
	but	3DL	STAT	stay	LOC.SP	Lēmērig
##	other	pro.h:s	lv	v:pred	adp	pn_np:l
		0002				0004

‘But the two lived at Lēmērig.’ [mc_veraa_as1_0003]

This profile for proper name subjects is precisely what we find reflected in the the regression model fit to the data, which is summarized in Table 8.2. Here, it is unsurprisingly humanness that comes out as highly significant ($e^{\beta} = 1 / 0.07 = 14.29$ times higher odds if the referent is human, $p < 0.001$). As such, this factor is not particularly elucidating, since there are very few non-human referents mentioned as proper names in subject position, and the overwhelming majority of proper names is human.

Clearly linked to humanness is protagonisthood (as represented by total mention frequency), which is similarly significant: Mentions of highly frequent referents are $e^{\beta} = 2.76$ times more likely to be realized with proper names ($p < 0.001$). As noted above, this means that only the most narratively central referents even have names. However, these also tend to be the ones that are the most topical and accessible, which according to accessibility theory (Ariel 1990, see Section 2.2.4) and other models of discourse anaphora should be less likely overall to require a form as informative and specifically referring as a proper name. Though these data are very much tentative, we here rather decidedly find the opposite to be the case (see also Schiborr 2017 for similar findings).

The remainder of the model makes this picture even clearer: Very few of the other factors in the model come anywhere near significance, and perhaps most notably, anaphoric distance is not among them – in fact, distance comes out as essentially irrelevant to the selection of proper names ($p < 0.780$). In other words, the factor that almost by itself explains the broader choice between lexical and non-lexical expressions (Chapter 7) has no bearing on the choice between proper names and other lexical expressions. A similar pattern shows for the form of the antecedent ($p < 0.701$), which likewise strongly affects lexical choices overall.

The only factors besides humanness and total mention frequency that pass the α-level of $p < 0.05$ in terms of significance are the number of recent co-referential mentions, sequence of mentions, and antecedent role, and those only barely. The rate of proper names increases with proportionally with recent mention frequency ($e^{\beta} = 1.21$ times higher odds for each mention, $p < 0.030$), again contra considerations of accessibility. As regards sequence of mention, third-in-sequence mentions are more likely to be proper names compared to the reference level of second-in-sequence mentions ($e^{\beta} = 1.75$, $p < 0.050$) – an entirely unexpected result and very likely a statistical artefact, as this is a very small subgroup. Lastly, role shifts from non-core positions to subject come out as significant in the model ($e^{\beta} = 1.72$, $p < 0.008$), but as we have noted earlier in Section 6.7.2, this transition is in fact quite rare, which, compounded with the relative rarity of proper name subjects, casts doubt on the validity of this particular association.

The regression model also indicates substantial variation between corpora and speakers, with random effect intercepts being $\sigma = 0.992$ for corpus and $\sigma = 2.124$ for speaker. This is again the result of major differences in content between texts (and hence speakers), as some texts have more proper names in them, some less, some none at all. As noted above in Section 4.2.1.1, this partly aligns with differences between text types: Biographical narratives placed in the real world generally have a higher proportion of proper name mentions than traditional narratives, in which characters are commonly based on narrative archetypes and may not be assigned names.

8.2 | Phrase weight

8.2.1 | Definition and methodological issues

The complexity of a noun phrase is a measure of informativity that is fairly straightforward to operationalize and quantify. Here, we employ a binary classification of phrase weights – heavy and light – among lexical expressions, the definition of which was already given in Section 4.2.1.2 above. To re-iterate, any lexical NP that contains an additional modifier, such as an adnominal adjective or possessor or a relative clause, or is coordinated is classified as heavy (e.g. *her very nice book that I read the other day*). Expressions that lack any such modifiers are classified as light. The exception to this rule are determiners (e.g. *the book*), which are not counted as contributing to phrase weight.

The binary classification used here is of course a rather coarse-grained one, as no further distinctions are made among heavy expressions – complexity as a measure of informativity spans a continuum, after all – but any finer grading of weights is fraught with conceptual issues, as different types of modifiers likely contribute differently to overall complexity. For this reason, we limit ourselves here to a binary weight distinction, and leave more nuanced approaches for future research. The subsample used for the following analysis is summarized in Table 8.3; unlike with proper names above, no corpora or texts are excluded from the sample.

8.2.2 | Selection of heavy expressions

First of all, note that the overall rates of heavy expression given in Table 8.3 span a fairly large range across corpora, from less than a fifth in Sanzhi Dargwa ($P = 0.15$) and Vera'a ($P = 0.18$), to over a third in English ($P = 0.37$) and Northern Kurdish ($P = 0.36$). While this appears to suggest that corpora differ substantially in their preference for heavy and informative subjects, greater variation is in fact found between the texts in each corpus, ranging from none to over 50% heavy expressions.[1] What we hence see here again is

1 Nevertheless, Levene's test of homogeneity of variances ($p < 0.375$) indicates that the variances across corpora are fairly homogenous; an ANOVA fit to the data ($F(9, 44) = 1.13$,

	total number of			referents		lexical mentions		
corpus	speakers	texts	clauses	all	smpl.	heavy	all	P(x)
C. Greek	1	3	1070	296	32	24	86	0.28
English	3	4	5649	1934	93	61	163	0.37
Mandarin	3	3	1194	466	57	71	208	0.34
Nafsan	4	9	1012	262	43	29	142	0.20
N. Kurdish	1	2	1841	298	46	49	138	0.36
S. Dargwa	4	8	1066	359	51	17	116	0.15
Tabasaran	2	5	1383	398	56	55	204	0.27
Teop	4	4	1303	267	45	31	167	0.19
Tulil	5	6	1264	383	59	43	151	0.28
Vera'a	10	10	3608	656	128	89	492	0.18
totals	37	54	19390	5319	610	469	1867	—

Table 8.3 | Overview of the subsample for heavy and light lexical expressions.
The 'all referents' column contains the total number of discourse entities referred to in the corpus data, while 'sampled referents' are those matching the sampling criteria outlined in this section (i.e. types); 'mentions' are the anaphoric instantiations (i.e. tokens) of the sampled referents.

the effect of textual content on referential choices, rather than differentiation between corpora and languages.

As regards the selection of heavy expressions, the logistic regression model summarized in Table 8.4 reveals that only a small handful of factors come out as influential: total mention frequency (i.e. protagonisthood), number of co-referential mentions, transitivity, and curiously, the number of mentions of related referents. Anaphoric distance also matters – if only marginally – though with a comparatively small effect, as every clause of distance to antecedent only raises the odds of a heavy expression $e^{\beta} = 1.05$ times ($p < 0.059$). But in combination with higher co-referential mention frequencies being associated with lighter expressions, with the odds of a heavy expression decreasing $e^{\beta} = 0.87$ times for each co-referential mention in recent discourse (i.e. in the

$p < 0.364$) likewise shows that the differences between the texts in each corpus are not statistically significant overall.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	phrase weight	(*light*, heavy)
fixed effects	humanness	(*human*, non-human)
	total freq.	(*bottom 95%*, top 5%)
	ante. distance	(*1*–10+)
	co-ref. ment.	(*0*–5+)
	related ment.	(*0*–2+)
	comp. ment.	(*0*–6+)
	ante. role	(*subject*, object, other)
	ante. form	(*non-lexical*, lexical)
	sequence	(*second* < third < subsequent)
	clause type	(*independent*, other)
	clause length	(*1;2*–12+)
	transitivity	(*transitive*, other)
random effects	corpus	
	speaker	

a. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.16	–1.810	0.370	–4.89	<0.001
(A_1)	humanness	= non-human	1.12	0.114	0.164	0.70	0.486
(B_1)	total freq.	= top 5%	0.70	–0.351	0.143	–2.45	0.014
(C)	ante. distance	* [0, 9]	1.05	0.047	0.025	1.89	0.059
(D)	co-ref. ment.	* [0, 5]	0.87	–0.138	0.066	–2.08	0.037
(E)	related ment.	* [0, 2]	1.33	0.285	0.087	3.28	0.001
(F)	comp. ment.	* [0, 6]	1.03	0.028	0.040	0.71	0.479
(G_1)	ante. role	= object	1.01	0.008	0.170	0.05	0.964
(G_2)		= other	0.91	–0.094	0.136	–0.69	0.488
(H_1)	ante. form	= lexical	1.08	0.081	0.134	0.61	0.543
(I_1)	sequence	= third	1.14	0.127	0.141	0.90	0.366
(I_2)		= subsequent	1.12	0.115	0.181	0.63	0.526
(J_1)	clause type	= other	1.31	0.270	0.146	1.85	0.065
(K)	clause length	* [0, 10]	1.03	0.031	0.024	1.28	0.200
(L_1)	transitivity	= other	1.38	0.320	0.142	2.25	0.024

Table 8.4 | Regression model results for the selection of heavy and light lexical expressions, with corpus and speaker as random effects.

b. | random effect intercepts

	groups	σ
corpus	10	0.310
speaker	37	0.818

c. | scaled residuals

min.	lower	median	upper	max.
−1.512	−0.579	−0.406	0.648	4.911

d. | correlation of fixed effects
(none above $r = |0.05|$)

e. | model evaluation

observations	1867	AIC	1961
model deviance	1927	log-likelihood	−964
residual d.f.	1850	conditional R^2	0.240
		marginal R^2	0.063

preceding six clauses, $p < 0.037$), this does suggest that the selection of heavier forms is in fact somewhat dependent on the structure of the local discourse. That is, heavier expressions become more likely the less recent and prominent the referent in question is. Higher total mention frequencies being noticeably predictive of lighter expressions ($e^{\beta} = 0.70$, $p < 0.014$) lends further support to this observation.

Transitivity also comes up as marginally significant in the model ($p < 0.024$. Subjects of transitive clauses are not only more likely to be non-lexical, as noted above, but if lexical, also more likely to be less complex and informative, with $e^{\beta} = 1 / 1.38 = 0.72$ times lower odds compared to subjects of intransitive clauses. Lastly, the effect of related mention frequency presents a bit of a mystery, as it suggests that heavier expressions become more common for each mereologically related referent mentioned recently ($e^{\beta} = 1.33$, $p < 0.001$), contrary both to expectations and the relatively weak effect of this factor on the choice of lexical expressions in general; this is most likely a statistical artefact.

But as with proper names above, what is most notable about the model results is which factors appear to either not matter at all, or only marginally. This applies to most factors related to properties of preceding discourse, and could be said to include anaphoric distance as well, given its minimal effect. Notably, humanness does not matter, despite the relatively significant effect of protagonisthood. There is also substantially less cross-corpus and inter-speaker variation (random effects intercept $\sigma = 0.310$ for corpora and $\sigma = 0.818$ for speakers) than among proper names, likely due to the more consistent occurrence of heavy expressions (vs. proper names) across texts, as unlike proper names, heavy expressions are presumably chosen independently of the content of a text.

In sum, the selection of heavier expressions for subject anaphors is somewhat affected by discourse structure and narrative prominence, but not conclusively so, and will require further investigation. Perhaps most notable is the association between lighter forms and transitive clauses (see Section 2.2.5.1).

8.3 | Demonstrative determiners

8.3.1 | Definition and methodological issues

The third and final distinction among lexical expression we will be examining here is those with and without demonstrative determiners (e.g. *this book* vs. *the book*). The classification criteria for this distinction were given in Section 4.2.1.3, and its status in the literature outlined in Section 2.5.3. The accessibility hierarchy (Ariel 1990) makes perhaps the most substantial claims about the selection of NPs with demonstrative determiners: All else being equal, expressions with such determiners supposedly indicate lower levels of accessibility than those without. Accessibility theory further suggests that proximal and distal demonstratives differ in terms of accessibility marking; it is not possible to test this claim here, for the reasons given in Section 4.2.1.3 above.

As with proper names and heavy expressions in the previous two sections, we will here be testing the selection of NPs with demonstrative determiners against all other lexical expressions. Unfortunately, however, not all of the corpora possess the requisite GRAID annotations for this analysis, as the symbols for demonstrative determiners (⟨ln_dem⟩ and ⟨rn_dem⟩) are not part of

	total number of			referents		lexical mentions		
corpus	speakers	texts	clauses	all	smpl.	dem.	all	P(x)
English	3	4	5649	1934	93	23	163	0.14
Mandarin	3	3	1194	466	57	43	208	0.21
N. Kurdish	1	2	1841	298	46	6	138	0.04
S. Dargwa	4	8	1066	359	51	23	116	0.20
Tabasaran	2	5	1383	398	56	36	204	0.18
Vera'a	10	10	3608	656	128	10	492	0.02
totals	23	32	14741	4111	431	141	1321	—

Table 8.5 | Overview of the subsample for lexical expressions with and without demonstrative determiners.
The 'all referents' column contains the total number of discourse entities referred to in the corpus data, while 'sampled referents' are those matching the sampling criteria outlined in this section (i.e. types); 'mentions' are the anaphoric instantiations (i.e. tokens) of the sampled referents.

base GRAID specification, and as such have not been applied universally to the Multi-CAST corpora.[2] As such, the following analysis is limited to those corpora that feature these annotations, in particular the English, Northern Kurdish, Mandarin, Sanzhi Dargwa, Tabasaran, and Vera'a corpora. Of course, as this renders the findings presented in this section even more tentative than the remainder of this study, the conclusions drawn here should be viewed in an appropriately cautious light. An overview of this subsample is provided in Table 8.5. It should be noted that a number of corpora in this subsample have rather low rates of demonstrate NPs; though this might hurt the predictive power of the regression analysis to some extent, we here nevertheless include all data.

2 Given knowledge of earlier findings related to the matter of demonstratives in Schiborr 2017, the addition of these annotation symbols has had rather low priority and hence could not be finished in time for submission, as it requires input from the original annotators.

subjects | generalized linear mixed-effects model
fit by maximum likelihood approximation (binomial, logit)

response	modified by demonstrative	(*no*, yes)
fixed effects	humanness	(*human*, non-human)
	total freq.	(*bottom 95%*, top 5%)
	ante. distance	(*1*–10+)
	co-ref. ment.	(*0*–5+)
	related ment.	(*0*–2+)
	comp. ment.	(*0*–6+)
	ante. role	(*subject*, object, other)
	ante. form	(*non-lexical*, lexical)
	sequence	(*second* < third < subsequent)
	clause type	(*independent*, other)
	clause length	(*1;2*–12+)
	transitivity	(*transitive*, other)
random effects	corpus	
	speaker	

a. | fixed effect coefficients

			e^{β}	β	SE	*z*-val.	*p*-val.
	(intercept)	—	0.10	–2.346	0.724	–3.24	0.001
(A_1)	humanness	= non-human	1.47	0.385	0.279	1.38	0.169
(B_1)	total freq.	= top 5%	1.02	0.023	0.246	0.09	0.925
(C)	ante. distance	* [0, 9]	0.94	–0.057	0.041	–1.38	0.167
(D)	co-ref. ment.	* [0, 5]	0.93	–0.077	0.103	–0.75	0.451
(E)	related ment.	* [0, 2]	0.72	–0.332	0.168	–1.98	0.048
(F)	comp. ment.	* [0, 6]	0.98	–0.019	0.067	–0.28	0.779
(G_1)	ante. role	= object	1.56	0.446	0.266	1.68	0.094
(G_2)		= other	0.85	–0.164	0.243	–0.67	0.501
(H_1)	ante. form	= lexical	1.07	0.063	0.221	0.29	0.775
(I_1)	sequence	= third	1.42	0.353	0.250	1.41	0.158
(I_2)		= subsequent	0.71	–0.337	0.293	–1.15	0.250
(J_1)	clause type	= other	1.43	0.355	0.241	1.48	0.140
(K)	clause length	* [0, 10]	1.05	0.046	0.041	1.13	0.258
(L_1)	transitivity	= other	0.94	–0.064	0.233	–0.28	0.783

Table 8.6 | Regression model results for the selection of lexical expressions with and without demonstrative determiners, with corpus and speaker as random effects.

b. | random effect intercepts

	groups	σ
corpus	6	0.987
speaker	23	1.105

c. | scaled residuals

min.	lower	median	upper	max.
−1.061	−0.362	−0.198	−0.094	7.581

d. | correlation of fixed effects
(none above $r = |0.05|$)

e. | model evaluation

observations	1321	AIC	791
model deviance	757	log-likelihood	−378
residual d.f.	1304	conditional R^2	0.424
		marginal R^2	0.039

8.3.2 | Selection of NPs with demonstrative determiners

The logistic regression model fit to this subsample is summarized in Table 8.6. Even more so than proper names and heavy NPs, we here find that no factor in the model comes out as strongly predictive – in fact, for demonstrative NPs, only a single predictor passes below an α-level of $p < 0.05$. This predictor is the frequency of mentions of mereologically related referents ($e^{\beta} = 0.72$, $p < 0.048$), which rather curiously has been significant in all three models presented in this chapter, but for whose apparent influence, as before, no ready explanation comes to mind – aside from attributing it to a statistical fluke. Besides this, we also find a substantial degree of inter-corpus (random effect intercept $\sigma = 0.987$) and inter-speaker ($\sigma = 1.105$) variation present in the model results, likely due to the small sample sizes.

9 | Synthesis

This chapter brings the observations in the previous four chapters together to discuss key findings, identify patterns, and ultimately draw a number of tentative conclusions on the selection of lexical expressions specifically and the mechanisms of referential choice in natural discourse in general. Section 9.1 focuses on the results of the predictive model presented in Chapter 7 and individual factor associations in Chapter 6, touching on the central position of recency in predicting form choices and the special role of same-role chain contexts. It also discusses variation between corpora, including as regards the sensitivity towards certain tested factors, in particular humanness.

Section 9.2 attempts to explain the cross-linguistically stable rates of lexical expression observed in Chapter 5 through juxtaposition with the highly prevalent contexts in which lexical expressions are the rarest, the aforementioned same-role chains. Lastly, Section 9.3 collects a number of smaller observations and more general conclusions, including on the selection of various types of lexical expressions as tested in Chapter 8, and the basicness of the basic lexical–non-lexical choice.

9.1 | Predicting lexical anaphors

9.1.1 | The primacy of recency

Perhaps the most significant take-away from the predictive model is that the lion's share of the variance in the data is explained by anaphoric distance alone. Of all contributing factors, distance is by far the most influential, indeed to such a degree that the model could still predict referential choices with only a marginally lower degree of accuracy solely on the basis of distance.[1] In the grand scheme of things, considerations of recency are thus the driving factor determining referential choices for anaphoric mentions, both in subject and object position, across all ten corpora tested. This fundamental association between recency and form of expression has been noted in the literature, as discussed in Chapter 2, and specifically also in Kibrik et al. (2013, 2016), who note the same centrality of recency effects in their modelling approaches.[2] Recency effects have been tied to volume constraints on short-term memory if distance is long, and the maintenance of discourse coherence if short (Chafe 1976; Clark & Sengul 1979; Givón 1983a; Ariel 1990; etc.).

While we find these associations confirmed here, we also see that as far as anticipating speakers' referential choices is concerned, anaphoric distance serves as a predictor par excellence. What the model suggests, then, is that referential choices are in essence stratified by the distance to the antecedent: All else being equal, very short distances lead to mostly non-lexical mentions, while longer distances in turn lead to increasing proportions of lexical expressions. At intermediate distances matters are more ambigious; this is where other properties of the local discourse and further relevant factors are at their most distinctive, and also where most of the variation between corpora can be found. In essence, this means that close mentions are reduced, and lexical mentions are distant; distant mentions, however, are not necessarily lexical (see Section 9.1.5).

1 Performance statistics for a gradient boosting model fit to the subject sample, with only distance and corpus as predictors: cross-corpus mean accuracy $acc = 0.83$, $\sigma = 0.042$; specificity $spc = 0.89$, $\sigma = 0.041$; precision $prc = 0.61$, $\sigma = 0.081$; $F_1 = 0.59$, $\sigma = 0.072$; Matthews correlation coefficient $MCC = 0.49$, $\sigma = 0.093$.

2 But note that according to these authors, recency effects are not limited to textual distance; see e.g. Kibrik & Krasavina (2005) on role of rhetorical structure and rhetorical distance.

9.1.2 | Stability in convergent contexts

Though themselves not as predictive, other factors reinforce the association of lexical and non-lexical referring expressions with distance, converging into specific extreme contexts. Short anaphoric distance, high recent mention frequency, and non-lexical antecedents in subject position delineate a context with especially high discourse coherence (Givón 1983a), in which speakers almost exclusively select reduced forms for subject anaphors in all corpora in the sample, with lexical expression consequently being quite rare, though also not categorically absent. This context is accounts for a substantial proportion of mentions ($P = 0.45$, $\sigma = 0.064$ of subject anaphors, though only $P = 0.26$, $\sigma = 0.046$ of objects), and often persists across multiple subsequent clauses, forming anaphoric chains (cf. Givón 1983a, 2017). Higher recent mention frequencies here serve as an indicator of longer chains and the continuous presence of the referent in preceding discourse. The centrality of this context is well supported by evidence from variationist typology, which shows that it is here that zero subjects are the most frequent, even in languages with a relatively low tolerance for zero such as English (Torres Cacoullos & Travis 2019: 672–674; Li & Bayley 2018: 151; and, for A only, Gipper 2016: 166–167; see also Schnell & Barth 2020).

The most frequently mentioned and most consistently topical referents are hence only very rarely lexically expressed. Only where they fall out of the focus of the narrative, and are then taken up again, do they require the use of a more informative (and hence likely lexical) expression. The diametrically opposed configuration is hence defined by long anaphoric distances and low recent mention frequency, which go hand-in-hand, as well as lexical, non-subject antecedents. The recovery from memory of referents in such context requires additional information, especially if these referents are also inherently non-salient (i.e. non-human and/or not narratively central). As such, this is where lexical expressions are the preferred choice, though as mentioned before, not the only one: While the majority of lexical subjects occur in context of low discourse coherence, of which the aforementioned is perhaps the most extreme, non-lexical forms may be selected for anaphors in any context – if at varying rates – though of course there are likely to be cross-linguistic differences in the selection of zero and pronominal forms in different contexts, which are not tested for here. We will discuss the universal availability of non-lexical expressions further in Section 9.1.5 below.

In sum, we find very few lexical subjects in high coherence contexts, and an increasing number at increasing distances to the antecedent. The rate of lexical expression is hence inversely proportional to the "topicality" of the referent (Chafe 1976), but is in fact better modelled as a function of anaphoric distance and the presence or absence of certain convergent factors, such as the role of the antecedent.[3]

Notably, this pattern is found in all corpora in the sample. The model indicates that there are no fundamental divergences in behaviour between the ten corpora. While there are a number of differences which lead to the relatively high influence of the corpus variable seen in Figures 7.3 and 7.7, they for the most part occur only within high-level interactions and affect only small subsets of the data. As noted above, the influence of cross-corpus variability is amplified by the corpus variable having many discrete levels, which give the modelling algorithm more opportunities for splits compared to, for instance, a binary variable.

We hence find a substantial degree of cross-corpus stability with regards to mechanisms guiding the selection of lexical expressions, especially in the most common (i.e. convergent) contexts outlined above. The proportions of lexical expressions at specific distances from the antecedent – ceteris paribus – can differ between languages, being shifted either higher or lower, based in part on the informativity of pronouns and similar considerations. This, however, appears to most strongly affect mentions at intermediate distances from the antecedent, which we will discuss in the next section; at the extreme ends of the distance scale – that is, in the most common cases – all corpora pattern essentially the same.

3 But there are of course also certain routinized associations with lexical NPs that are independent from recency and coherence concerns, such as the implication of non-co-reference with high accessible referents, as well as at episodic boundaries, which we will touch upon further below.

9.1.3 | Variability in ambivalent contexts

As noted above, a substantial proportion of anaphoric mentions in subject and object position occurs at very low and very high distance values from the antecedent. The remainder falls somewhere inbetween, at intermediate distances. Anaphors at these distances are substantially more ambivalent in terms of form selection, and so it is here that other factors besides distance bear out their influence. Whereas at extreme distances predictability (or lack thereof) is chiefly determined by recency, inbetween other factors start to play a disambiguatory role, and may flip outcomes one way or another.

In other words, if the speaker cannot rely on topic continuity alone to carry the identification of the referent, but said referent has also been mentioned recently enough to remain active in working memory, form choices rely more noticeably on properties such as the inherent salience of the referent. In the predictive model, this is most clearly visible in the interactions of distance with certain other factors, where the predictiveness of the latter for either lexical or non-lexical expressions is lowest at the low and high ends of distance spectrum, and highest inbetween. This pattern is especially evident for the factors of humanness and antecedent form, though it does not seem to apply to the role of the antecedent, presumably since this factor together with distance is involved in defining the highly predictive same-role chain context.

Precisely which factors make how much of a difference at intermediate distances appears to differ to some degree from language to language, however. Specifically, certain properties appear to make much less of a difference in certain corpora; in the present sample, this is most obviously the case with English, which, as seen throughout Chapter 6, is mostly or fully insensitive to the effects of humanness and protagonisthood, role and form of antecedent, recent mention frequency, sequence of mention, and others. Perhaps most surprising is the apparent insensitivity of English to humanness distinctions, which cannot be reduced to considerations of accessibility; we will discuss this further in the next section.

9.1.4 | Inherent salience of referents

As noted above, anaphoric distance exerts by far the greatest effect within the predictive models. Factors related to the inherent salience of referent, such as humanness and protagonisthood, conversely come out as much less predictive than might be expected (cf. e.g. Givón 1983a; Ariel 2001), in part due to their strong associations with distance, which leaves little of the variance in the data for them to explain.

But as the interaction between humanness and distance in Figure 7.5d above reveals, humanness does in fact affect the selection of lexical expression, shifting choices towards reduced forms across contexts, and doing so at all distances from the antecedent. The effect of low recency on lexicality is in essence "delayed" for human referents, so that human referents remain non-lexical at longer distances than non-human referents, that is, when not mentioned, the activation of salient referents decays at a slower rate. As noted in the previous section, this effect is greatest at short to intermediate distances, but levels out as the distance to the antecedent increases, as there the inherent salience of human referents begins to add less to their identifiability.

Generally speaking, human and narratively central referents in the corpus data are substantially more likely to be subjects, recurrent, and hence non-lexical, indicating their favoured status in discourse. The organization of discourse is hence strongly anthropocentric; this reflects the well-known association between humanness and topicality (Chafe 1994; Dahl 2000, 2008), and more generally, the favoured position of human entities in human cognition (Ariel 2001; Fukumura & van Gompel 2011; Fraurud 1996; Dahl & Fraurud 1996). All else being equal, the higher the salience of a referent, the more likely it is to be realized non-lexically; non-topical, non-human, and not narratively central (i.e. non-protagonist) referents are conversely more likely to receive a lexical expression.

As such, humanness distinctions do not themselves drive referential choices; rather, humanness confers a cognitively preferred status to human referents, which in ambivalent contexts in particular may tip the balance of referential choices one way or another. But unlike other, contextual factors, the effect of humanness (and, to a lesser degree, protagonisthood), being inherent properties of the referent, is also felt ubiquitously, and applies to a degree in all contexts. In terms of prediction, however, this effect is generally outweighed by more influential factors (i.e. anaphoric distance) in the con-

vergent contexts described above. Humanness hence does not come out as particularly influential by itself in the models presented in Chapter 7, as in these contexts its effects are subsumed by other, convergent factors such as distance, recent mention frequency, and antecedent role and form, with which humanness and protagonisthood are strongly associated.

It is important to note, however, that the corpora in the sample differ noticeably in how sensitive they are to humanness distinctions as regards the selection of lexical expressions. As already noted in Section 6.1, form choices in certain corpora, most notably Northern Kurdish and Tabasaran, appear to be substantially more sensitive to humanness than the cross-corpus mean, either for mentions in subject or object position, or for both. Others, such as English and Sanzhi Dargwa, are less sensitive; English in particular appears to be almost completely indifferent to humanness distinctions. The English data also have a comparatively lower proportion of human subjects overall ($P = 0.63$ vs. cross-corpus mean $P = 0.85$, $\sigma = 0.109$), likely owing to the prevalence of first-person references in the texts (see Section 4.6.3). An essentially opposite pattern is found for objects in the Tulil corpus (see Section 6.1), where it is human mentions that are noticeably less likely to be lexical compared to the cross-corpus mean.

One possible avenue for explaining the apparent parametricization of humanness effects across languages is via cross-linguistic differences in the informativity of pronouns and the preference for zero anaphors. If we assume that humanness distinctions are a crucial part of any message in any language, then if a language has no means of communicating this distinction through reduced forms (i.e. pronouns, or else agreement), then we might expect rates of lexicality to be more sensitive to the humanness of referents, as these are then the only referring expressions that can express humanness distinctions. In English, the conditions for the selection of zero are far more restrictive than in any other language in the sample, so that speakers are in most cases presented with an essentially binary choice between pronominal and and lexicals forms, both of which happen to indicate humanness. Speakers are hence able to make humanness distinctions essentially anywhere, irrespective of their actual form choices, so that the attenuating effect of humanness on lexicality found in other languages does not play out in English. Conversely, for many other languages, humanness may only be expressed overtly on lexical expressions; here, speakers are by necessity forced to switch to lexical expressions in order to express humanness distinctions.

Variation in the sensitivity to humanness between languages could also be attributed to potential language-specific parameter weightings. Both humanness and anaphoric distance (or more generally, topicality) affect the selection of lexical expressions; where both line up, the result is the stable low or high levels of lexicality noted above, but where they do not, the outcome depends on language-specific weightings: If a given language places greater weight on humanness, it will pattern like Northern Kurdish and Tabasaran, where non-human referents are substantially more likely to be lexical; if it instead ranks topicality more highly, it will pattern like English, where humanness distinctions are subdued and distance effects on lexicality delayed, as noted in Section 6.3.

Furthermore, languages differ in their tolerance of non-human pronouns, which is high in English ($P = 0.55$ of all pronouns) but low in Northern Kurdish ($P = 0.04$). In the latter, this narrows the choices for non-human referents to just zero versus lexical, which might further tip the proportions in favour of lexical expressions if the referent in question is insufficiently accessible.

Given the limited number of languages in the sample and overall small sample size, the causes for the apparent parametricization of humanness effects across languages will ultimately have to remain speculation at this point, though the cross-linguistic variability of these patterns does underscore the importance of typological diversity in the sample.[4]

9.1.5 | Non-categoricity of referential choices

Another key observation is that there is no set of circumstances in any of the tested corpora, including in the convergent contexts described above as well as the more varied ones that fall inbetween, in which referential choices prefer lexical or non-lexical expressions to the total or near-total exclusion of the alternative.[5] Either type of expression can occur under any circumstances with some likelihood, if at substantially different proportions. This echoes Kibrik et al. (2013, 2016), who conclude that referential choices are

4 That is, assuming that the observed patterns are not a quirk of the corpus data, but actual typological properties of the tested languages.

5 But compare this to clause-internal anaphora, where hard constraints can and do hold (e.g. reflexive binding).

inherently non-categorical. In terms of accessibility, this means that reduced forms may refer to comparatively less accessible referents than their positions on various referential scales may suggest at first glance (Givón 1983a; Ariel 1990). This distributional property of pronouns has been noted in Ariel (1996: 22), and is there explained in terms of the persistent salience of certain referents: Highly salient referents continue to license pronominal forms at higher distances, whereas less salient referents switch to lexical expressions instead. But the data from Multi-CAST suggest that this association goes beyond referent accessibility, as pronominal expressions are available even for referents with minimal accessibility (i.e. at very long distances) and low inherent salience (i.e. non-human), as noted above in Section 6.3.3.

That being said, lexical expressions are noticeably more restricted in terms of availability than non-lexical ones, which can occur with appreciable frequency essentially anywhere. There are contexts in which lexical expressions are strongly dispreferred (i.e. anaphoric chains), but there are no corresponding contexts in which the same is true for non-lexical expressions, at least not to the same extent. For subject anaphors in particular, non-lexical forms are essentially always a viable option: Even if they are not the preferred choice in a given context, they are not dispreferred to the same extent that lexical expressions are in contexts that strongly favour non-lexical expressions. This dichotomy can be explained at least in part through considerations of economy of production, but lexical overcoding may also lead to unintended interpretations of non-co-reference (cf. Ariel 1990; Grosz et al. 1995).

In sum, this and corresponding conclusions in Kibrik et al. (2013, 2016) indicate that there are practical limits to the predictability of referential choices. Given the above-noted prevalence of non-lexical expressions in essentially all contexts, it is easier to say when to expect a non-lexical expression than a lexical one. Of course, this applies only to the binary choice of lexical and non-lexical forms, and does not take into account the selection of pronominal and zero expressions, which exhibit a well-documented degree of cross-linguistic variation (Perlmutter 1971; Bickel 2003; Neeleman & Szendröi 2007; Travis & Lindstrom 2016; etc.).[6]

6 Preliminary counts show that rates of zero anaphora (vs. pronominal) for third-person subjects do not drop consistently with increasing distance for distances below $d < 10$ clauses from the antecedent (cross-corpus mean Spearman's rank correlation coefficient $\rho = -0.05$, $\sigma = 0.460$). The exception are languages in which zero expression is largely restricted to

9.1.6 | Differences between roles

In terms of predicting the selection of lexical expression, many of the same considerations hold for both subject and object anaphors. The fact that form selection in both positions is sensitive to largely the same factors (and considerations of discourse coherence) is surprising in a way, given that objects are generally associated with largely unexpected – and hence not very cohesive – material. This expectation perhaps stems from the strong association of the object role with new information ($P = 0.29$, $\sigma = 0.083$ of objects are new across corpora, but only $P = 0.08$, $\sigma = 0.032$ of subjects); for anaphoric objects, this is not as pronounced.

Nevertheless, there are still notable differences in how subject and object anaphors pattern. There are a number of language-specific and generally applicable factors affecting lexical choices, such as the above-mentioned dispreference for lexical human objects in Tulil, shifted thresholds for form associations with anaphoric distance in English, and in all corpora, a less pronounced influence of role and form continuity from antecedent to anaphor, all of which are reflected in the much less cross-linguistically stable baseline lexicality rate for objects. Even so, the informativity of object expressions is largely determined by the recoverability of the referent, and same the contexts that make subjects predominantly non-lexical (low anaphoric distance, role continuity, high recent mention frequency, etc.) also affect the form of objects.

As such, "topical" objects behave much the same as "topical" subjects, but since objects are not the primary carriers of topical information in the clause – on the contrary, as the prototypical object is indefinite, inanimate, and rhematic (Comrie 1979: 19) – much fewer objects occur in contexts of high discourse coherence. The conditions for these contexts are also more limited, in that, for instance, objects become predominantly lexical at shorter distances from their antecedent compared to subjects. Subjects are overall much less likely to be expressed lexically due to being inherently more salient

same-role chains (English, Vera'a), where there is a notably drop in rates from distances of $d = 1$ to $d = 2$ clauses.

and their central role in structuring discourse (Foley & Van Valin 1984; Van Valin 2005; Dalrymple & Nikolaeva 2011).

But where subject anaphors pattern in remarkably similar ways across corpora, object anaphors exhibit substantially more cross-corpus variation, and so the generalizations made above are much more tenuous. One possible explanation for the variability of objects is that they, unlike the chiefly human and narratively central subjects, are more strongly dependent on the actual content of discourse, and hence the number and nature of the referents mentioned in object position; this idea is developed further in Section 9.2.4 below. Lastly, it should be mentioned that the present study has only tested subjects and objects, but it is probable that similar considerations hold for mentions in other positions (such as oblique arguments) as well.

9.2 | Explaining the lexical baseline rate

9.2.1 | Clause chains and topic shifts

In Chapter 5 we observed that the overall rates of lexical expression are suprisingly stable across the corpora in the sample, which is true especially for (anaphoric, third-person) subjects, for which the ten corpora all cluster around a baseline rate of about 20% lexical expressions. As noted there, this finding runs counter to the conclusions drawn in Stoll & Bickel (2009), who note that there are substantial differences in the frequency and modalities of use of lexical expressions in Russian and Belhare, which has implications for cross-linguistic variation overall.

As the predictive model shows, we find that for third-person anaphoric subjects, all languages select lexical expressions under essentially the same circumstances, with only a small degree of cross-linguistic variability, in particular at intermediate distances to the antecedent. Since the overall rates of lexical expression in subject position are also essentially the same across languages, this means that the circumstances that condition lexical subjects come up at the same rate across languages. As noted above, lexical expressions are selected mostly in contexts of low discourse coherence and at longer distances to the antecedent. That is, where lexical expressions tend to not find use is for mentions of referents with an active role in the current stretch of discourse; compare this to pronouns, which, as seen above, are an option in all contexts,

including for distantly mentioned referents. The lowest lexicality rates are found at very short distances, specifically in same-role clause chains, where discourse is at its most cohesive (Givón 2017).

It is this context of same-subject chains that I would argue is the most central mechanism of discourse structure (cf. Chafe 1976; Givón 1983a, 2017; Huang 2000a; etc.), and hence at the core of an explanation of the observed usage rates of referring expression types. Discourse in essence consists of sequences of more or less connected clauses, and is tied together by the predictability of repeated reference (Nariyama 2003: 45), with the subject functioning as a common pragmatic pivot (in terms of Foley & Van Valin 1984: 108–134). In this view, subjects are not just one of the roles in an individual clause (as implied by the standard definitions of subjecthood), but in actual usage a vector of continuity and coherence; this is the idea behind the notions of the privileged syntactic argument and the aforementioned pivots in functionalist literature (Van Valin 2005). The data presented in the present study, as discussed in the preceding sections, reinforces this notion, and further highlights the importance of usage and discourse-based approaches to grammar.

Numerous studies have observed the special role of same-clause chains across languages, noting further that this is also the context in which zero expression is the most common, even in languages that usually exhibit a marked dispreference for zero, such as English (e.g. Owens et al. 2013: 263 for Arabic; Haeri 1989 for Persian; Meyerhoff 2009 for Bislama and Tamambo; Cameron 1994: 32; Silva-Corvalán 2001: 154; Torres Cacoullos & Travis 2019: 674 for Spanish; and Harvie 1998: 21; Leroux & Jarmasz 2005: 7; Torres Cacoullos & Travis 2014: 24, 2019: 674; Travis & Lindstrom 2016: 112; Schiborr 2017: 51–52 for English; as well as McKee et al. 2011: 388 for Australian Sign Language; Schnell & Barth 2020 for Vera'a; and many more). Shifts to a different referent conversely tend to favour more informative forms. As such, lexical subjects are dispreferred under conditions of topic continuity, and only become likely where topic continuity is interrupted and the topic switches to another, potentially less accessible referent.

Given the centrality of the clause chain context, we can hence define referent accessibility strictly in terms of co-referentiality with the subject of the preceding clause. If a subject anaphor occurs in such a context (i.e. at a distance of $d = 1$ clauses from its subject antecedent), it is maximally accessible; if not, its accessibility is a function of anaphoric distance (i.e. $d \geq 2$ clauses), attenuated by the salience of the referent (humanness, narrative centrality) and

a number of other, contextual factors. Crucially, this allows us to use clause chain status as a quantifiable operationalization of topicality (cf. Chafe 1976), and makes breaks in chains correspond to shifts in topic (cf. switch reference marking, Jacobsen 1967: 268; Falk 2006: 66–71; Hale 1992; etc.).

As such, the selection of lexical and non-lexical expressions – for subjects in particular – is primarily a matter of a mention's position in clause chains, and the cross-linguistically stable lexical baseline rate in large part comes down to the frequency of chain breaks and topic shifts. There are additional forces at work here as well of course, such as the use of lexical expressions at episodic boundaries and other special circumstances, which will be discussed further below, but topic chains and anaphoric distance account for the vast majority of cases.

We have already provided a working definition of clause chaining in Section 6.7.4, but it is worthwhile to expand the associated terminology going forward: Any two immediately adjacent clauses with co-referential subjects form a chain. For the figures shown above and below, this also counts syntactically embedded clauses. A clause chain consists of at least two linked clauses, and can reach theoretically infinite length. If the reference of the subject changes, that is, if it is not co-referential with the subject of previous clause, this constitutes a break in the chain, and the potential start of a new chain. In the following, the first co-referential subject in a chain will be referred to as a "chain starter", and subsequent co-referential mentions as "chain links". If a subjects breaks a chain but does not initialize a new one because reference immediately switches to another referent in the following clause (i.e. forms a chain of length $l = 1$), it is referred to as an "unchained mention", or simply a "one-off".[7] Note that this definition of clause chaining differs from the more nuanced ones in Givón (2017) and related work on discourse coherence, in that it takes an extremely simplified structural view, aiming to be as readily operationalizable and quantifiable as possible, given adequately annotated data. See also Torres Cacoullos & Travis (2019: 672–673) for further semantic refinements in defining same-subject chain contexts.

Subject mentions can hence be divided into three groups: those that initialize a chain of co-referential subject mentions (starters), those that continue

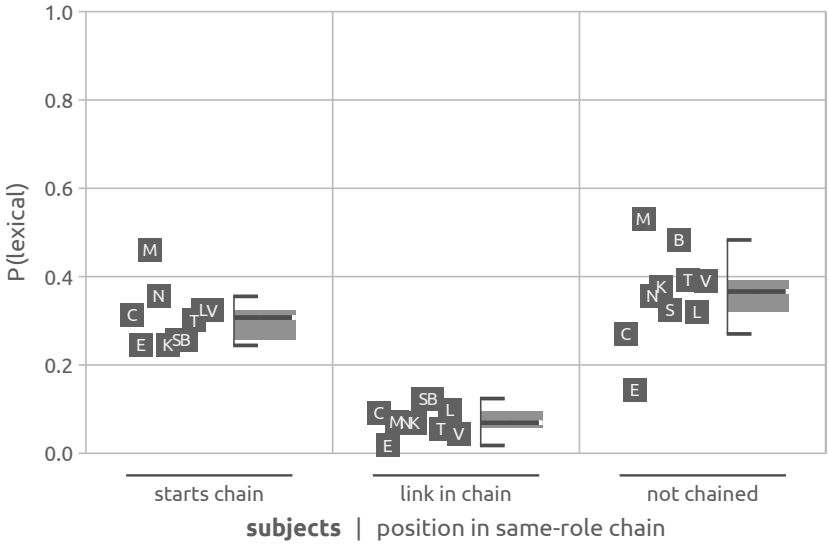
7 Note that this does not imply that mentions of these referents do not recur, only that they do not occur in a sequence.

the chain (links), and those that do neither (one-offs). Figure 9.1 shows the proportion of lexical expressions in each of these groups across the ten corpora; these figures recontextualize those shown earlier in Figure 6.24 in Section 6.7.4. They confirm that chain links – being highly accessible and topical, with their antecedent in the same position in the previous clause – are expectedly very rarely lexical in all corpora in the sample (cross-corpus mean $P = 0.08$, $\sigma = 0.034$), though notably also not categorically non-lexical (see Section 9.1.5 above). Since the overwhelming majority of chain links is non-lexical, most anaphoric lexical subjects consequently occur when a clause chain is broken. Lexical subjects are hence above all associated with shifts from one referent and topic to another, and the rates of lexical subject expression observed in Chapter 5 are as such a direct consequence of the rate at which clause chains are broken. We will discuss possible explanations for the cross-linguistically stable frequency of chain breaks and topic shifts futher below in Section 9.2.3, but first take a detour to look closer at the properties of chain starters and one-off mentions, and what differences exist between them.

9.2.2 | Properties of chain breaks

As seen in Figure 9.1, subjects that start clause chains and subjects that are unchained mentions are about equally likely to be lexical from a broad cross-corpus perspective, with the latter having a somewhat higher rate of lexical expression overall (cross-corpus mean $P = 0.36$, $\sigma = 0.107$) than the former ($P = 0.31$, $\sigma = 0.066$) in all corpora except English, where they are instead much less likely to be so. Other notable outliers include Mandarin, where all mentions that break chains have a comparatively higher rate of lexicality, a pattern that we already noted in Chapter 5, and Tabasaran, which shows the largest difference in lexicality rates between chain starters and unchained mentions. In general, the greatest degree of cross-corpus variation is found among unchained subjects, with links and chain starters being comparatively stable.

But overall – ignoring English for now – there are no substantial differences between those subjects that continue on as a topic and those that are immediately switched out again and hence do not persist beyond a single mention. This suggests that whether a topic shift continues as a chain or not has little bearing on its lexicality. Taken together with the fact that the majority of chain breaks are still non-lexical (and that some chain links are lexical),



	corpus	starts chain N(lex)	starts chain N(all)	starts chain P(lex)	link in chain N(lex)	link in chain N(all)	link in chain P(lex)	not chained N(lex)	not chained N(all)	not chained P(lex)
C	C. Greek	27	86	0.31	19	207	0.09	40	148	0.27
E	English	60	244	0.25	8	452	0.02	95	659	0.14
M	Mandarin	70	152	0.46	25	354	0.07	113	213	0.53
N	Nafsan	54	152	0.36	25	369	0.07	63	177	0.36
K	N. Kurdish	32	131	0.24	19	281	0.07	87	231	0.38
S	S. Dargwa	27	105	0.26	29	234	0.12	60	185	0.32
B	Tabasaran	43	167	0.26	47	383	0.12	114	236	0.48
T	Teop	43	143	0.30	19	341	0.06	105	267	0.39
L	Tulil	37	114	0.32	21	216	0.10	93	290	0.32
V	Vera'a	151	467	0.32	53	1225	0.04	288	738	0.39
	totals	544	1761	—	265	4062	—	1058	3144	—

Figure 9.1 | Lexicality of anaphoric subjects by position in same-role chains, across corpora.

this suggests that lexical expressions are not selected specifically for topic shifts, and that there is hence no structural association between topic shifts and lexicality.

Rather, anaphoric subjects are only lexical if the referent is not accessible enough for a non-lexical form, that is, if it has not been mentioned recently enough. As such, while lexical expressions are very likely to indicate chain breaks, chain breaks are not necessarily lexical; at the same time, lexical NPs are avoided for chain links since the referent is already maximally predictable and hence accessible, and so does not require an informative expression to aid identification; selection of an overly informative expression might in fact be misleading. The selection of lexical expressions for subject anaphors hence does not appear to have a specific discourse-structuring function, but is simply a consequence of limitations on memory capacity and considerations of referent accessibility, which might also explain the strong cross-linguistic regularities found in the data. The same is not true for objects – or at least not to the same extent – which we will get to further below in Section 9.2.4. What small difference in lexicality rates exists between starters and one-offs can in part be explained by the composition of the two groups of subjects, in particular by the mean distance to their antecedents, their humanness and narrative centrality, and the transitivity of the clauses they are situated in.

While chain links by definition occur at a distance of $d = 1$ clause from the antecedent, other mentions can have any distance to antecedent. The mean anaphoric distance for chain starters is $d = 3.50$ clauses ($\sigma = 0.523$), for one-offs $d = 3.83$ clauses ($\sigma = 0.523$), meaning that while the latter tend to be slightly less recent, the difference is negligible; as such, there is essentially no difference between starters and one-offs in terms of recency. For both, the association between anaphoric distance and lexicality is the same (cross-corpus mean Spearman's rank correlation coefficient $\rho = 0.80$, $\sigma = 0.082$ for starters vs. $\rho = 0.81$, $\sigma = 0.078$ for one-offs).

We noted in Figure 6.23 in Section 6.7.4 that clause chains are about three clauses long on average (cross-corpus mean $l = 2.92$ clauses, $\sigma = 0.238$; i.e. a chain starter and two links). Since the average distance to the antecedent of both chains starters and one-offs is roughly the same, this means that referents that break a clause chain are likely to have been the subject of the previous chain, or at the very least have been mentioned in it. In fact, a little over half of chain starters were subjects on their previous mention ($P = 0.52$, $\sigma = 0.137$ subject antecedents and $P = 0.14$, $\sigma = 0.053$ object antecedents), compared to

a little less than half of one-offs ($P = 0.45$, $\sigma = 0.120$ and $P = 0.19$, $\sigma = 0.053$). While these are mostly very similar profiles, chain starters are slightly more likely to have last been mentioned in a more prominent position, which might explain their slightly lower rate of lexical expression, given the association between lexicality and antecedent role noted ealier. Notably, about a third of both starters and one-offs was last mentioned in neither subject nor object position. These figures also reveal that the strong rates of role persistence from mention to mention for subjects, as seen in Section 6.7), are mostly driven by links in same-role chains, as at the start of chains, there is a roughly 50% chance that referent was not a subject on its previous mention.

Subjects are predominantly human (see Section 6.1), and this is especially true for those occurring in clause chains ($P = 0.90$, $\sigma = 0.090$). The latter are also more likely to be protagonists ($P = 0.74$, $\sigma = 0.086$), and hence tend to have both higher total and local mention frequencies. In fact, humanness and protagonisthood are where chained and unchained mentions differ the most, with the latter being both less likely to be human ($P = 0.78$, $\sigma = 0.123$) and narratively central ($P = 0.59$, $\sigma = 0.093$) than the former by an appreciable margin. The nature of the referent hence influences its likelihood of persisting as a topic over multiple clauses; this follows from the association between topicality and humanness already discussed above (Section 9.1.4). While most referents that occur in chains tend to be highly frequent and narratively central referents, unchained mentions are considerably less likely to be so, instead being more incidental, involving referents whose perspective is secondary and who contribute less to narrative advancement.

As concerns lexicality, while non-human and non-protagonist referents are more likely to receive lexical expression in general, this is especially true for unchained mentions ($P = 0.61$, $\sigma = 0.196$ for one-offs vs. $P = 0.42$, $\sigma = 0.262$ for starters). Notably, this also applies to chain links ($P = 0.24$, $\sigma = 0.173$ vs. $P = 0.08$, $\sigma = 0.034$ overall), though as noted earlier, this type of referent is quite rare in chain-medial position. This suggests that for the least prototypical kinds of topics (non-human, not narratively central), mentions in subject chains do not carry the same overwhelming level of accessibility, likely due to relative unpredictability of this configuration.

To summarize, there appears to be no inherent structural association between mentions that start clause chains (or not) and specific referential choices; rather, what differences in lexicality rates exist between these two groups of subjects follow chiefly from which kinds of referents are liable to

start (and continue) clause chains in the first place, and which are not: human, narratively central referents, especially those that already had topic status in the recent discourse, which are generally more likely to receive a favoured position in speakers' memory, and hence be realized non-lexically. Where marked variation in referential choices occurs is, unsurprisingly, between chain-initial and chain-medial mentions. As such, whether a subject is realized lexically or non-lexically is primarily determined by whether the referent in question can be identified without the need for additional cues. Chain links are inherently maximally accessible, while the accessibility of chain breaks results from how recently they were last mentioned and how inherently salient the referent is; how the discourse continues is evidently irrelevant.

What about the English data then, which goes against the tendencies laid out by the other corpora in the sample? Figure 9.1 shows that the lexicality of chain starters and links in English tends towards lower end of the cross-corpus distribution, but it is one-offs that are the most notable outlier, as English is the only corpus in the sample in which unchained mentions are less likely to be lexical than chain starters. This means that the lower overall lexicality rate for subjects in English noted in Chapter 5 and elsewhere is the consequence of its low lexicality rate among one-offs. This difference is both a matter of how English treats incidental perspective shifts, and of what kind of referents come up in these shifts in the texts. For one, English has the shortest average chain length in the sample ($l = 2.48$, $\sigma = 0.885$), shifting topics somewhat more frequently than most other corpora. This means chain breaks tend be more recent overall and hence more accessible, and that one-offs are cycled more rapidly. For another, while one-offs are themselves not more likely to be non-human in English than in other corpora, we have already commented on the overall lower proportion of human subjects in the English data overall ($P = 0.63$ vs. the cross-corpus mean of $P = 0.85$, $\sigma = 0.109$).

The insensitivity of English to humanness distinctions, already discussed above, also appears to play a role: In other corpora, humanness has a substantial effect on lexicality among unchained mentions ($P = 0.30$, $\sigma = 0.101$ for humans vs. $P = 0.60$, $\sigma = 0.203$ for non-humans), but no such effect exists in English ($P = 0.14$ vs. $P = 0.15$). Among chain starters, however, human referents are about twice as likely to be realized lexically than this is the case for non-human referents ($P = 0.29$ vs. $P = 0.14$) – a reversal of the expected pattern. While the subsample for the latter is unsurprisingly quite small, raising the likelihood of these observations being only a statistical artefact, a similar

pattern is also found in the Mandarin data ($P = 0.47$ for humans vs. $P = 0.36$ for non-humans).

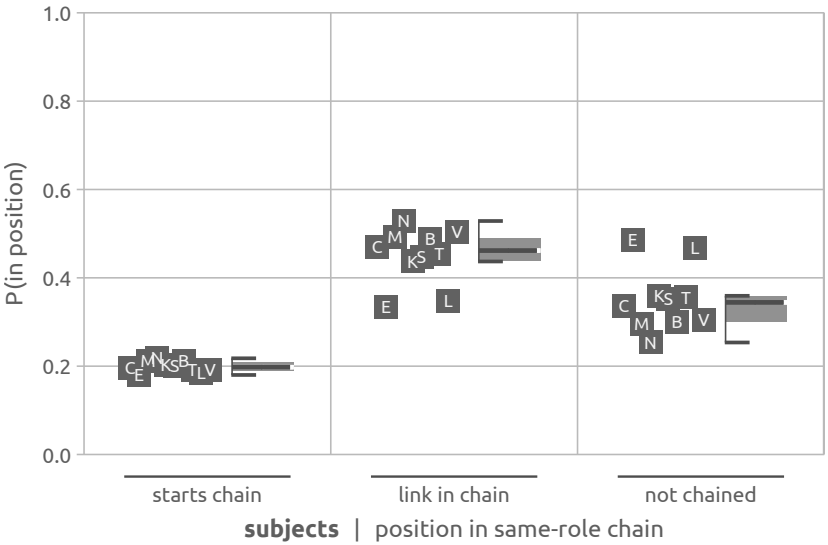
9.2.3 | Accounting for the frequency of topic shifts

As discussed earlier, lexical subjects occur mostly where topics shift from one referent to another and clause chains are broken. Chain-medial mentions, that is chain links, are conversely very consistently non-lexical, owing to their recency and high predictability. An explanation of the cross-linguistically stable lexical baseline rate for subjects noted in Chapter 5 hence has to account above all for the rate of topic shifts in natural discourse.

We have already touched on the topic of average chain lengths in Section 6.7 and in the previous section; chains are of roughly the same length across corpora, with only two minor outliers, English and Tulil, in which clause chains are on average about half a clause shorter. The average length of clause chains is tied to the relative proportions of the three types of subjects – chain starters, chain links, and one-off mentions – which are shown for the ten corpora in Figure 9.2. This is a perspective on the data which we have neglected in the discussion of lexicality rates above, but provides a crucial piece to the puzzle.

Note, first of all, that in most corpora the majority anaphoric subjects occur in same-role chain contexts, either initiating or continuing chains (cross-corpus mean $P = 0.45$, $\sigma = 0.064$), as already mentioned earlier. Unchained mentions account for less than half of subject mentions; in light of Figure 9.1, this means they make up the majority of lexical subjects in most corpora. Since the average length of clause chains is quite similar across corpora, the proportion of links accordingly patterns similarly as well, with a couple of notable outliers.

But perhaps the most striking observation here is that the proportion of chain starters relative to chain links and one-offs is almost exactly the same across all corpora, at around a fifth of the total (cross-corpus mean $P = 0.20$, $\sigma = 0.013$). In other words, this means that on average, a stretch of discourse of ten clauses (with anaphoric subjects) will contain about two same-subject chains of variable length, in addition to a somewhat variable number of unchained subject anaphors. While this disregards the presence of referent introductions, as noted in Chapter 5, these do not make up a substantial proportion of subject mentions. This consistent rate suggests that narrative discourse



	corpus	starts chain N(pos)	starts chain P(pos)	link in chain N(pos)	link in chain P(pos)	not chained N(pos)	not chained P(pos)	N(all)
C	C. Greek	86	0.20	207	0.47	148	0.34	441
E	English	244	0.18	452	0.33	659	0.49	1355
M	Mandarin	152	0.21	354	0.49	213	0.30	719
N	Nafsan	152	0.22	369	0.53	177	0.25	698
K	N. Kurdish	131	0.20	281	0.44	231	0.36	643
S	S. Dargwa	105	0.20	234	0.45	185	0.35	524
B	Tabasaran	167	0.21	383	0.49	236	0.30	786
T	Teop	143	0.19	341	0.45	267	0.36	751
L	Tulil	114	0.18	216	0.35	290	0.47	620
V	Vera'a	467	0.19	1225	0.50	738	0.30	2430
	totals	1761	—	4062	—	3144	—	8967

Figure 9.2 | Proportions of anaphoric subjects in specific positions of same-role chains, across corpora.

is structured along very similar lines across languages (at least as regards third-person anaphoric references), especially when taken together with the apparent universality of the key form selection criteria (anaphoric distance, antecedent role) discussed above. This confirms our initial hypothesis that, contra Stoll & Bickel (2009), cross-linguistic variation in referential choices is largely limited to variation in the proportions of zero and pronominal anaphors.

That being said, there is still a degree of cross-corpus variation evident in the data; in particular, the English and Tulil data share the distinction of having comparatively lower proportions of chain links, and correspondingly higher proportions of unchained mentions. These differences are tied to the average chain length in these corpora, which is accordingly slightly lower compared to the cross-corpus mean ($l = 2.48$ in English, $l = 2.53$ in Tulil). It is, however, not immediately clear what the cause of this variation could be; it might potentially be related to differences in narrative structure or content and the types of referents involved.

Of course, all this just shift the onus of the explanation from lexicality rates to rates of chain breaks; so how come it is that topics change at roughly this rate in all corpora in the sample? This is an open question, but there a number of potential avenues for explication.

The upper boundary of topic shift frequencies is likely bounded by considerations of narrative coherence. Switching topics too often would make narrative discourse difficult to follow, as it would lack the cohesiveness created by recurrent reference (cf. Foley & Van Valin 1984: 108–134; Givón 1983a, 2017; Kehler 2002, 2004; etc.). Conversely, switching topic too infrequently would result in a narrative that lacks in tellability (Labov & Waletzky 1967), in the sense of the Gricean maxim of quantity (Grice 1975). Narratives require multiple actors to be tellable; a narrative consisting of nothing but an unbroken sequence of actions by the same entity would not be worth telling.

As an actor in a narrative encounters and interacts with other entities, discourse may shift perspective to put their appearance and responses in focus, though as noted above, in many cases the perspective shifts immediately back (or away) again. This aligns with clauses that break chains being more likely to be intransitive (see Section 9.3.1); rather than directly advancing the narrative, they may instead serve to frame changes in perspective.

In this view, the frequency of topic shifts is not exclusively bounded by considerations of information structure, but also by more general social,

phatic, and communicative functions of language. Discourse is hence subject to a set of global pragmatic constraints in a Gricean sense, so that communicatively successful narratives imply certain quantitative limits, both upper and lower, on the number and realization of discourse referents (Haig et al. 2021). Speakers hence strike a balance between topic shifts and longer sequences of clauses involving same topic.

This relates to the notion of the uniform information density in language production (UID, Jaeger 2006; Levy & Jaeger 2007). In general terms, the UID hypothesis states that speakers aim to smooth the information density distribution of their utterances over time to achieve optimal communication.[8] In this view, the structure of discourse needs to be sufficiently differentiated to be informative and tellable, but not so much that the necessary level of redundancy encoded in it becomes too low. Distinctiveness and frequency are hence inversely correlated (Meylan 2018: 105; cf. also Zipf's law of abbreviation and its refinements, e.g. Piantadosi et al. 2011; Baayen et al. 2016). This has parallels in, for instance, the size of vowel and consonant inventories and the distribution of vowels in vowel space, which are approximately normally distributed in the languages of world (Gordon 2016).

Speakers hence try to keep the interest curve of their discourse stable; this regularity leads to a degree of routinization, so that interlocutors expect a clause chain to only go on for so long before the topic switches. A discourse containing only a single topic would hence have an information density that is too low, running into issues with tellability; conversely, a discourse with too frequent topic switches in too close succession would have too high a density, and as a result not be predictable enough. Of course, speakers do not mechanically break chains every three clauses; there may be local extrema in the rate of topic shifts – the longest chain in the sample contains 21 links, and there are passages with few if any chains – but over longer stretches of discourse, the rates will even out.

The same principle also affects referential choices directly (cf. Levy & Jaeger 2007; Jaeger 2010; Orita et al. 2015). If interloctors expect a topic chain to continue, they will consequently predict that the next subject will be

8 Compare also the entropy rate constancy principle (Genzel & Charniak 2002; cf. also Levshina 2019 for a similar notion of entropy applied to word order variation) and the smooth signal redundancy principle (Aylett & Turk 2004), which has been applied to phonetics.

co-referential with the previous if it is reduced in form; if the next subject is lexical, it will be assumed to be non-co-referential and hence constitute a shift in topic. This is the reason for the routinized association between topicality and reduced forms, which constrains the availability of lexical forms in high coherence contexts and makes lexical overcoding lead to ambiguity.

What is remarkable, then, is that the balance between switching and continuing topics appears to be as stable across languages as the data from Multi-CAST – which are hardly homogenous – appear to suggest. This indicates that what constitutes an appropriate information density in narration (and hence tellable discourse) appears to be very similar cross-linguistically, which in turn suggests that discourse structure is guided in large part by universal cognitive mechanisms.

What this does not account for, however, are differences in lexicality rates (and hence rates of topic shifts) between text types. To briefly recapitulate Section 5.2, a cursory review of studies on referential choice and anaphora suggests that subject lexicality rates in conversational texts (e.g. Kärkkäinen 1996; Francis et al. 1999) are substantially lower than those in the monologic narratives examined here and in various studies of Pear story retellings (e.g. Givón 1990; Ashby & Bentivoglio 1993; Payne 1993; Himmelmann 1997; Arnold 2003; Kumagai 2006; Haig & Schnell 2016); written texts conversely exhibit a well-known tendency to be highly explicit (e.g. Prince 1992; cf. also Fox 1987b; Toole 1996). It should be kept in mind, however, that evidence is rather sparse and largely limited to English and other highly-studied languages, and that cross-study comparability in general is quite poor.

There are multiple candidates for variables that can potentially differ systematically between text types – the accessibility thresholds for the use of lexical expressions may vary, or else the rates of topic shifts and lengths of clause chains – but lacking sufficient data, all we can do at this point is speculate. Whatever it may be, it is likely that any variation between text types results from high-level rather than low-level differences, that is from matters of discourse organization and the routinized expectations tied to genres rather than the underlying mechanisms of discourse coherence and form selection (cf. Travis & Lindstrom 2016). We might further suspect that for spoken discourse, differences will most clearly align between monologues and dialogues/multilogues (Ozerov 2018); see also Shor (2020) for a discussion of monologue-oriented and conversation-oriented approaches to referential choice.

9.2.4 | Object form is content-dependent

While, as noted above in Section 9.1.6, subject and objects are expressed via lexical expressions under similar circumstances, for the latter, the lexicality rates show much greater variation both across corpora and individual texts (see Section 5.4). Consequently, no strong central tendency for lexicality can be identified for them, as it can for subjects.

This in turn means we cannot appeal to the same universal aspects of discourse structure to explain referential choices in object position. The object role is not as strongly associated with topicality, and hence has a much lesser role in establishing and maintaining discourse coherence; same-role chains are hence not only rarer for objects (cross-corpus mean $P = 0.26$, $\sigma = 0.046$ for objects vs. $P = 0.45$, $\sigma = 0.064$ for subjects), but also do less to tie sequences of clauses together. While the proportion of objects in same-role chains is very similar across the ten corpora (see Section 6.7.4), it is not nearly as clear-cut in terms of lexicality. The considerations that very clearly delineate subject expression inside and outside same-role chains hence do to apply to the same extent for mentions in object position.

Instead, the rate at which objects are lexical is more strongly dependent on the properties of the referent on the one hand, and more generally the composition of the referents that come up in a text on the other. Humanness and protagonisthood shift accessibility thresholds upwards or downwards in both roles, but unlike subjects, objects are neither predominantly human nor narratively central, so that differences in the inherent salience of referents exert greater influence. The proportion of human objects, for instance, ranges between 12% in the English and 49% the Teop corpora (see Section 6.1). As such, where a lot of the regularity in subject forms is due to the relative homogeneity of the referents mentioned in subject position, objects appear to pattern more randomly, as their referents exhibit far greater diversity. Likewise, while the majority of subject mentions are instantiations of the same few referents ($N = 7.98$, $\sigma = 2.578$ mentions per unique referent), the number and frequency of recurrence of the referents mentioned in object positions may vary substantially across texts ($N = 2.94$, $\sigma = 0.717$ mentions). What determines the lexicality profile of objects in a corpus (or text) is hence to a large degree a matter of which kinds of entities come up as patients and stimuli. Texts that predominantly involve humans interacting with other humans will pattern differently from texts in which the majority of entities interacted

with are inanimate. Similarly, a text with only small number of more or less persistent referents that recur multiple times in objects position will pattern differently from a text with numerous, more incidental referents that are encountered, interacted with, and then immediately discarded again.

In sum, subject lexicality is homogeneous largely because of a strong association with topicality and humanness. As the semantics of objecthood in general lead to fewer restrictions, object lexicality is conversely more heterogeneous and is hence more strongly influenced by matters of narrative content.

9.3 | Further observations

9.3.1 | Topic shifts and transitivity

There is one more dimension that is worth investigating in the context of clause chains, which is the transitivity of clauses that break chains vis-à-vis those that continue them. Overall, clauses that break chains (and possibly start a new one) are slightly more likely to be intransitive than clauses that continue an ongoing chain (cf. Hopper & Thompson 1980), with chain-medial clauses being on average 55% intransitive ($\sigma = 6\%$), chain-starting clauses 65% ($\sigma = 9\%$), and one-off clauses 67% ($\sigma = 8\%$). Notably, there is very little difference between one-offs and starters, so that whether or not a new topic continues forward (or indeed is human or a protagonist) has no bearing on how the previous chain was broken.

This (small) association between chain breaks and intransitivity in essence parallels the claims of the principle of the separation of role and reference (Lambrecht 1994) and preferred argument structure (Du Bois 1987b, 2003b, 2017; see Section 2.2.5.1 above) for the introduction of new information into discourse: In this view, the task of introducing new referents (or in this case, retrieving non-topical referents) is cognitively demanding, and speakers hence employ intransitive clauses to flatten the information density curve of their utterances. However, the observed difference is not nearly as large as might be expected from a fundamental functional association of this kind.

An alternative view might instead attribute the slightly higher rate of intransitive predicates in chain breaks not to mechanisms of information structure, but the needs of narrative structure (cf. Schnell et al. 2021b). In this view, before a new topic can be involved in a transitive event, it may have

to be made to arrive or appear first via an intransitive one. Speakers hence may use intransitive predicates to focus on the change in referent, rather than immediately combine said topic shift with a transitive action.

One could also argue in favour of an inverse interpretation of the association between transitivity and chain status, so that it is not that chain breaks are more likely to be intransitive, but that chain-medial clauses are inherently more conducive to the expression of transitive events. This would in turn imply that most of the heavy lifting in advancing a narrative is done within clause chains, which also links up with unchained mentions being more frequently non-human and non-protagonists, as mentioned above.

9.3.2 | Episodic breaks

As mentioned above, there are a number of specific circumstances under which the usual considerations of discourse coherence and recency effects do not apply, one of which involves shifts from one scene in a narrative to another, and the beginning of new narrative strands (called "world boundaries" in Kibrik 2011). These often also involve a temporal or spatial shift. Taking into account episodic structure is important especially in narrative discourse (Kibrik 2000: 78); it has been noted that the beginning of a new narrative episode is a borderline after which speakers tend to use lexical expressions even for recently mentioned referents (Marslen-Wilson et al. 1982; Tomlin 1987a; Fox 1986, 1987a; see also Anderson et al. 1983; Vonk et al. 1992). These boundaries can also split clause chains: There is small percentage of lexical subjects within clause chains in the data, some of which occur at boundaries between episodes; we noted an example of this in (86) in Section 7.2.5 above.

While the data used for the present study do not allow episodic boundaries to be identified and examined systematically, we can nevertheless forward a possible account of their effect on referential choices, as noted in the literature and anecdotally in the Multi-CAST data. First, it is likely that episodic breaks involve topic shifts in the sense discussed above, even when the new topic happens to be co-referential the one in the previous clause. Second, their association with lexical expression might then follow from a routinized reassessment of the accessibility of any active referents at the start of the new episode, in effect "wiping the slate clean" to a degree. As such, when speakers set up a new scene, old referents may not be carried over as is, but may need to be re-established, being too "distant" in narrative or conceptual space

because they were last mentioned in a different place, at a different time, or under different circumstances within the narrative.

In any case, a systematic examination will be necessary to confirm these musings; given the relative rarity of this phenomenon (compared to discourse anaphora in general), this would require substantial amounts of narrative data, ideally from narratives of considerable length, and specialized annotations that (in all likelihood manually) mark out the beginnings of new episodes.

9.3.3 | Types of lexical expressions

In Chapter 8, we examined whether three major types of lexical expressions – proper names (vs. common nouns), complex NPs with additional modifiers (vs. simple), and NPs with demonstrative modifiers specifically (vs. without) – are subject to the same mechanisms that determine the broader choice between lexical and non-lexical expressions (third person, subjects only). These are usually understood to be in essence special cases of the more general lexical NP category, varying in informativity and specificness of reference along a continuous cline: On the accessibility scale (Ariel 1990), for instance, all three types (among a number of others) are distinguished as distinct levels whose selection is determined by the same considerations of referent accessibility, which in turn is largely determined by the properties of the discourse and the salience of the referent.

The findings in Chapter 8, however tentative though they may be, indicate that, of the factors that manage to predict the lexicality of an anaphor with greater than chance accuracy, none conclusively associate with the choice of specific types of lexical expressions. This suggests that their selection is not based on local discourse structure in an obvious manner, if at all, contrary to the claims of accessibility theory and related views.

For proper names, their use is instead largely conditioned by whether the referent in question is narratively central enough to be identified by name in the first place; Haghighi & Klein (2010) come to similar conclusions, noting that their selection is "governed more by entity frequency than antecedent distance". There is consequently a large degree of variability between corpora and texts, which shows that the presence of proper name mentions (in subject position, at the very least) is highly dependent on the actual content of the discourse, and is therefore best investigated on the basis of data sets that

are either controlled for content to ensure their use, or are at the very least substantially larger.

The distinction between simple (e.g. *the book*) and complex (e.g. *the small book with the green cover*) is perhaps the one with the clearest correlation with informativity. We would consequently expect that heavy expressions would be selected by the much same considerations as lexical expressions in general, only under more extreme circumstances of inaccessibility. Givón (1995), for instance, notes that speakers chiefly use simple lexical NPs at distances below $d < 7$ clauses, and more complex expressions at $d \geq 10$ clauses and above; the relative arrangement of simple and complex expressions on the accessibility scale suggests similarly (Ariel 1990: 73, ex. 1). In the Multi-CAST data, however, while more complex (and informative) lexical expressions are associated with higher anaphoric distances and related factors, the effect only has marginal strength.

Lastly, for NPs with demonstrative modifiers (e.g. *this book*) the model results in Section 8.3 suggest that modification by a demonstrative is not influenced by any of the tested factors in an appreciable fashion, neither by the properties of the local discourse (anaphoric distance, etc.) nor by the inherent salience of the referent (humanness, protagonisthood). This echoes corresponding findings in Schiborr (2017), but, as mentioned above, contradicts accessibility theory and the accessibility scale (Ariel 1990: 73), which claim a nuanced association between (various kinds of) demonstratives and accessibility. While Ariel (1990: 53) does note that the differences are not expected to be substantial in English specifically, we here find that discourse-related factors appear to play no role whatsoever across texts from multiple languages, in addition to English. As such, it is likely that speakers make use of demonstrative NPs for their deictic or stance-taking function, rather than their informativity vis-à-vis other lexical expressions.

In sum, the exact circumstances of the selection of the three types of lexical expression examined here cannot be predicted from factors tested here. Ultimately, this suggests that, as far as lexical expressions are concerned, referential choices in discourse are less attuned to the specific degree of informativity of an expression than expected by many referential hierarchies, and instead appear to primarily revolve around the broad choice between lexical and non-lexical forms (Kibrik 2011; Kibrik et al. 2016). We will continue this line of thought in the next section.

9.3.4 | Referential choice is about lexicality first

One of the key findings of this study is that rates of lexical expressions are similar across languages, especially for third-person subject anaphors. As discussed above, this is in part explained by rates of topic shifts in narrative discourse following cross-linguistically stable patterns, as lexical NPs are selected under the same circumstances.

Among non-lexical expressions, conversely, matters are far more varied (see Chapter 5), with the choice between zero between various pronominal forms being a well-known typological variable. Notably, a language's preference for either reduced form does not affect its rate of lexical expression, and as briefly noted above, zero anaphors are not highly sensitive to anaphoric distance, the main driver of the lexical–non-lexical choice. Among lexical expressions, major distinctions among form types (proper names, phrasal complexity, etc.) do not align significantly with the properties of local discourse (see Chapter 8), and are as such not made on the basis of the same considerations that guide the selection of lexical expressions in general. The use of proper names in particular appears to be more a matter of the identity of the referent than discourse structure. Moreover, the single most common type of lexical expression are simple, common lexical NPs.

In sum, these observations suggest that models of referential choice that place zero, pronouns, and various types of lexical NPs on a single scale (e.g. Givón 1983a; Ariel 1990) conflate distinct referential categories and misrepresent the modalities of their selection by making it dependent on a single variable (topicality/givenness, accessibility, etc.). A more appropriate hierarchization of referential choice that is applicable across languages might instead work in multiple (i.e. at least two) stages, with different sets of factors operating at different stages of the process. In this view, the first and most fundamental (and in all likelihood typologically universal, given the preliminary evidence examined in the present study) choice is between lexical and non-lexical expressions; this in essence echoes the notion of basic referential choice in Kibrik (2011: 39–42). As discussed above, this choice is based on considerations of discourse coherence and referent accessibility, and as such sensitive above all to recency and related factors, and to constraints on memory reflecting on local discourse structure. That being said, as discussed above in the context of the English data in the sample, there are also differences between how languages pattern, for instance regarding the influence of

particular factors such as humanness. Furthermore, if pronouns are relatively informative in a language, this can result in a higher share of pronouns in contexts where lexical expressions would be expected, though as noted earlier, reduced expressions (chiefly, but not exclusively, pronouns) are available in all contexts anyway.

Once the choice between lexical and non-lexical forms has been made, speakers then decide on a specific kind of expression in the second stage. If the referent to be expressed is accessible enough for a reduced form, the choice is between zero or a pronoun; if not, speakers decide on whichever type of lexical expression is most appropriate for the present circumstances. Discourse coherence and accessibility likely still play a role at this stage, for instance in the selection of zero in same-role chains, but are not the sole motivators of referential choices; instead, semantic and other factors come to the fore here, as do language-specific constraints (see, e.g., Schnell & Barth 2018; Schwenter 2016, 2014 on zero objects) as well as co-occuring agreement and interactions with other morphological material. Of course, this two-stage model a only explanatory abstraction, and does not imply that this is how referential choices are necessarily made on a cognitive level.

In any case, this is not the right place to further develop this idea, as doing so would far exceed the scope of this study. It will require dedicated attention in future work, alongside the other ideas proposed here; see Section 10.2 further below.

10 | Conclusions

This final section provides a summary of the study, its main findings, and conclusions (Section 10.1) and an outlook on possible future avenues of research (Section 10.2).

10.1 | Summary of the main findings

This study has examined two aspects of referential choice in natural discourse: the overall proportions of various referring expressions, in particular of lexical NPs, and the specific circumstances under which lexical expressions are selected for anaphors in subject and object position. This examination has focused on mentions in the third person (i.e. those for which lexical NPs are available in free variation) and on anaphoric references (i.e. no introductions) specifically, and as such only touches on a limited subset of all cases of referential choice; the specific sample selection criteria were outlined in Section 4.1 above. The analyses draw from spoken data from ten languages, the but the size and typological representativity of the data are limited, so it is important to stress the tentative nature of these following conclusions.

Chapters 6 and 7 and Section 9.1 explored one mechanism of referential choice, the selection of lexical expressions, on the basis of an array of twelve factors, including various properties of the local discourse including anaphoric distance, recent mention frequency, and the role of the antecedent, as well as the inherent properties of the referent such as humanness and narrat-

ive centrality. These factors were calculated programmatically on the basis of comparatively simple annotations applied uniformly to the corpus data (Section 3.3).

In Chapter 6, we examined the influence of each of these factors individually and further noted certain key interactions between them. Using these insights as a foundation, Chapter 7 then presented multifactorial analyses of the data using methods from the machine learning toolkit: classification trees (Breiman et al. 1984) and two comprehensive predictive models using the gradient boosting algorithm (Friedman et al. 2000; Friedman 2001, 2002), one for mentions in subject position, and another for mentions in object position for comparison.

Overall, the models suggest that the selection of lexical expressions is conditioned by more or less the same factors in all corpora in the sample (cf. Torres Cacoullos & Travis 2019), and can hence be predicted with better than chance accuracy from considerations of local discourse structure and referent salience, though it is anaphoric distance that does most of the heavy lifting. The primacy of recency effects is especially noteworthy, as it appears to far exceed the already central influence given to it in the literature (Clark & Sengul 1979; Givón 1983a; Ariel 1990; Kibrik 2011; Kibrik et al. 2016; Arnold 2010; etc.). Very short distances and longer distances in particular converge with certain other factors into configurations of extreme accessibility or inaccessibility, the most important of these being same-role chains (cf. Givón 1983a, 2017). Many of the tested factors, however, are at their most distinctive at intermediate distances. Referent-inherent properties such as humanness apply everywhere, but are less predictive in same-role chains and at very short distances to the antecedent, not being as distinctive there due to the prevalence of human referents in these contexts.

The data also suggest that the effects of certain factors may be parametricized across languages, particularly as regards their sensitivity to humanness distinctions, which are not reducible to considerations of accessibility. This parametricization results in different profiles especially at intermediate distances, as mentioned above. One hypothetical explanation for these patterns may attribute them to differences in the informativity of pronominal expressions and language-specific factor weightings, but further research is needed for a proper account.

As regards differences between roles, similar patterns obtain for subjects and objects, but the latter exhibit far greater variation between corpora. While

considerations of recency and related factors apply to both, the relative homogeneity of form selection in the subject role can attributed to its role as a pragmatic pivot (Foley & Van Valin 1984: 108–134), establishing and maintaining discourse coherence through the predictability of recurrent reference. Object form, conversely, is more content-dependent, being more strongly influenced by the properties and composition of the referents in a text. While the proportions of lexical expressions in any position depend to a certain degree on the content of a text, this is much less pronounced for subject anaphors, as they draw from a much smaller and much more homogenous pool of referents compared to objects. If anything, the variability among object anaphors highlights special status of the subject role.

While the core of this study has been focused on the broad choice between lexical and non-lexical expressions, in Chapter 8 and Section 9.3.3 we have also tested whether different types of lexical NP are selected on the basis of the same considerations. Specifically, we examined proper names, light and heavy expressions (i.e. NPs with additional semantic content beyond the head noun), and the presence or absence of demonstrative modifiers. For all three types, the selection of the marked type is not conclusively based on considerations of discourse structure or accessibility. The choice of proper names is instead mostly a matter of a referent's frequency and narrative centrality (cf. Haghighi & Klein 2010), phrasal complexity shows only a marginal association with anaphoric distance (which as noted above, is strongly predictive of lexicality overall), and the use of demonstrative modifiers is likely a matter of semantic content, since none of the discourse factors tested in this study bear any influence on it. As such, the selection of specific types of lexical expressions is not simply an extension of the basic lexical–non-lexical choice, as suggested by many of the more nuanced referential hierarchies (e.g. Givón 1983a; Ariel 1990; etc.).

Where Chapters 6–8 investigated the mechanisms of referential choices across languages, Chapter 5 and Section 9.2 addressed the complementary question of the proportions of different referring expressions. Where Stoll & Bickel (2009) conclude that there is substantial variation between languages (Russian and Belhare in their case) in this regard, we here found that for subject anaphors in particular, rates of lexical expressions in the ten corpora are remarkably similar, clustering more or less tightly around the 20% lexical (vs. reduced) mark. This observation led to the identification of a lexical baseline rate of spoken narrative discourse, for which a tentative explanation might be

that over long stretches, discourse tends towards an ideal level of explicitness of subject expression.

Notably, the rates of pronominal and zero anaphors vary substantially between corpora (cf. pro-drop, Perlmutter 1971; Bickel 2003; Neeleman & Szendröi 2007; Travis & Lindstrom 2016; etc.). On the basis of these observations, we proposed the foundations of a simple two-stage model of referential choice in Section 9.3.4, ultimately following the notion of a "basic referential choice" in (Kibrik 2011): This approach reduces speakers' choices among the range of available referential options – zero, pronouns, simple and complex lexical NPs, proper names, and so on – first down to a binary choice between lexical and non-lexical expressions, which operates on the principles outlined above; in a second step, this choice is then differentiated further – in idiosyncratic and language-specific ways – to yield specific types of lexical or non-lexical expressions.

In Section 9.2, we related the stable lexical baseline to the frequency of topic shifts, which we operationalized as breaks in same-role chains. Given that subject mentions that are co-referential with the subject of the previous clause are predominantly – though not exclusively – non-lexical, this means that most lexical subjects occur when clause chains are broken and topics switch from one referent to another. Since mentions in clause chains are maximally predictable and a very large proportion of subjects occurs in such a context, we further proposed a tentative definition of accessibility (for subject anaphors) in terms of occurrence in same-role chains, whereby the most accessible referent is one located in a chain-medial position. For mentions in other contexts, accessibility decreases proportionally with increasing anaphoric distance, and is further modulated by other factors such as humanness. Accessibility is tied directly to the likelihood of lexical expression, but not in a strict one-to-one entailment, since as noted, referential choices are inherently non-categorical (cf. Kibrik et al. 2013, 2016). Non-lexical expressions are available, in varying proportions, in all contexts and at all distances from the antecedent, and there is a small percentage of lexical expressions within clause chains. As such, lexical expression are not used to mark topic shifts specifically, given that lexical and non-lexical forms are about equally likely in this context. Rather, speakers resort to using more explicit forms if the new topic is too inaccessible (i.e. not recent enough) for a less explicit expression.

Tying lexicality rates to rates of breaks in clause chains – that is, contexts where referents have non-maximal accessibility – of course in turn necessit-

ates explaining the rate of chain breaks. In Section 9.2.3, we argued that there is an optimal density of referential information encoded in subject anaphors (cf. Jaeger 2006; Levy & Jaeger 2007) that spoken narrative discourse tends towards, from which follows the rate of topic shifts observed in the data. This runs counter to the claims that languages differ fundamentally in the density of information carried by discourse (cf. Huang 2000a: 262; Bickel 2003: 710). In this view, if the rate of topic shifts is too high, then the discourse will consequently not be coherent enough, as the lack of recurrent subject reference will fail to tie events together. Conversely, if the rate of topic shifts is too low, then the resulting discourse will not be informative enough and hence lack in tellability (Labov & Waletzky 1967). As such, the rate at which topics shift from one referent to another strikes a balance between being too frequent and too infrequent; there may be local extrema in information density, but over longer stretches of discourse, the rates will even out to those observed in the data. Given the absence of major cross-linguistic variation, this pattern is likely a matter of fundamental cognitive and communicative constraints, as it is present across unrelated languages with different narrative conventions.

In conclusion, this study lends further support to predictive modelling being a viable way of gaining insights into the mechanisms of referential choice, following Kibrik et al. (2013, 2016). It also shows that the use of spoken multilingual corpus data can provide valuable insights of a kind that are regrettably still quite rare in work on the subject, though this type of data is admittedly substantially more difficult to acquire and prepare in sufficient quantities. The importance of typological diversity in one's sample can hardly be understated; while one of the key take-aways of this study is that in essence the same principles apply across languages, it would not have been possible to conclusively identify which mechanisms are the most fundamental on the basis of data from only a single language, as certain language-specific characteristics such as the preference or dispreference for zero anaphors may confound potential universal properties of discourse structure.

10.2 | Outlook and future research

Lastly, this final section evaluates what could be done better, refined, and expanded on, and discusses a number of promising vanues for future research.

The predictive models in Chapter 7 were deliberately designed to be rather inclusive in terms of factor selection. While this is in and of itself not a detriment, as the modelling algorithms used are not confounded by unpredictive factors (unlike, e.g., regression models), a substantial proportion of the tested factors essentially came out as irrelevant. Some genuinely have no effect on referential choices, but others lack influence because they explain the same variance as other, more dominant factors. This most noticeable affects any factor associated with anaphoric distance, which, as we briefly noted above, is almost as predictive of lexicality by itself as the full multifactorial model. A future revision of the modelling approach might hence refine factor selection to minimize redundancy, and in this way streamline the explanatory power of the model.

But conversely, there are also a number of factors that were not tested here but could improve the accuracy of the predictions and offer better explanations. Chief among these is the involvement of semantic roles and the properties of the predicate (beyond transitivity); a system for annotating predicate semantics is currently in the early testing stages. Another possibly fruitful expansion might target considerations of rhetorical structure, as argued for in Kibrik & Krasavina (2005) and Kibrik et al. (2016), such as measures of hierarchical rhetorical distance to complement measures of linear textual distance. Lastly, it would be good to find better ways of accounting for cross-linguistic variation in the informativity of pronominal and basic lexical expressions.

While this study is a step forward in terms of typological representativity, especially compared to the prevailingly monolingual work on this subject – a lot of which relies on data from English, which, as seen here, might not be the best choice for drawing generalized conclusions – it is still a long way off from making robust claims about the universality of the observed patterns, given that only four of the world's major language families are represented here (in addition to a few isolates).

The question of representativity also extends to the types of texts examined, which here comprise only a single kind, spoken monologic narratives. As noted previously, the extent to which different text types affect rates of lex-

ical expression remains uncertain, though a preliminary assessment suggests that these differences may be substantial (see Section 9.2.3 and the summary in Section 5.2.1 above), and this will need to be tested (with consistent methodology) in the future. Doing so might involve the juxtaposition of lexicality rates with the rates of chain breaks across various text types, with the aim of determining whether it is differences in accessibility thresholds for specific referring expressions, the rate at which topic shifts occur, the types of referents mentioned, or other mechanisms (or a combination of multiple) that are responsible. Of course, this is assuming that other text types even pattern as consistently across diverse languages as spoken narratives do. A first step might be to contrast (spoken) monologues with dialogues/multilogues; the family problems picture task used for the SCOPIC project (Barth & Evans 2017a), for instance, has both a solo and a group component, a so might provide fruitful ground for such an investigation. A second step might further differentiate exlusively third person and autobiographical monologic narratives (i.e. where the speaker is external to the narrative vs. part of it, i.e. extradiegetic vs. diegetic, cf. Genette 1980); this is a distinction that is not consequently made in the present data set, but is likely to bear some influence owing to differences in the frequency of first person references (see Section 4.6.3).

Another issue that could not be accounted for are the discourse functions (and narrative functions) of one-off reference switches. Are they just incidental perspective shifts, intended to show the arrival of entities on scene or their response to actions performed by other entities, or do they serve a more structural role in managing discourse? For this, future work would need to investigate what kinds of predicates are involved in chain breaks (and specifically, chain starters and unchained mentions), which might rely on the aforementioned system of annotations of predicate semantics.

This study has very deliberately focused on anaphoric mentions – as per its title – largely for the sake of analytic simplicity and constraining its scope. For a more complete picture, however, we would also need to account for referent introductions, that is mentions of referents with no prior exponence in a text. If we take a production-oriented view of discourse, then new referents are not categorically different from given referents, but merely delineate an extreme case of inaccessibility (cf. views in Ariel 1990), meaning they should more or less seamlessly slot into the model of referential choice presented here.

Lastly, confirmation of our hypotheses regarding the optimal density of topic shifts and the tellability of discourse might require resorting to experi-

mental setups, as there are likely to be limits on what insights can be gleaned from the exclusive analysis of corpus data. One possible avenue for setting up a psycholinguistic experiment might involve some form of a discourse continuation task (cf. Kehler 2022): Here, respondents are presented with snippets of discourse, and tasked with providing a short continuation the narrative presented therein; the aim is to see at which rate the topic of the continuation matches that of last clause chain of trigger the narrative. The key parameters to vary there are the length of the final clause chain in the sequence, its transitivity, and the humanness its subject and, if present, object and other arguments. Of particular interest here are trigger narratives that defy narrative conventions and hence predictability: chains whose length far exceeds the average, sequences of exclusively one-off mentions, and so on.

Of course, the topics outlined above represent only a small subset of the possible avenues for research on referential choice, anaphora, and lexical expression. As it stands, the present study has raised more questions than it has answered, but it can hopefully serve as a starting point for future investigations, and as an example of the viability of modern analytical tools in corpus linguistics and typology.

References

Adamou, Evangelia & Haude, Katharina & Vanhove, Martine (eds.). 2018. *Information structure in lesser-described languages: Studies in prosody and syntax*. Amsterdam: John Benjamins.

Almor, Amit. 1999. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review* 106(4). 748–765. (https://doi.org/10.1037/0033-295X.106.4.748).

Anderson, Anthony & Garrod, Simon C. & Sanford, Anthony J. 1983. The accessibility of pronominal antecedents as a function of episode shifts in narrative texts. *Quarterly Journal of Experimental Psychology* 35(1). 427–440.

Anderson, John R. & Hastie, Reid. 1974. Individuation and reference in memory: Proper names and definite descriptions. *Cogntive Psychology* 6(4). 495–514. (https://doi.org/10.1016/0010-0285(74)90023-1).

Anderwald, Lieselotte & Wagner, Susanne. 2007. FRED: The Freiburg English Dialect Corpus. In Beal, Joan C. & Corrigan, Karen P. & Moisl, Hermann L. (eds.), *Creating and digitizing language corpora*, *Vol. 1*: *Synchronic databases*, 35–53. London: Macmillan.

Andrews, Avery. 2007. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description*, vol. 1: Clause structure, 132–223. Cambridge: Cambridge University Press.

Appelt, Douglas E. 1985. Planning English referring expressions. *Artificial Intelligence* 26(1). 1–33. (https://doi.org/10.1016/0004-3702(85)90011-6).

Appelt, Douglas E. & Kronfeld, Amichai. 1987. A computational model of referring. *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI)*. 640–674.

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(1). 65–87.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. 1996. Referring expressions and the +/– coreference distinction. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 13–36. Amsterdam: John Benjamins.

Ariel, Mira. 2001. Accessibility theory: An overview. In Sanders, Ted & Schilperoord, Joost & Spooren, Wilbert (eds.), *Text representation*: *Linguistic and psycholinguistic aspects*, 29–87. Amsterdam: John Benjamins.

Ariel, Mira. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes* 37(2). 91–116.

Ariel, Mira. 2006. Accessibility theory. In Brown, Keith (ed.), *Encyclopedia of language & linguistics*, 15–18. Amsterdam: Elsevier Science.

Ariel, Mira. 2009. Discourse, grammar, discourse. *Discourse Studies* 11(1). 5–36.

Arnold, Jennifer E. 1998. *Reference form and discourse patterns*. Unpublished Ph.D. dissertation, Stanford University.

Arnold, Jennifer E. 1999. *Marking salience: The similarity of topic and focus*. Unpublished manuscript, University of Pennsylvania.

Arnold, Jennifer E. 2003. Multiple constraints on reference form: Null, pronominal, and full reference in Mapudungun. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*: *Grammar as architecture for function*, 225–245. Amsterdam: John Benjamins.

Arnold, Jennifer E. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4(4). 187–203. (https://doi.org/10.1111/j.1749-818X.2010.00193.x).

Arnold, Jennifer E. & Bennetto, Loisa & Diehl, Joshua J. 2009. Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition* 110(2). 131–146. (https://doi.org/10.1016/j.cognition.2008.10.016).

Arnold, Jennifer E. & Griffin, Zenzi M. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language* 56(4). 521–536. (https://doi.org/10.1016/j.jml.2006.09.007).

Arnold, Jennifer E. & Losongco, Anthony & Wasow, Thomas & Ginstrom, Ryan. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1). 28–55. (https://doi.org/10.1353/lan.2000.0045).

Arvanti, Amalia. 2006. Erasure as a means of maintaining diglossia in Cyprus. *San Diego Linguistic Papers* 2. 25–38.

Ashby, William J. & Bentivoglio, Paola. 1993. Preferred argument structure in spoken French and Spanish. *Language Variation and Change* 5. 61–76.

Asher, Nicholas & Lascarides, Alex. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Austin, John L. 1962. *How to do things with words: The William James Lectures delivered at Harvard University in 1955*. Oxford: Clarendon Press.

Aylett, Matthew & Turk, Alice. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic promin-

ence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56. (https://doi.org/10.1177/00238309040470010201).

Baayen, Harald R. & Arppe, Antti. 2011. Statistical classification and principles of human learning. In Zeldes, Amir & Lüdeling, Anke (eds.), *Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (QITL-4)*, 8–11. Berlin: Humboldt-Universität zu Berlin.

Baayen, Harald R. & Janda, Laura A. & Nesset, Tore & Endresen, Anna & Makarova, Anastasia. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37. 253–291.

Baayen, Harald R. & Milin, Peter & Ramscar, Michael. 2016. Frequency in lexical processing. *Aphasiology* 30(11). 1174–1200. (https://doi.org/10.1080/02687038.2016.1147767).

Baddeley, Alan D. & Hitch, Graham. 1974. Working memory. In Bower, Gordon H. (ed.), *The psychology of learning and motivation*: *Advances in research and theory*, 47–89. New York: Academic Press.

Bard, Ellen G. & Anderson, Anne H. & Sotillo, Catherine & Aylett, Matthew & Doherty-Sneddon, Gwyneth & Newlands, Alison. 2000. Controlling intelligibility of referring expressions in dialogue. *Journal of Memory and Language* 42(1). 1–22. (https://doi.org/10.1006/jmla.1999.266).

Barth, Danielle & Evans, Nicholas. 2017a. The Social Cognition Parallax Corpus (SCOPIC): Design and overview. In Barth, Danielle & Evans, Nicholas (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (*Language Documentation & Conservation* special publication 12), 1–21. Honolulu, HI: University of Hawai'i Press.

Barth, Danielle & Evans, Nicholas (eds.). 2017b. *The Social Cognition Parallax Corpus (SCOPIC)* (*Language Documentation & Conservation* special publication 12). Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/45042).

Barth, Danielle & Evans, Nicholas & Arka, I Wayan & Bergqvist, Henrik & Forker, Diana & Gipper, Sonja & Hodge, Gabrielle & Kashima, Eri & Kasuga, Yuki & Kawakami, Carine & Kimoto, Yukinori & Knuchel, Dominique & Kogura, Norikazu & Kurabe, Keita & Mansfield, John & Narrog, Heiko & Pratiwi, Desak Putu Eka & van Putten, Saskia & Senge, Chikako & Tykhostup, Olena. 2021. Language vs. individuals in cross-linguistic corpus typology. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 179–232. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74661).

Bates, Douglas & Maechler, Martin & Bolker, Ben & Walker, Steve. 2020. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. (https://doi.org/10.18637/jss.v067.i01).

Bates, Elizabeth & MacWhinney, Brian. 1989. Functionalism and the competition model. In MacWhinney, Brian & Bates, Elizabeth (eds.), *The crosslinguistic study of sentence processing*, 3–76. Cambridge: Cambridge University Press.

Baumann, Stefan. 2006. Information structure and prosody: Linguistic categories for spoken language annotation. In Sudhoff, Stefan & Lenertova, Denisa & Meyer, Roland & Pappert, Sandra & Augurzky, Petra & Mleinek, Ina & Richter, Nicole & Schließer, Johannes (eds.), *Methods in empirical prosody research*, 153–180. Berlin: Mouton de Gruyter.

Belz, Anya & Kow, Eric & Varges, Sebastian. 2009. The GREC main subject reference generation challenge 2009: Overview and evaluation results. *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*.

Belz, Anya & Kow, Eric & Viethen, Jette & Gatt, Albert. 2008. The GREC challenge 2008: Overview and evaluation results. *Proceedings of the 5th International Natural Language Generation Conference (INLG 2008)*. 183–199.

Belz, Anya & Kow, Eric & Viethen, Jette & Gatt, Albert. 2010. Generating referring expressions in context: The GREC task evaluation challenges. In Krahmer, Emiel J. & Theune, Mariët (eds.), *Data-oriented methods and empirical evaluation*, 294–327. Berlin: Springer.

Bender, Emily. 1999. Constituting context: Null objects in English recipes revisited. *University of Pennsylvania Working Papers in Linguistics* 6(1). 53–68.

Bentz, Christian & Ferrer-i-Cancho, Ramon. 2016. Zipf's law of abbreviation as a language universal. In Bentz, Christian & Jäger, Gerhard & Yanovich, Igor (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tübingen: University of Tübingen. (http://hdl.handle.net/10900/68639).

Berman, Ruth A. & Slobin, Dan I. (eds.). 1994. *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Erlbaum.

Bhat, D. N. S. 2004. *Pronouns*. Oxford: Oxford University Press.

Biber, Douglas & Conrad, Susan. 2009. *Register, genre, and style* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Biber, Douglas & Conrad, Susan & Reppen, Randi. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Bickel, Balthasar. 2005. *Referential density in discourse: Typological and sociological factors*. Paper presented at ZAS, Berlin, February 2005.

Bickel, Balthasar. 2011. Grammatical relations typology. In Song, Jae Jung (ed.), *The Oxford handbook of linguistic typology*, 399–444. Oxford: Oxford University Press.

Bickel, Balthasar. 2013. Distributional biases in language families. In Bickel, Balthasar & Grenoble, Lenore & Peterson, David & Timberlake, Alan (eds.), *Language typology and historical contingency*: *In honor of Johanna Nichols*, 415–444. Amsterdam: John Benjamins.

Bischoffberger, Julia & Schnell, Stefan. 2014. *Thematic prominence and referential choice*. Paper presented at the 2014 Conference of the Linguistic Society of New Zealand, Hamilton, New Zealand, 23–25 November 2014.

Bock, Kathryn H. & Warren, Richard K. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21(1). 47–67. (https://doi.org/10.1016/0010-0277(85)90023-X).

Bogomolova, Natalia. 2021. Ergativity in Tabasaran: A reply to Woolford (2015). *Linguistic Inquiry*. (https://doi.org/10.1162/ling_a_00420).

Bogomolova, Natalia & Ganenkov, Dmitry & Schiborr, Nils N. 2021. Multi-CAST Tabasaran. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#tabasaran) (Accessed 2021-01-27).

Bolinger, Dwight. 1979. Pronouns in discourse. In Givón, Talmy (ed.), *Discourse and syntax*, vol. 12, 289–308. New York: Academic Press.

Bornkessel-Schlesewsky, Ina & Schlesewsky, Matthias. 2013. Neurotypology: Modeling crosslinguistic similarities and differences in the neurocognition of language comprehension. In Sanz, Montserrat & Laka, Itziar & Tanenhaus, Michael K. (eds.), *Language down the garden path*: *The cognitive and biological basis for linguistic structures*, 241–252. Oxford: Oxford University Press.

Botley, Simon P. 1999. *Corpora and discourse anaphora: Using corpus evidence to test theoretical claims*. Ph.D. dissertation, Lancaster University.

Botley, Simon P. & McEnery, Tony (eds.). 2000a. *Corpus-based and computational approaches to discourse anaphora*. Amsterdam: John Benjamins.

Botley, Simon P. & McEnery, Tony. 2000b. Discourse anaphora: The need for synthesis. In Botley, Simon P. & McEnery, Tony (eds.), *Corpus-based and computational approaches to discourse anaphora*, 1–43. Amsterdam: John Benjamins.

Botley, Simon P. & McEnery, Tony. 2001. Proximal and distal demonstratives: A corpus-based study. *Journal of English Linguistics* 29(3). 214–233.

Branco, António & McEnery, Tony & Mitkov, Ruslan (eds.). 2005. *Anaphora Processing: Linguistic, cognitive and computational modelling*. Amsterdam: John Benjamins.

Breiman, Leo & Friedman, Jerome H. & Stone, Charles J. & Olshen, Richard A. 1984. *Classification and regression trees*. New York: Chapman and Hall.

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive Processes* 10(2). 137–167.

Brennan, Susan E. & Friedman, Marilyn W. & Pollard, Carl J. 1987. A centering approach to pronouns. *Proceedings of the 25th Meeting of the Association for Computational Linguistics (ACL'87)*. 155–162.

Bresnan, Joan & Mchombo, Sam A. 1987. Topic, pronoun, and agreement in Chichewa. *Language* 63(4). 741–782.

Brickell, Timothy C. 2016. Multi-CAST Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#tondano) (Accessed 2019-03-08).

Brown, Gillian & Yule, George. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Butt, Miriam. 2003. The light verb jungle. *Harvard Working Papers in Linguistics* 9(1). 1–49.

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Amberber, Mengistu & Harvey, Mark & Baker, Brett (eds.), *Complex predicates in cross-linguistic perspective*, 48–78. Cambridge: Cambridge University Press.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733. (https://doi.org/10.1353/lan.2006.0186).

Bybee, Joan. 2009. Language universals and usage-based theory. In Christiansen, Morten H. & Collins, Christopher & Edelman, Shimon (eds.), *Language universals*. Oxford: Oxford University Press. (https://doi.org/10.1093/acprof:oso/9780195305432.003.0002).

Caballero, Gabriela & Kapatsinski, Vsevolod M. 2015. Perceptual functionality of morphological redundancy in Choguita Rarámuri (Tarahumara). *Cognition and Neuroscience* 30(9). 1134–1143.

Callaway, Charles B. & Lester, James C. 2002. Pronominalization in generated discourse and dialogue. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*. 88–95.

Cameron, Richard. 1994. Switch reference, verb class and priming in a variable syntax. *Papers from the Regional Meeting of the Chicago Linguistics Society: Parasession on Variation in Linguistic Theory* 30(2). 27–45.

Cameron, Richard & Flores-Ferrán, Nydia. 2004. Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context* 1(1). 41–65. (https://doi.org/10.1075/sic.1.1.05cam).

Campbell, Lyle & Lee, Nala H. & Okura, Eve & Simpson, Sean & Ueki, Kaori (eds.). 2010. *The catalogue of endangered languages (ElCat)*. (http://endangeredlanguages.com/userquery/download/) (Accessed 2020-09-01).

Carlson, Lynn & Marcu, Daniel & Okurowski, Mary Ellen. 2002. *RST Discourse Treebank (LDC2002T07)*. Philadelpha, PA: Linguistic Data Consortium. (https://doi.org/10.35111/4w31-m996).

Chafe, Wallace. 1974. Language and consciousness. *Language* 50. 111–133.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, Charles N. (ed.), *Subject and topic*, 25–55. New York: Academic Press.

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Tomlin, Russell S. (ed.), *Coherence and grounding in discourse*, 21–51. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago, IL: The University of Chicago Press.

Chafe, Wallace. 1996. Inferring identifiability and accessibility. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 37–46. Amsterdam: John Benjamins.

Chambers, Craig G. & Smyth, Ron. 1998. Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language* 39(4). 593–608.

Chomsky, Noam. 1981. *Lectures on government and binding: The Pisa lectures*. Dordrecht: Foris.

Chomsky, Noam. 1982. *Some concepts and consequences of the theory of government and binding*. MIT Press: Cambridge, MA.

Chomsky, Noam. 1993. *Lectures on government and binding: The Pisa lectures*. Berlin: Mouton de Gruyter.

Christiansen, Thomas. 2011. *Cohesion: A discourse perspective*. Bern: Peter Lang.

Clancy, Patricia M. 1980. Referential choice in English and Japanese narrative discourse. In Chafe, Wallace (ed.), *The Pear Stories*: *Cognitive, cultural, and linguistic aspects of narrative production*, 127–202. Norwood, NJ: Ablex.

Clark, Hebert H. & Haviland, Susan E. 1977. Comprehension and the given–new contract. In Freedle, Roy O. (ed.), *Discourse production and comprehension*, 1–40. Norwood, NJ: Ablex.

Clark, Herbert H. & Schreuder, Robert & Buttrick, Samuel. 1983. Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior* 22(2). 245–258. (https://doi.org/10.1016/S0022-5371(83)90189-5).

Clark, Herbert H. & Sengul, C. J. 1979. In search of referents for nouns and pronouns. *Memory and Cognition* 7(1). 35–41.

Clark, Herbert H. & Wasow, Thomas. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37(3). 201–242. (https://doi.org/10.1006/cogp.1998.0693).

Clark, Herbert H. & Wilkes-Gibbs, Deanna. 1986. Referring as a collaborative process. *Cognition* 22(1). 1–39. (https://doi.org/10.1016/0010-0277(86)90010-7).

Comrie, Bernard. 1979. Definite and animate direct objects: A natural class. *Linguistica silesiana* 3. 13–21.

Comrie, Bernard. 1989. *Language universals and linguistic typology*. 2nd edn. Oxford: Blackwell.

Contemori, Carla & Dussias, Paola E. 2016. Referential choice in a second language: Evidence for a listener-oriented approach. *Language, Cognition and Neuroscience* 31(10). 1–16. (https://doi.org/10.1080/23273798.2016.1220604).

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville. 2000. *Number*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2003. Agreement: The range of the phenomenon and the principles of the Surrey Database of Agreement. In Brown, Dunstan & Corbett, Greville G. & Tiberius, Carole (eds.), *Agreement*: *A typological perspective*, 155–202.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Cowles, H. Wind & Walenski, Matthew & Kluender, Robert. 2007. Linguistic and cognitive prominence in anaphor resolution: Topic, constrative focus and pronouns. *Topoi* 26. 3–18. (https://doi.org/10.1007/s11245-006-9004-6).

Cristea, Dan & Ide, Nancy & Marcu, Daniel & Tablan, M. Valentin. 2000. Discourse structure and co-reference: An empirical study. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'00)*.

Cristea, Dan & Ide, Nancy & Romary, Laurent. 1998. Veins theory: A model of global discourse cohesion and coherence. *Proceedings of the 17th COLING and the 36th Annual Meeting of the ACL (COLING-ACL'98)*. 281–285.

Croft, William. 2013. Agreement as anaphora, anaphora as coreference. In Bakker, Dik & Haspelmath, Martin (eds.), *Languages across boundaries*: *Studies in memory of Anna Siewierska*, 107–129. Berlin: Mouton De Gruyter.

Croft, William A. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Culotta, Aron & Wick, Michael & Hall, Robert & McCallum, Andrew. 2007. First-order probabilistic models for co-reference resolution. *Proceedings of HLT-NAACL 2007*. 81–88.

Cysouw, Michael. 2003. *The paradigmatic structure of person*. Oxford: Oxford University Press.

Cysouw, Michael & Wälchli, Bernhard. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung (STUF)* 60(2). 95–99. (https://doi.org/10.1524/ stuf.2007.60.2.95).

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(1). 37–77.
Dahl, Östen. 2008. Animacy and egophoricity: Grammar, ontogeny, and phylogeny. *Lingua* 118(2). 141–150.
Dahl, Östen. 2015. *How WEIRD are WALS languages?* Paper presented at the Closing Conference of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 1–3 May 2015.
Dahl, Östen & Fraurud, Kari. 1996. Animacy in grammar and discourse. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 47–64. Amsterdam: John Benjamins.
Dale, Robert & Reiter, Ehud. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2). 233–263.
Dalrymple, Mary & Nikolaeva, Irina. 2011. *Objects and information structure*. Cambridge: Cambridge University Press.
The agreement cross-reference continuum. 2005. Person marking in functional grammar. In de Groot, Caspar & Hengeveld, Kees (eds.), *Morphosyntactic expression in functional grammar*, 203–248. Berlin: de Gruyter.
Dell, Gary S. 1986. A spreading activation theory of retrieval in language production. *Psychological Review* 93(3). 283–321.
Diessel, Holger. 1999. *Demonstratives: Form, function and grammaticalization*. Amsterdam: John Benjamins.
Dingemanse, Mark & Torreira, Francisco & Enfield, Nick J. 2013. Is "Huh?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE* 8(11). (https://doi.org/10.1371/journal.pone.0078273).
Dipper, Stefanie & Rieger, Christine & Seiss, Melanie & Zinsmeister, Heike. 2011. Abstract anaphors in German and English. In Hendrickx, Iris & Lalitha Devi, Sobha & Branco, António & Mitkov, Ruslan (eds.), *Anaphora processing and applications*: *8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011*, 96–107. Dordrecht: Springer.
Dixon, R. M. W. 1979. Ergativity. *Language* 55. 59–138.
Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.
Dixon, R. M. W. 2009. *Basic linguistic theory: Grammatical topics*. Oxford: Oxford University Press.
Dixon, R. M. W. & Aikhenvald, Alexandra Y. 2002. *Word: A cross-linguistic typology*. Cambridge: Cambridge University Press.
Doddington, George & Mitchell, Alexis & Przybocki, Mark & Ramshaw, Lance & Strassel, Stephanie & Weischedel, Ralph. 2004. The Automatic Content Extraction (ACE) program — Tasks, data, and evaluation. *Proceedings of the 4nd Inter-*

national Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 26–28 May 2004.

Donohue, Mark. 2008. Semantic alignment systems. In Donohue, Mark & Wichmann, Søren (eds.), *The typology of semantic alignment*, 24–75. Oxford: Oxford University Press.

Dooley, Robert A. & Levinsohn, Stephen H. 2001. *Analyzing discourse: A manual of basic concepts*. Dallas, TX: SIL.

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (`http://wals.info`) (Accessed 2016-02-05).

Du Bois, John. 1980. Beyond definiteness: The trace of identity in discourse. In Chafe, Wallace (ed.), *The Pear Stories*: *Cognitive, cultural, and linguistic aspects of narrative production*, 203–274. Norwood, NJ: Ablex.

Du Bois, John. 1987a. Absolutive zero: Paradigm adaptivity in Sacapultec Maya. *Lingua* 71(2). 203–222.

Du Bois, John. 1987b. The discourse basis of ergativity. *Language* 63(4). 805–855.

Du Bois, John. 2003a. Argument structure: Grammar in use. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*: *Grammar as architecture for function*, 11–60. Amsterdam: John Benjamins.

Du Bois, John. 2003b. Discourse and grammar. In Tomasello, Michael (ed.), *The new psychology of language*: *Cognitive and functional approaches to language structure*, vol. 2, 47–88. Mahwah, NJ: Erlbaum.

Du Bois, John. 2017. Ergativity in discourse and grammar. In Coon, Jessica & Massam, Diane & Travis, Lisa D. (eds.), *The Oxford handbook of ergativity*, 23–57. Oxford: Oxford University Press.

Du Bois, John W. & Chafe, Wallace L. & Meyer, Charles & Thompson, Sandra A. & Englebretson, Robert & Martey, Nii. 2005. *Santa Barbara Corpus of Spoken American English, parts 1–4*. Philadelphia, PA: Linguistic Data Consortium.

Du Bois, John W. & Schuetze-Coburn, Stephan & Cumming, Susanna & Paolino, Danae. 1993. Outline of discourse transcription. In Edwards, Jane A. & Lampert, Martin D. (eds.), *Talking data*: *Transcription and coding in discourse research*, 45–89. Hillsdale, NJ: Erlbaum.

Durie, Mark. 2003. New light on information pressure: Information conduits, 'escape valves', and role alignment stretching. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*: *Grammar as architecture for function*, 159–196. Amsterdam: John Benjamins.

Efimova, Zoya V. 2006. *Referencial'naja struktura narrativa v Japonskom jazyke (v sopostavlenii s Russkim)* [Referential structure of narrative in Japanese (as compared to Russian)]. Ph.D. dissertation, Russian State University for the Humanities.

Elith, Jane & R., Leathwick John & Hastie, Trevor. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4). 802–813. (https://doi.org/10.1111/j.1365-2656.2008.01390.x).

Engelhardt, Paul E. & Bailey, Karl G. D. & Ferreira, Fernanda. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language* 54(4). 554–573. (https://doi.org/10.1016/j.jml.2005.12.009).

English Dialects Research Group. 2005. *Freiburg English Dialect Corpus (FRED)*. (http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/) (Accessed 2016-02-03).

Evans, Nicholas. 2020. Why the comparability problem is still central in typology. *Linguistic Typology* 24(3). 417–425. (https://doi.org/10.1515/lingty-2020-2055).

Everett, Caleb. 2009. A reconsideration of the motivations for preferred argument structure. *Studies in Language* 33(1). 1–24. (https://doi.org/10.1075/sl.33.1.02eve).

Falk, Yehuda N. 2006. *Subjects and universal grammar: An explanatory theory*. Cambridge: Cambridge University Press.

Farrell, Patrick (ed.). 2005. *Grammatical relations*. Oxford: Oxford University Press.

Fernandez-Vest, M. M. Jocelyne & Van Valin, Robert D., Jr. 2016. *Information structuring spoken language from a cross-linguistic perspective*. Berlin: Mouton de Gruyter.

Féry, Caroline & Fanselow, Gisbert & Krifka, Manfred (eds.). 2006. *The notions of information structure*. Potsdam: Universitätsverlag Potsdam.

Féry, Caroline & Ishihara, Shinichirō (eds.). 2014. *The Oxford handbook of information structure*. Oxford: Oxford University Press.

Féry, Caroline & Krifka, Manfred. 2008. Information structure: Notional distinctions, ways of expression. In van Sterkenburg, Piet (ed.), *Unity and diversity of languages*, 123–135. Amsterdam: John Benjamins.

Fiedler, Ines & Schwarz, Anne (eds.). 2010. *The expression of information structure: A documentation of its diversity across Africa*. Amsterdam: John Benjamins.

Fillmore, Charles J. 1982. Frame semantics. In The Linguistic Society of Korea (ed.), *Linguistics in the morning calm*: *Selected papers from SICOL-1981*, 111–137. Seoul: Hanshin.

Foley, William A. 2008. The place of Philippine languages in a typology of voice systems. In Austin, Peter K. & Musgrave, Simon (eds.), *Voice and grammatical relations in Austronesian languages*, 22–44. Stanford, CA: CSLI Publications.

Foley, William A. & Van Valin, Robert D., Jr. (eds.). 1984. *Functional syntax and universal grammar*. Cambridge: Cambridge University Press.

Foraker, Stephani & McElree, Brian. 2007. The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language* 50(3). 357–383. (https://doi.org/10.1016/j.jml.2006.07.004).

Ford, Cecilia. 2004. Contingency and units in interaction. *Discourse Studies* 6(1). 27–52. (https://doi.org/10.1177/1461445604039438).

Ford, Cecilia & Fox, Barbara A. & Thompson, Sandra A. 1996. Practices in the construction of turns: The 'TCU' revisited. *Pragmatics* 6(3). 427–454. (https://doi.org/10.1075/prag.6.3.07for).

Forker, Diana. 2011. Finiteness in Hinuq. *Linguistic Discovery* 9. 3–29.

Forker, Diana. 2013. Hinuq verb forms and finiteness. *Acta Orientalia Academiae Scientarium Hungaricae* 66. 69–93.

Forker, Diana. 2020. *A grammar of Sanzhi Dargwa* (Languages of the Caucasus 2). Berlin: Language Science Press. (https://doi.org/10.5281/zenodo.3339225).

Forker, Diana & Mutalov, Rasul & Kaliszewska, Iwona & Belyaev, Oleg. 2019. *Shiri / Sanzhi. DoBeS archive at The Language Archive*. (https://hdl.handle.net/1839/81fdc5ba-be3a-4695-9337-f77159d6705a).

Forker, Diana & Schiborr, Nils N. 2019. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#sanzhi) (Accessed 2019-04-05).

Fox, Barbara. 1986. Local patterns and general principles in cognitive processes. *Text* 16(1). 25–51.

Fox, Barbara. 1987a. Anaphora in popular written English narratives. In Tomlin, Russell S. (ed.), *Coherence and grounding in discourse*, 157–174. Amsterdam: John Benjamins.

Fox, Barbara. 1987b. *Discourse structure and anaphora: Written and conversational English*. Cambridge: Cambridge University Press.

Fox, Barbara (ed.). 1996. *Studies in anaphora*. Amsterdam: John Benjamins.

Francis, Hartwell & Gregory, Michelle & Michaelis, Laura. 1999. Are lexical subjects deviant? *Proceedings of the Chicago Linguistics Society* 35(1). 85–97.

Fraurud, Kari. 1996. Cognitive ontology and NP form. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 65–87. Amsterdam: John Benjamins.

Fretheim, Thorstein. 1996. Accessing contexts with intonation. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 89–112. Amsterdam: John Benjamins.

Fretheim, Thorstein & Gundel, Jeanette K. (eds.). 1996. *Reference and referent accessibility*. Amsterdam: John Benjamins.

Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5). 1189–1232. (https://doi.org/10.1214/aos/1013203451).

Friedman, Jerome H. 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38(4). 367–378. (https://doi.org/10.1016/S0167-9473(01)00065-2).

Friedman, Jerome H. & Hastie, Trevor & Tibshirani, Robert. 2000. Additive logistic regression: A statistical view of boosting. *Annals of Statistics* 28(2). 337–407. (https://doi.org/10.1214/aos/1016218223).

Friedman, Jerome H. & Meulman, Jacqueline J. 2003. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine* 22(9). 1365–1381. (https://doi.org/10.1002/sim.1501).

Fries, Peter H. 1994. On theme, rheme and discourse goals. In Coulthard, Malcolm (ed.), *Advances in written text analysis*, 229–249. London: Routledge.

Fukumura, Kumiko & van Gompel, Roger P. G. 2009. Speakers use their own discourse model to determine referents' accessibility during the production of referring expressions. In van Deemter, Kees & Gatt, Albert & van Gompel, Roger P. G. & Krahmer, Emiel J. (eds.), *Proceedings of the workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009)*. Amsterdam: University of Tilburg.

Fukumura, Kumiko & van Gompel, Roger P. G. 2011. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes* 26(10). 1472–1504. (https://doi.org/10.1080/01690965.2010.506444).

Fukumura, Kumiko & van Gompel, Roger P. G. 2012. Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? *Cognitive Science* 36(7). 1289–1311. (https://doi.org/10.1111/j.1551-6709.2012.01255.x).

Fukumura, Kumiko & van Gompel, Roger P. G. & Harley, Trevor & Pickering, Martin J. 2011. How does similarity-based interference affect the choice of referring expression? *Journal of Memory and Language* 65(3). 331–344. (https://doi.org/10.1016/j.jml.2011.06.001).

Fukumura, Kumiko & van Gompel, Roger P. G. & Pickering, Martin J. 2010. The use of visual context during the production of referring expressions. *The Quarterly Journal of Experimental Psychology* 63(9). 1700–1715. (https://doi.org/10.1080/17470210903490969).

Futrell, Richard & Dyer, William & Scontras, Greg. 2020a. What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks. *Proceedings of the 58nd Annual Meeting of the Association for Computational Linguistics (ACL'20)*. 2003–2012. (https://doi.org/10.18653/v1/2020.acl-main.181).

Futrell, Richard & Levy, Roger P. & Gibson, Edward. 2020b. Dependency locality as an explanation principle for word order. *Language* 76(2). 371–412.

Futrell, Richard & Mahowald, Kyle & Gibson, Edward. 2015. Quantifying word order freedom in dependency corpora. *Proceedings of the 3rd International Conference on Dependency Linguistics (Depling 2015), Uppsala, Sweden, 24–26 August 2015*. 91–100.

Garnham, Alan. 1987. Eeffects of antecedent distance and intervening text structure in the interpretation of ellipses. *Language and Speech* 30(1). 59–68. (https://doi.org/{10.1177/002383098703000105}).

Garnham, Alan. 2001. *Mental models and the interpretation of anaphora*. London: Psychology Press.

Garrod, Simon & Sanford, Anthony J. 1982. The mental representation of discourse in a focused memory system: Implications for the interpretation of anaphoric noun-phrases. *Journal of Semantics* 1(1). 21–41.

Garrod, Simon & Trabasso, Tom. 1973. A dual-memory information processing interpretation of sentence comprehension. *Journal of Verbal Learning and Verbal Behaviour* 12(2). 155–167. (https://doi.org/10.1016/S0022-5371(73)80005-2).

Gatt, Albert & Krahmer, Emiel J. & van Deemter, Kees & van Gompel, Roger P. G. 2014. Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience* 29(8). 899–911.

Ge, Niyu & Hale, John & Charniak, Eugene. 1998. A statistical approach to anaphora resolution. *Proceedings of the Workshop on Very Large Corpora* 6. 161–171.

Genette, Gérard. 1980. *Narrative discourse: An essay in method*. Ithaca, NY: Cornell University Press.

Genetti, Carol & Crain, Laura. 2003. Beyond preferred argument structure: Sentences, pronouns, and given referents in Nepali. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*: *Grammar as architecture for function*, 197–223. Amsterdam: John Benjamins.

Genzel, Dmitriy & Charniak, Eugene. 2002. Entropy rate constancy in text. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*. 199–206.

Gernsbacher, Morton Ann & Givón, Talmy (eds.). 1995. *Coherence in spontaneous texts*. Amsterdam: John Benjamins.

Giangoullis, Konstantinos G. 2009. *Kypriaka paradosiaka paramytha: Ek stomatos Elenis Mich, Satsia, Apo to Geri-Pyroi (1887–1982)* (Viviothiki Kyprion Laikon Poiiton 71). Leukosia: Theopress Publications.

Gipper, Sonja. 2016. Constraints on choice of referring expression in Yurakaré. In Holler, Anke & Suckow, Katja (eds.), *Empirical perspectives on anaphora resolution*, 143–168. Berlin: de Gruyter.

Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In Li, Charles N. (ed.), *Subject and topic*, 149–188. New York: Academic Press.
Givón, Talmy. 1978. Referentiality and definiteness. In Greenberg, Joseph H. & Ferguson, Charles & Moravcsik, Edith (eds.), *Universals of human language*: *Syntax*, 291–330. Stanford, CA: Stanford University Press.
Givón, Talmy (ed.). 1979. *Discourse and syntax* (Syntax and semantics 12). New York: Academic Press.
Givón, Talmy (ed.). 1983a. *Topic continuity in discourse* (Typological Studies in Language 3). Amsterdam: John Benjamins.
Givón, Talmy. 1983b. Topic continuity in discourse: An introduction. In Givón, Talmy (ed.), *Topic continuity in discourse*, vol. 3, 1–42. Amsterdam: John Benjamins.
Givón, Talmy. 1983c. Topic continuity in spoken English. In Givón, Talmy (ed.), *Topic continuity in discourse*, vol. 3, 343–364. Amsterdam: John Benjamins.
Givón, Talmy. 1990. *Syntax: A functional typological introduction*. Amsterdam: John Benjamins.
Givón, Talmy. 1995. Coherence in text versus coherence in mind. In Gernsbacher, Morton Ann & Givón, Talmy (eds.), *Coherence in spontaneous texts*, 59–115. Amsterdam: John Benjamins.
Givón, Talmy. 1995. *Functionalism and grammar*. Amsterdam: John Benjamins.
Givón, Talmy. 2001. *Syntax: An introduction*. Amsterdam: John Benjamins.
Givón, Talmy. 2017. Zero, pronouns and clause-chaining: Toward a diachronic understanding. *Lingua* 185(1). 96–120. (https://doi.org/10.1016/j.lingua.2016.08.001).
Givón, Talmy. 2020. *Coherence*. Amsterdam: John Benjamins.
Godfrey, John & Holliman, Edward & McDaniel, Jane. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In IEEE (ed.), *Proceedings of ICASSP-92*: *1992 IEEE International Conference on Acoustics, Speech and Signal Processing, 23–26 March 1992, The San Francisco Marriot, San Francisco, California*, 517–520. New York: Institute of Electrical and Electronics Engineers. (https://doi.org/10.1109/ICASSP.1992.225858).
Gordon, Matthew K. 2016. *Phonological typology*. Oxford: Oxford University Press.
Gordon, Peter C. & Chan, Davina. 1994. Pronouns, passives, and discourse coherence. *Journal of Memory and Language* 34(2). 216–231.
Gordon, Peter C. & Grosz, Barbara J. & Gilliom, Laura A. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science* 17(3). 311–347.
Greenbacker, Charles F. & McCoy, Kathleen F. 2009. Feature selection for reference generation by psycholinguistic research. In van Deemter, Kees & Gatt, Albert & van Gompel, Roger P. G. & Krahmer, Emiel J. (eds.), *Proceedings of the workshop on the Production of Referring Expressions: Bridging the gap between computa-*

tional and empirical approaches to reference (PRE-CogSci 2009). Amsterdam: University of Tilburg.
Greenbaum, Sidney (ed.). 1996. *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon Press.
Greenberg, Joseph H. (ed.). 1963. *Universals of language*. Cambridge, MA: MIT Press.
Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.
Greenwell, Brandon & Boehmke, Bradley & Cunningham, Jay & GBM Developers. 2020. *gbm: Generalized boosted regression models. R package version 2.1.8*. (`http://CRAN.R-project.org/package=gbm`) (Accessed 2020-09-16).
Grice, H. P. 1975. Logic and conversation. In Cole, Peter & Morgan, Jerry L. (eds.), *Syntax and semantics 3*: *Speech acts*, 41–58. New York: Academic Press.
Gries, Stefan T. 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*. 1–31. (`https://doi.org/10.1515/cllt-2018-0078`).
Grimes, Joseph. 1975. *The thread of discourse*. The Hague: Mouton.
Grosz, Barbara J. & Joshi, Aravind K. & Weinstein, Scott. 1983. Providing a unified account of definite noun phrases in discourse. *Proceedings of the 21th Meeting of the Association for Computational Linguistics (ACL'83)*. 44–50.
Grosz, Barbara J. & Joshi, Aravind K. & Weinstein, Scott. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2). 203–225.
Grosz, Barbara & Sidner, Candace. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics* 12(3). 175–204.
Grüning, André & Kibrik, Andrej A. 2005. Modelling referential choice in discourse: A cognitive calculative approach and a neural network approach. In Branco, António & McEnery, Tony & Mitkov, Ruslan (eds.), *Anaphora Processing*: *Linguistic, cognitive and computational modelling*, 163–198. Amsterdam: John Benjamins.
Guérin, Valérie. 2019. *Bridging constructions*. Berlin: Language Science Press.
Gundel, Jeanette K. 1985. Shared knowledge and topicality. *Journal of Pragmatics* 9(1). 83–107.
Gundel, Jeanette K. 1988. Universals of topic-comment structure. In Hammond, Michael & Moravscik, Edith A. & Wirth, Jessica R. (eds.), *Studies in syntactic typology*, 209–239. Amsterdam: John Benjamins.
Gundel, Jeanette K. 2003. Information structure and referential givenness/newness: How much belongs in the grammar? In Müller, Stefan (ed.), *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar, Michigan State University*, 122–142. Stanford, CA: CSLI Publications.

Gundel, Jeanette K. 2010. Reference and accessibility from a givenness hierarchy perspective. *International Review of Pragmatics* 2(2). 148–168. (https://doi.org/10.1163/187731010X528322).

Gundel, Jeanette K. & Bassene, Mamadou & Gordon, Bryan J. & Humnick, Linda. 2010. Testing predictions of the givennes hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics* 42(7). 1770–1785. (https://doi.org/10.1016/j.pragma.2009.09.010).

Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 1993a. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307.

Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 1993b. On the generation and interpretation of demonstrative expressions. *Proceedings of the 12th International Conference on Computational Linguistics (COLING'93)*.

Gundel, Jeanette K. & Hegarty, Michael & Borthen, Kaja. 2003. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information* 12(3). 281–299.

Hadjidas, Harris & Vollmer, Maria C. 2015. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#cypgreek) (Accessed 2019-03-08).

Hadjioannou, Xenia & Tsiplakou, Stavroula & Kappler, Matthias. 2011. Language policy and language panning in Cyprus. *Current Issues in Language Planning* 12(4). 503–569. (https://doi.org/10.1080/14664208.2011.629113).

Haeri, Niloofar. 1989. Overt and non-overt subjects in Persian. *IPrA Papers in Pragmatics* 3(1). 155–166.

Haghighi, Aria & Klein, Dan. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*.

Haghighi, Aria & Klein, Dan. 2010. Coreference resolution in a modular, entity-centered model. *Proceedings of HLT '10*. 385–393.

Haig, Geoffrey. 2018a. Northern Kurdish (Kurmanjî). In Haig, Geoffrey & Khan, Geoffrey (eds.), *The languages and linguistics of Western Asia*: *An areal perspective*, 106–158. Berlin: Mouton de Gruyter.

Haig, Geoffrey. 2018b. The grammaticalization of object pronouns: Why differential object indexing is an attractor state. *Linguistics* 56(4). 781–818. (https://doi.org/10.1515/ling-2018-0011).

Haig, Geoffrey & Forker, Diana. 2018. Agreement in grammar and discourse: A research overview. *Linguistics* 56(4). 715–734. (https://doi.org/10.1515/ling-2018-0014).

Haig, Geoffrey & Nau, Nicole & Schnell, Stefan & Wegener, Claudia (eds.). 2011a. *Documenting endangered languages: Achievements and perspectives*. Berlin: Mouton de Gruyter.

Haig, Geoffrey & Öpengin, Ergin. 2018. Kurmanji in Turkey: Structure, varieties, and status. In Bulut, Christiane (ed.), *Linguistic minorities in Turkey and Turkic-speaking minorities of the peripheries*, vol. 111. Wiesbaden: Harrassowitz.

Haig, Geoffrey & Schiborr, Nils N. & Schnell, Stefan. 2020. *On potential statistical universals of grammar in discourse: Evidence from Multi-CAST*. Paper presented at the Workshop Corpus-based typology: Spoken language from a cross-linguistic perspective, as part of the 42nd Annual Conference of the German Linguistic Society (DGfS 2020), Hamburg, Germany, 4–-6 March 2020.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (https://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2019-03-08).

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/) (Accessed 2019-03-08).

Haig, Geoffrey & Schnell, Stefan. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618. (https://doi.org/10.1353/lan.2016.0049).

Haig, Geoffrey & Schnell, Stefan & Schiborr, Nils N. 2021. Universals of reference in discourse and grammar: Evidence from the Multi-CAST collection of spoken corpora. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 141–177. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74660).

Haig, Geoffrey & Schnell, Stefan & Wegener, Claudia. 2011b. Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Haig, Geoffrey & Nau, Nicole & Schnell, Stefan & Wegener, Claudia (eds.), *Documenting endangered languages*: *Achievements and perspectives*, 55–86. Berlin: Mouton de Gruyter.

Haig, Geoffrey & Vollmer, Maria C. & Thiele, Hanna. 2019. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#nkurd) (Accessed 2019-07-05).

Hajičová, Eva & Panevová, Jarmila & Sgall, Petr. 2000. Coreference in annotating a large corpus. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece, 31 May–2 June 2000*.

Hale, Kenneth. 1992. Subject obviation, switch reference, and control. In Larson, Richard K. & Iatridou, Sabine & Lahiri, Utpal & Higginbotham, James (eds.), *Control and grammar*, 51–77. Dordrecht: Springer.

Halliday, M. A. K. & Hasan, Ruqaiya. 1976. *Cohesion in English*. London: Longman.

Halmari, Helena. 1996. On accessibility and coreference. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 155–178. Amsterdam: John Benjamins.

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin. 2020. *Glottolog 4.2.1*. Jena: Max Planck Institute for the Science of Human History. (https://doi.org/10.5281/zenodo.3754591). (http://glottolog.org) (Accessed 2020-08-24).

Harvie, Dawn. 1998. Null subject in English: Wonder if it exists? *Cahiers Linguistiques d'Ottawa* 16. 15–25.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(4). 663–687.

Haspelmath, Martin. 2011. On S, A, P, T, and R as comparative concepts for alignment typology. *Linguistic Typology* 15(3). 535–567.

Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntactic status of bound person forms. In Bakker, Dik & Haspelmath, Martin (eds.), *Languages across boundaries*: *Studies in memory of Anna Siewierska*, 197–226. Berlin: Mouton De Gruyter.

Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633. (https://doi.org/10.1017/S0022226720000535).

Haspelmath, Martin & Calude, Andreea & Spagnol, Michael & Narrog, Heiko & Bamyaci, Elif. 2014. Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.

Hawkins, John A. 1978. *Definiteness and indefiniteness*. London: Croom Helm.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

He, Haibo & Garcia, Edwardo A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9). 1263–1284. (https://doi.org/10.1109/TKDE.2008.239).

He, Haibo & Ma, Yunqian. 2013. *Imbalanced learning: Foundations, algorithms, and applications*. New York: Wiley.

Hearst, Marti. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, 9–16. La Cruces, NM: Association for Computational Linguistics. (https://doi.org/10.3115/981732.981734).

Heath, Jeffrey G. 1975. Some functional relationships in grammar. *Language* 51(1). 89–104.

Hedberg, Nancy & Gundel, Jeanette K. & Zacharski, Ron. 2007. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In Branco, António & McEnery, Tony & Mitkov, Ruslan & Silva, Fátima (eds.), *Proceedings of DAARC-2007*: *The 6th Discourse Anaphora and Anaphor Resolution Colloquium*, 31–36. Porto: Centro de Linguistica da Universidade do Porto.

Helasvuo, Marja-Liisa & Huumo, Tuomas. 2015. Canonical and non-canonical subjects in constructions: Perspectives from cognition and discourse. In Helasvuo, Marja-Liisa & Huumo, Tuomas (eds.), *Subjects in constructions*: *Canonical and non-canonical*, 1–9. Amsterdam: John Benjamins.

Heller, Daphna & Gorman, Kristen S. & Tanenhaus, Michael K. 2012. To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science* 4(2). 290–305. (https://doi.org/10.1111/j.1756-8765.2012.01182.x).

Helmbrecht, Johannes & Denk, Lukas & Thanner, Sarah & Tonetti, Ilenia. 2018. Morphosyntactic coding of proper names and its implications for the Animacy Hierarchy. In Cristofaro, Sonia & Zúñiga, Fernando (eds.), *Typological hierarchies in synchrony and diachrony*, 377–402. Amsterdam: John Benjamins.

Hernández, Nuria. 2006. *User's guide to FRED: Freiburg Corpus of English Dialects*. (https://www.freidok.uni-freiburg.de/data/2489) (Accessed 2015-12-20).

Himmelmann, Nikolaus P. 1996. Demonstratives in narrative discourse: A taxonomy of universal uses. In Fox, Barbara (ed.), *Studies in anaphora*, 205–254. Amsterdam: John Benjamins.

Himmelmann, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(2). 161–195. (https://doi.org/10.1515/ling.1998.36.1.161).

Himmelmann, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960. (https://doi.org/10.1353/lan.2014.0105).

Hinds, John. 1978. Anaphora in Japanese conversation. In Hinds, John (ed.), *Anaphora in Discourse*, 136–179. Alberta: Linguistic Research.

Ho, Tin Kam. 1995. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, 14–16 August 1995*. 278–282.

Hobbs, Jerry R. 1976. Pronoun resolution. *Research reports of the Department of Computer Science, City College, City University of New York* 76(7).

Holler, Anke & Suckow, Katja (eds.). 2016. *Empirical perspectives on anaphora resolution*. Berlin: de Gruyter. (https://doi.org/10.1515/9783110464108).

Holmberg, Anders. 2009. Null subject parameters. In Biberauer, Theresa & Holmberg, Anders & Roberts, Ian & Sheehan, Michelle (eds.), *Parametric variation*: *Null subjects in minimalist theory*, 88–124. Cambridge: Cambridge University Press.

Hopper, Paul. 1987. Stability and change in VN/NV alternating languages: A study in pragmatics and linguistic theory. In Bertuccelli Papi, Marcella & Verschueren, Jef (eds.), *The pragmatic perspective*, 455–476. Amsterdam: John Benjamins.

Hopper, Paul & Thompson, Sandra A. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–299.

Hothorn, Torsten & Zeileis, Achim. 2015. partykit: A modular toolkit for recursive partitioning in R. *Journal of Machine Learning* 16(118). 3905–3909. (`https://jmlr.org/papers/v16/hothorn15a.html`) (Accessed 2020-09-16).

Huang, Charles T.-J. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry* 15(4). 531–574.

Huang, Yan. 1994. *The syntax and pragmatics of anaphora: A study with special reference to Chinese*. Cambridge: Cambridge University Press.

Huang, Yan. 1995. On null subjects and null objects in generative grammar. *Linguistics* 33(6). 1081–1123.

Huang, Yan. 2000a. *Anaphora: A cross-linguistic study*. Oxford: Oxford University Press.

Huang, Yan. 2000b. Discourse anaphora: Four theoretical models. *Journal of Pragmatics* 32(2). 151–176.

Huddleston, Rodney D. & Pullum, Geoffrey K. (eds.). 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hunt, Kellogg W. 1977. Early blooming and late blooming syntactic structures. In Cooper, Charles R. & Odell, Lee (eds.), *Evaluating writing*: *Describing, measuring and judging*, 91–104. Urbana, IL: NCTE.

Ide, Nancy & Cristea, Dan. 2000. A hierarchical account of referential accessibility. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*.

Iida, Ryū & Komachi, Mamoru & Inui, Kentarō & Yuji, Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. *Proceedings of the 2007 Linguistic Annotation Workshop (LAW'07)*. 132–139. (`https://doi.org/10.5555/1642059.1642081`).

Izre'el, Shlomo. 2005. Intonation units and the structure of spontaneous spoken language: A view from Hebrew. In Auran, Cyril & Bertrand, Roxanne & Chanet, Catherine & Colas, Annie & Di Cristo, Albert & Portes, Cristel & Reynier, Alain & Vion, Monique (eds.), *Proceedings of the IDP05 International Symposium on Discourse–Prosody Interfaces*.

Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jacobsen, William H., Jr. 1967. Switch-reference in Hokan-Coahuiltecan. In Dell, Hymes H. & Bittle, William E. & Hoijer, Harry (eds.), *Studies in southwestern ethnolinguistics*, 238–263. The Hague: Mouton.

Jaeger, Florian T. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Unpublished Ph.D. dissertation, Stanford University.

Jaeger, Florian T. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62. (https://doi.org/10.1016/j.cogpsych.2010.02.002).

Jaeggli, Osvaldo & Safir, Ken. 1989. The null subject parameter and parametric theory. In Jaeggli, Osvaldo & Safir, Ken (eds.), *The null subject parameter*, 1–44. Dordrecht: Kluwer.

Jarbou, Samir O. & Migdadi, Fathi. 2012. Testing the limits of anaphoric distance in Classical Arabic: A corpus-based study. *Research in Language* 10(4). 423–444. (https://doi.org/10.2478/v10015-012-0003-y).

Jarvella, Robert J. 1971. Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behaviour* 10(4). 409–416. (https://doi.org/10.1016/S0022-5371(71)80040-3).

Johnson-Laird, Philip N. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Just, Erika & Čéplö, Slavomír. 2022. Differential object indexing in Maltese – a corpus based pilot study. In Turek, Przemyslaw & Nintemann, Julia (eds.), *Maltese: Contemporary changes and historical innovations*, 105–132. Berlin: Mouton de Gruyter. (https://doi.org/10.1515/9783110783834-005).

Kail, Michele. 1989. Cue validity, cue cost, and processing types in French sentence comprehension. In MacWhinney, Brian & Bates, Elizabeth (eds.), *The crosslinguistic study of sentence processing*, 77–117. Cambridge: Cambridge University Press.

Kameyama, Megumi. 1998. Intrasentential centering: A case study. In Walker, Marilyn A. & Joshi, Aravind K. & Prince, Ellen F. (eds.), *Centering Theory in discourse*, 89–112. Oxford: Clarendon.

Kamp, Hans. 1981. A theory of truth and semantic representation. In Groenendijk, Jeroen A. G. & Janssen, Theo M. V. & Stokhof, Martin J. B. (eds.), *Formal methods in the study of language*, 277–322. Amsterdam: Mathematisch Centrum.

Kamp, Hans & Reyle, Uwe. 1993. *From discourse to logic*. Dordrecht: Kluwer.

Kärkkäinen, Elise. 1996. Preferred argument structure and subject role in American English conversational discourse. *Journal of Pragmatics* 25(5). 675–701.

Karmiloff-Smith, Annette. 1981. The grammatical marking of thematic structure in the development of language production. In Deutsch, Werner (ed.), *The child's construction of language*, 121–148. London: Academic Press.

Karttunen, Lauri. 1976. Discourse referents. In McCawley, James D. (ed.), *Syntax and Semantics 7*: *Notes from the Linguistic Underground*, 363–385. New York: Academic Press.

Kawahara, Daisuke & Kurohashi, Sadao & Hashida, Kōichi. 2002. Construction of a Japanese relevance-tagged corpus. *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*. 495–498.

Keenan, Edward L. 1976. Towards a universal definition of "subject". In Li, Charles N. (ed.), *Subject and topic*, 303–333. New York: Academic Press.

Keenan, Edward L. & Comrie, Bernard. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8(1). 63–99.

Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.

Kehler, Andrew. 2004. Discourse coherence. In Horn, Laurence & Ward, Gregory (eds.), *Handbook of pragmatics*, 241–265. Malden, MA: Blackwell.

Kehler, Andrew. 2022. Coherence establishment as a source of explanation in linguistic theory. *Annual Review of Linguistics* 8. 123–142. (`https://doi.org/10.1146/annurev-linguistics-011619-030357`).

Kempf, Luise & Nübling, Damaris & Schmuck, Mirjam (eds.). 2020. *Die Linguistik der Eigennamen*. Berlin: Mouton de Gruyter.

Khamis, James. 2008. Measures of Association: How to choose? *Journal of Diagnostic Medical Sonography* 24(3). 155–162.

Khudiakova, Mariya V. & Dobrov, Grigory B. & Kibrik, Andrej A. & Loukachevitch, Natalia V. 2011. Computational modeling of referential choice: Major and minor referential options. *Proceedings of the PRE-CogSci-11 Workshop on the Production of Referring Expressions: Bridging the Gap Between Computational, Empirical, and Theoretical Approaches to Reference, Boston, USA, 20 July 2011*.

Kibrik, Andrej A. 1996. Anaphora in Russian narrative prose: A cognitive calculative account. In Fox, Barbara (ed.), *Studies in anaphora*, 253–303. Amsterdam: John Benjamins.

Kibrik, Andrej A. 1999. Cognitive inferences from discourse observations: Reference and working memory. In van Hoek, Karen & Kibrik, Andrej A. & Noordman, Leo (eds.), *Discourse studies in cognitive linguistics*: *Proceedings of the 5th International Cognitive Linguistics Conference*, 29–52. Amsterdam: John Benjamins.

Kibrik, Andrej A. 2000. A cognitive calculative approach towards discourse anaphora. In Baker, Paul & Hardie, Andrew & McEnery, Tony & Siewierska, Anna (eds.), *Proceedings from the 3rd Discourse Anaphora and Reference Resolution Colloquium (DAARC 2000)*, 72–82. Lancaster: Lancaster University Centre for Computer Corpus Research on Language.

Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.

Kibrik, Andrej A. & Khudyakova, Mariya V. & Dobrov, Grigory B. & Linnik, Anastasia S. & Zalmanov, Dmitrij A. 2016. Referential choice: Predictability and its limits. *Frontiers in Psychology* 7(1429). (https://doi.org/10.3389/fpsyg.2016.01429).

Kibrik, Andrej A. & Khudyakova, Mariya & Dobrov, Grigory B. & Linnik, Anastasia S. 2013. Referential choice: A cognitively based modeling study. *Proceedings of the PRE-CogSci-13 Workshop on the Production of Referring Expressions: Bridging the Gap between Cognitive and Computational Approaches to Reference, Berlin, Germany, 31 July 2013*.

Kibrik, Andrej A. & Krasavina, Olga N. 2005. A corpus study of referential choice: The role of rhetorical structure. *Proceedings of DIALOG'05*.

Kimoto, Yukinori. 2019. Multi-CAST Arta. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#arta) (Accessed 2019-07-05).

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit* 6. 79–86.

Krahmer, Emiel J. & van Deemter, Kees. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics* 38(1). 173–218.

Krahmer, Emiel & Theune, Mariët. 2002. Efficient context-sensitive generation of referring expressions. In van Deemter, Kees & Kibble, Rodger (eds.), *Information sharing*: *Givenness and newness in language processing*, 223–264. Stanford, CA: CSLI Publications.

Krahmer, Emiel & Theune, Mariët & Viethen, Jette & Hendrickx, Iris. 2008. GRAPH: The costs of redundancy in referring expressions. *Proceedings of the 5th International Natural Language Generation Conference (INLG 2008)*. 227–229.

Krasavina, Olga N. & Chiarcos, Christian. 2007. PoCoS: Potsdam Coreference Scheme. *Proceedings of the Linguistic Annotation Workshop, Prague, June 2007*. 156–163.

Krifka, Manfred & Musan, Renate (eds.). 2012. *The expression of information structure*. Berlin: Mouton de Gruyter.

Krug, Manfred & Lucas, Christopher. 2018. Definite article (omission) in British, Maltese, and other Englishes. *Language Typology and Universals (STUF)* 71(2). 261–305. (https://doi.org/10.1515/stuf-2018-0012).

Kuhn, Max & Johnson, Kjell. 2013. *Applied predictive modeling*. New York: Springer.

Kumagai, Yoshiharu. 2006. Information management in intransitive subjects: Some implications for the preferred argument structure theory. *Journal of Pragmatics* 38(5). 670–694.

Kwon, Nayoung & Lee, Yoonhyoung & Gordon, Peter C. & Kluender, Robert. 2010. Cognitive and linguistic factors affecting the subject/object asymmetry: An eye-

tracking study of prenominal relative clauses in Korean. *Language* 86(3). 546–582. (https://doi.org/10.1353/lan.2010.0006).

Labov, William. 1994. *Principles of linguistic change: Internal factors*. Malden, MA: Blackwell.

Labov, William & Waletzky, Joshua. 1967. Narrative analysis. In Helm, June (ed.), *Essays on the verbal and visual arts*, 12–44. Seattle, WA: University of Washington Press.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*. Cambridge: Cambridge University Press.

Lappin, Shalom & Leass, Herbert J. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4). 535–561.

Lascarides, Alex & Asher, Nicholas. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy* 16(5). 437–493.

Lehmann, Christian. 1982. Directions for interlinear morphemic translations. *Folia Linguistica* 16(2). 193–224.

Lehmann, Christian. 1988. Towards a typology of clause linkage. In Haiman, John & Thompson, Sandra A. (eds.), *Clause combining in grammar and discourse*, vol. 18, 181–225. Amsterdam: John Benjamins.

Leroux, Martine & Jarmasz, Lidia-Gabriela. 2005. A study about nothing: Null subjects as a diagnostic of the convergence between English and French. *University of Pennsylvania Working Papers in Linguistics* 12(2). 1–14.

Levinson, Stephen C. 1987. Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics* 23(2). 379–434.

Levinson, Stephen C. 1991. Pragmatic reduction of the binding conditions revisited. *Journal of Linguistics* 27(1). 107–161.

Levshina, Natalia. 2016. Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast* 16(1). 84–117. (https://doi.org/10.1075/lic.16.1.04lev).

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Languages in Contrast* 23(3). 533–572. (https://doi.org/10.1515/lingty-2019-0025).

Levshina, Natalia. 2021. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*. (https://doi.org/10.1515/lingty-2020-0118).

Levy, Roger & Jaeger, Florian T. 2007. Speakers optimize information density through syntactic reduction. In Schölkopf, Bernhard & Platt, John & Hoffmann, Thomas (eds.), *Advances in neural information processing systems 19*: *Proceedings of the 2006 Conference*, 849–856. Cambridge, MA: MIT Press.

Li, Charles N. & Thompson, Sandra A. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley, CA: University of California Press.

Li, Xiaoshi & Bayley, Robert. 2018. Lexical frequency and syntactic variation: Subject pronoun use in Mandarin Chinese. *Asia-Pacific Language Variation* 4(2). 135–160. (https://doi.org/10.1075/aplv.17005.li).

Lichtenberk, František. 1996. Patterns of anaphora in To'aba'ita narrative discourse. In Fox, Barbara (ed.), *Studies in anaphora*, 379–411. Amsterdam: John Benjamins.

Linnik, Anastasia S. & Dobrov, Grigory B. 2011. Protagonism as a factor affecting referential choice in discourse. *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, 6–7 October 2011*.

Lockwood, Hunter & Macaulay, Monica. 2012. Prominence hierarchies. *Language and Linguistics Compass* 6(7). 431–446.

Loukachevitch, Natalia V. & Dobrov, Grigory B. & Kibrik, Andrej A. & Khudiakova, Mariya V. & Linnik, Anastasia S. 2011. Factors of referential choice: Papers from Annual Internation Conference "Dialogue" (2011). In Kibrik, Alexander E. (ed.), 458–467. Moscow: RGGU.

Lowder, Matthew W. & Maxfield, Nathan D. & Ferreira, Fernanda. 2018. Processing of self-repairs in stuttered and non-stuttered speech. *Language, Cognition and Neuroscience* 35(1). 93–105. (https://doi.org/10.1080/23273798.2019.1628284).

Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4(226). (https://doi.org/10.3389/fpsyg.2013.00226).

MacWhinney, Brian & Bates, Elizabeth. 1978. Sentential devices for conveying givenness and newness: A cross-cultural developmental study. *Journal of Verbal Learning and Verbal Behaviour* 17(5). 539–558.

Mahmoudveysi, Parwin & Bailey, Denise & Paul, Ludwig & Haig, Geoffrey. 2012. *The Gorani language of Gawraju (Gawrajuyi), a village of West Iran: Texts, grammar, and lexicon* (Beiträge zur Iranistik 35). Wiesbaden: Reichert.

Malchukov, Andrej L. & Haspelmath, Martin & Comrie, Bernard. 2010. Ditransitive constructions: A typological overview. In Malchukov, Andrej L. & Haspelmath, Martin & Comrie, Bernard (eds.), *Studies in ditransitive constructions*: *A comparative handbook*, 1–60. Berlin: Mouton de Gruyter.

Mann, William C. & Matthiessen, Christian M. I. M. & Thompson, Sandra A. 1992. Rhetorical structure theory and text analysis. In Mann, William C. & Thompson, Sandra A. (eds.), *Discourse description*: *Diverse linguistic analyses of a fundraising text*, 39–78. Amsterdam: John Benjamins.

Mansfield, John & Stoll, Sabine & Bickel, Bathasar. 2020. Category clustering: A probabilistic bias in the morphology of verbal agreement marking. *Language* 96(2). 255–293. (https://doi.org/10.1353/lan.2020.0021).

Margetts, Anna. 2007. Three-participant events in the languages of the world: Towards a crosslinguistic typology. *Linguistics* 45(3). 393–451.

Marslen-Wilson, William & Levy, Elena & Tyler, Lorraine K. 1982. Producing interpretable discourse: The establishment and maintenance of reference. In Jarvella, Robert J. & Klein, Wolfgang (eds.), *Speech, place, and action*: *Studies in deixis and related topics*, 339–378. Chichester: Wiley.

Matras, Yaron. 1997. Clause combining, ergativity, and coreferent deletion in Kurmanji. *Studies in Language* 21(3). 613–653. (https://doi.org/10.3758/s13414-015-0899-0).

Matthiessen, Christian M. I. M. & Thompson, Sandra A. 1988. The structure of discourse and 'subordination'. In Haiman, John & Thompson, Sandra A. (eds.), *Clause combining in grammar and discourse*, vol. 18, 275–329. Amsterdam: John Benjamins.

Mayer, Mercer. 1969. *Frog, where are you?* New York: Dial Books for Young Readers.

McCoy, Kathleen F. & Strube, Maria. 1999. Generating anaphoric expressions: Pronoun or definite description? *Proceedings of the ACL workshop on Discourse and Reference Structure*. 63–71.

McDaniel, Dana & McKee, Cecile & Cowart, Wayne & Garret, Merrill F. 2015. The role of the language production system in shaping grammars. *Language* 91(2). 415–441. (https://doi.org/10.1353/lan.2015.0021).

McEnery, Tony. 1995. *Computational pragmatics*. Ph.D. dissertation, Lancaster University.

McEnery, Tony & Tanaka, Izumi & Botley, Simon P. 1997. Corpus annotation and reference resolution. In Mitkov, Ruslan & Boguraev, Branimir (eds.), *ANARESOLUTION '97*: *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 57–74. Stroudsburg, PA: Association for Computational Linguistics.

McKee, Rachel & Schembri, Adam & McKee, David & Johnston, Trevor. 2011. Variable 'subject' presence in Australian Sign Language and New Zealand Sign Language. *Language Variation and Change* 21(3). 297–317.

McLuhan, Marshall. 1964. *Understanding media: The extension of man*. New York: New American Library.

McNamara, Danielle S. & Kintsch, Walter. 1996. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes* 22(3). 247–288. (https://doi.org/10.1080/01638539609544975).

Menardi, Giovanna & Torelli, Nicola. 2012. Training and assessing classification rules with unbalanced data. *Data Mining and Knowledge Discovery* 28. 92–122. (https://doi.org/10.1007/s10618-012-0295-5).

Meng, Chenxi. 2014. *Recordings of Tulil (2012–2014). Collection CM2 at PARADISEC*. (https://doi.org/10.4225/72/5af5be0adb4f9).

Meng, Chenxi. 2018. *A grammar of Tulil*. Ph.D. dissertation, La Trobe University.

Meng, Chenxi. 2019. Multi-CAST Tulil. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#tulil) (Accessed 2019-07-05).

Mettouchi, Amina & Vanhove, Martine & Caubet, Dominique (eds.). 2015. *Corpus-based studies of lesser-described languages: The CorpAfroAs corpus of spoken AfroAsiatic languages* (Studies in Corpus Linguistics 68). Amsterdam: John Benjamins.

Meyerhoff, Miriam. 2009. Replication, transfer, and calquing: Using variation as a tool in the study of language contact. *Language Variation and Change* 21(3). 297–317.

Meylan, Stephan C. 2018. *Representing linguistic knowledge with probabilistic models*. Ph.D. dissertation, University of California.

Michaelis, Laura & Hartwell, Francis. 2007. Lexical subjects and the conflation strategy. In Hedberg, Nancy & Zacharski, Ron (eds.), *The grammar-pragmatics interface*: *Eassays in honor of Jeanette K. Gundel*, 19–48. Amsterdam: John Benjamins.

Miller, Jim. 1988. Does spoken language have sentences? In Palmer, Frank R. (ed.), *Grammar and meaning*: *Essays in honour of Sir John Lyons*, 116–135. Cambridge: Cambridge University Press.

Miltsakaki, Eleni. 2002. Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics* 28(3). 319–355.

Miltsakaki, Eleni. 2003. *The syntax–discourse interface: Effects of the main–subordinate distinction on attention structure*. Ph.D. dissertation, University of Pennsylvania.

Mithun, Marianna. 2003. Pronouns and agreement: The information status of pronominal affixes. *Transactions of the Philological Society* 101(2). 235–278.

Mithun, Marianne. 1991. The role of motivation in the emergence of grammatical categories: The grammaticalization of subjects. In Traugott, Elizabeth C. & Heine, Bernd (eds.), *Approaches to grammaticalization*, 159–194. Amsterdam: John Benjamins.

Mithun, Marianne. 1996. Prosodic cues to accessibility. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 223–234. Amsterdam: John Benjamins.

Mithun, Marianne. 2015. Discourse and grammar. In Hamilton, Heidi & Schiffrin, Deborah & Tannen, Deborah (eds.), *Handbook of discourse analysis*, 9–41. Malden, MA: Blackwell.

Mitkov, Ruslan. 2000. Pronoun resolution: The practical alternative. In Botley, Simon P. & McEnery, Tony (eds.), *Corpus-based and computational approaches to discourse anaphora*, 128–142. Amsterdam: John Benjamins.

Mitkov, Ruslan. 2002. *Anaphora resolution*. London: Longman.

Mitkov, Ruslan & Barbu, Catalina. 2001. Evaluation tool for rule-based anaphora resolution methods. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01), Toulouse, France*.

Moosegard Hanse, Maj-Britt. 1998. *The function of discourse particles: A study with special reference to spoken Standard French*. Amsterdam: John Benjamins.

Morrow, Daniel G. 1985. Prominent characters and events organize narrative understanding. *Journal of Memory and Language* 24(3). 304–319.

Mosel, Ulrike. 2019. A multifunctional Teop-English dictionary. *Dictionaria* 4(1-6488). (`https://doi.org/10.5281/zenodo.3257580`). (`https://dictionaria.clld.org/contributions/teop`) (Accessed 2019-08-19).

Mosel, Ulrike & Schnell, Stefan. 2015. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (`https://multicast.aspra.uni-bamberg.de/#teop`) (Accessed 2019-03-08).

Mosel, Ulrike & Thiesen, Yvonne. 2007. *The Teop sketch grammar*. Unpublished manuscript, University of Kiel. (`https://hdl.handle.net/1839/00-0000-0000-0008-24F6-3@view`) (Accessed 2016-05-14).

Myachykov, Andriy & Posner, Michael I. 2005. Attention in language. In Itti, Laurent & Rees, Geraint & Tsotsos, John K. (eds.), *Neurobiology of Attention*, 324–329. Amsterdam: Elsevier.

Myhill, John. 1997. *Viewpoint, sequencing, and pronoun usage in Javanese short stories*. Guy, Gregory R. & Feagin, Crawford & Schiffrin, Deborah & Baugh, John (eds.). Amsterdam: John Benjamins. 237–258.

Nariyama, Shigeko. 2003. *Ellipsis and reference tracking in Japanese*. Amsterdam: John Benjamins.

Navarretta, Constanza. 2011. Antecedent and referent types of abstract pronominal anaphora. *Bochumer Linguistische Arbeitsberichte* 3. 99–110. (Accessed 2017-02-04).

Neeleman, Ad & Szendröi, Kriszta. 2007. Radical pro drop and the morphology of pronouns. *Linguistic Inquiry* 38(4). 671–714.

Nicolae, Cristina & Nicolae, Gabriel & Roberts, Kirk. 2010. C-3: Coherence and Coreference Corpus. In Calzolari, Nicoletta and Choukri, Khalid and Maegaard, Bente and Mariani, Joseph and Odijk, Jan and Piperidis, Stelios and Rosner, Mike and

Tapias, Daniel (ed.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, 136–143. Valetta: European Language Resources Association.

Nicolis, Marco. 2008. The null subject parameter and correlating properties: The case of creole languages. In Biberauer, Theresa (ed.), *The limits of syntactic variation*, 271–294. Amsterdam: John Benjamins. (http://www18.georgetown.edu/data/people/mn256/publication-32861.pdf) (Accessed 2015-08-16).

Noonan, Michael. 2003. *A crosslinguistic investigation of referential density*. Unpublished manuscript, University of Wisconsin-Milwaukee. (http://crossasia-repository.ub.uni-heidelberg.de/190/) (Accessed 2016-02-08).

O'Grady, William. 1997. *Syntactic development*. Chicago, IL: University of Chicago Press.

Öpengin, Ergin & Haig, Geoffrey. 2014. Regional variation in Kurmanji: A preliminary classification of dialects. *Kurdish Studies* 2(2). 143–176.

Orita, Naho & Vornov, Eliana & Feldman, Naomi H. & Daumé, Hal, III. 2015. Why discourse affects speakers' choice of referring expressions. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015*. 1639–1649.

Osgood, Charles E. 1971. Where do sentences come from? In Steinberg, Danny D. & Jakobovits, Leon A. (eds.), *Semantics*: *An interdisciplinary reader in philosophy, linguistics and psychology*, 497–529. Cambridge: Cambridge University Press.

Owens, Jonathan & Dodsworth, Robin & Kohn, Mary. 2013. Subject expression and discourse embeddedness in Emirati Arabic. *Language Variation and Change* 25(2). 255–285.

Ozerov, Pavel. 2018. Tracing the sources of information structure: Towards the study of interactional management of information. *Journal of Pragmatics* 138(1). 77–97. (https://doi.org/10.1016/j.pragma.2018.08.017).

Paschen, Ludger & Delafontaine, François & Draxler, Christoph & Fuchs, Susanne & Stave, Matthew & Seifart, Frank. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'20), Marseille, France, 13–16 May 2020*. 2657–2666.

Passonneau, Rebecca & Litman, Diane J. 1997. Discourse segmentation by human and automated means. *Computational Linguistics* 23(1). 103–139. (https://doi.org/10.5555/972684.972689).

Payne, Doris L. 1987. Information structuring in Papago narrative discourse. *Language* 63(4). 783–804.

Payne, John & Pullum, Geoffrey K. & Scholz, Barbara C. & Berlage, Eva. 2013. Anaphoric *one* and its implications. *Language* 89(4). 794–829.

Payne, Thomas E. 1993. *The Twins stories: Participant coding in Yagua narrative*. Los Angeles: University of California Press.

Perlmutter, David. 1971. *Deep and surface constraints in syntax*. New York: Holt, Rinehart and Winston.

Piaget, Jean. 1955. *The language and thought of the child*. New York: World.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21. 1112–1130. (https://doi.org/10.3758/s13423-014-0585-6).

Piantadosi, Steven T. & Tily, Harry J. & Gibson, Edward. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526–3529. (https://doi.org/10.1073/pnas.1012551108).

Poesio, Massimo. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece, 31 May–2 June 2000*.

Poesio, Massimo. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In Webber, Bonnie L. and Byron, Donna (ed.), *DiscAnnotation '04*: *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, 72–79. Stroudsburg, PA: Association for Computational Linguistics.

Poesio, Massimo & Artstein, Ron. 2008. Anaphoric annotation in the ARRAU corpus. In Calzolari, Nicoletta and Choukri, Khalid and Maegaard, Bente and Mariani, Joseph and Odijk, Jan and Piperidis, Stelios and Tapias, Daniel (ed.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association.

Poesio, Massimo & di Eugenio, Barbara. 2001. Discourse structure and anaphoric accessibility. In Kruijff-Korbayová, Ivana & Steedman, Mark (eds.), *Proceedings of ESSLLI 2001*: *Workshop on Information Structure and Discourse Semantics*, 129–144. Helsinki, Finland.

Poesio, Massimo & Mehta, Rahul & Maroudas, Axel & Hitzeman, Janet. 2004. Learning to resolve bridging references. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*. 143–150. (https://doi.org/10.3115/1218955.1218974).

Polanyi, Livia. 1995. *The linguistic structure of discourse*. Stanford, CA: CSLI Publications.

Polanyi, Livia & van den Berg, Martin & Ahn, David D. 2003. Discourse structure and sentential information structure: An initial proposal. *Journal of Logic, Language and Information* 12(3). 337–350.

Portele, Yvonne & Bader, Markus. 2016. Accessibility and referential choice: Personal pronouns and d-pronouns in written German. *Discours* 18. 1–41. (https://doi.org/10.4000/discours.9188).

Prasad, Rashmi & Lee, Alan & Dinesh, Nikhil & Miltsakaki, Eleni & Campion, Geraud & Joshi, Aravind & Webber, Bonnie. 2008. *Penn Discourse Treebank Verson 2.0 (LDC2008T05)*. Philadelpha, PA: Linguistic Data Consortium. (https://doi.org/10.35111/nbvh-1n26).

Prasad, Rashmi & Webber, Bonnie & Lee, Alan & Joshi, Aravind. 2019. *Penn Discourse Treebank Verson 3.0 (LDC2019T05)*. Philadelpha, PA: Linguistic Data Consortium. (https://doi.org/10.35111/qebf-gk47).

Prince, Ellen F. 1981a. On the inferencing of indefinite-*this* NPs. In Joshi, Aravind K. & Webber, Bonnie L. & Sag, Ivan A. (eds.), *Elements of discourse understanding*, 231–250. Cambridge: Cambridge University Press.

Prince, Ellen F. 1981b. Toward a taxonomy of given-new information. In Cole, Peter (ed.), *Radical pragmatics*, 223–255. New York: Academic Press.

Prince, Ellen F. 1992. The ZPG letter: Subjects, definiteness, and information-status. In Mann, William C. & Thompson, Sandra A. (eds.), *Discourse description*: *Diverse linguistic analyses of a fund-raising text*, vol. 16, 295–325. Amsterdam: John Benjamins.

Ridgeway, Greg. 2020. *Generalized boosted models: A guide to the gbm package*. (https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf) (Accessed 2020-09-16).

Riesberg, Sonja. 2014. *Symmetrical voice and linking in Western Austronesian languages*. Berlin: Mouton de Gruyter.

Riesberg, Sonja & Shiohara, Asako & Utsumi, Atsuko. 2018. *Perspectives on information structure in Austronesian languages*. Berlin: Language Science Press.

Riester, Arndt & Baumann, Stefan. 2014. *RefLex scheme: Annotation guidelines*. Unpublished manuscript, University of Stuttgart / University of Cologne. (https://doi.org/10.18419/opus-9011). (Accessed 2018-03-03).

Riester, Arndt & Baumann, Stefan. 2017. *The RefLex scheme — Annotation guidelines* (SinSpeC: Working papers of the SFB 732 14). Stuttgart: University of Stuttgart. (http://elib.uni-stuttgart.de/handle/11682/9028) (Accessed 2018-03-01).

Ross, Malcolm D. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra: Pacific Linguistics.

Rupp, Laura & Tagliamonte, Sali A. 2019. "They used to follow Ø river": The zero article in York English. *Journal of English Linguistics* 47(4). 279–300. (https://doi.org/10.1177/0075424219865933).

Sachs, Jacqueline S. 1967. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psycholinguistics* 2(9). 437–442.

Sacks, Harvey & Schegloff, Emanuel A. & Jefferson, Gail. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language* 50(4). 696–735. (https://doi.org/10.2307/412243).

San Roque, Lila & Rumsey, Alan & Gawne, Lauren & Spronck, Stef & Hoenigman, Darja & Carroll, Alice & Miller, Julia & Evans, Nicholas. 2012. Getting the story straight: Language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation* 6. 134–173.

Sanford, Anthony J. & Garrod, Simon. 1981. *Understanding written language: Explorations of comprehension beyond the sentence*. Chichester: Wiley.

Schiborr, Nils N. 2015. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#english) (Accessed 2016-02-28).

Schiborr, Nils N. 2017. *Antecedent distance and the accessibility hierarchy: A quantitative approach*. Master's thesis, University of Bamberg.

Schiborr, Nils N. 2018. multicastR: A companion to the Multi-CAST collection. R package version 2.0.0. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://cran.r-project.org/package=multicastR) (Accessed 2020-02-22).

Schiborr, Nils N. 2019. Multi-CAST collection overview. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/) (Accessed 2019-07-05).

Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines (v1.1)*. Unpublished manuscript, University of Bamberg. (https://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2019-03-08).

Schiering, René & Bickel, Balthasar & Hildebrandt, Kristine A. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics* 46(3). 657–709. (https://doi.org/10.1017/S0022226710000216).

Schlücker, Barbara & Ackermann, Tanja. 2017. The morphosyntax of proper names: An overview. *Folia Linguistica* 51(2). 309–339.

Schnell, Stefan. 2011. *A grammar of Vera'a*. Ph.D. dissertation, University of Kiel. (https://www.academia.edu/2317752/Schnell_00002011_A_grammar_of_Veraa_an_Oceanic_language_of_North_Vanuatu) (Accessed 2016-02-22).

Schnell, Stefan. 2015. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#veraa) (Accessed 2019-03-08).

Schnell, Stefan & Barth, Danielle. 2018. Discourse motivations for pronominal and zero objects across genres in Vera'a. *Language Variation and Change* 30(1). 51–81. (https://doi.org/10.1017/S0954394518000054).

Schnell, Stefan & Barth, Danielle. 2020. Expression of anaphoric subjects in Vera'a: Functional and structural factors in the choice between pronoun and zero. *Language Variation and Change*.

Schnell, Stefan & Haig, Geoffrey & Schiborr, Nils N. & Vollmer, Maria C. 2020. *Introducing new referents: A corpus-based cross-linguistic perspective*. Paper presented at the 53rd Annual Meeting of the Societas Linguistica Europaea (SLE 2020), Bucharest, Romania, 26 August–1 September 2020.

Schnell, Stefan & Haig, Geoffrey & Seifart, Frank. 2021a. The role of language documentation in corpus-based typology. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 1–28. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74656).

Schnell, Stefan & Schiborr, Nils N. 2018. Corpus-based typological research in discourse and grammar: GRAID and Multi-CAST. *Asian and African Languages and Linguistics* 12. 1–16. (http://hdl.handle.net/10108/91145).

Schnell, Stefan & Schiborr, Nils N. 2022. Crosslinguistic corpus studies in linguistic typology. *Annual Review of Linguistics* 8. 171–191. (https://doi.org/10.1146/annurev-linguistics-031120-104629).

Schnell, Stefan & Schiborr, Nils N. & Haig, Geoffrey. 2021b. Efficiency in discourse processing: Does morphosyntax adapt to accommodate new referents? *Linguistics Vanguard* 7(s3). (https://doi.org/10.1515/lingvan-2019-0064).

Schwenter, Scott. 2014. Two kinds of object marking in Portuguese and Spanish. In Amaral, Patrícia & Carvalho, Ana M. (eds.), *Portuguese–Spanish interfaces*: *Diachrony, synchrony, and contact*, 237–260. Amsterdam: John Benjamins.

Schwenter, Scott. 2016. Null objects across South America. In Face, Timothy L. & Klee, Carol A. (eds.), *Selected proceedings of the 8th Hispanic Linguistics Symposium*, 23–36. Somerville, MA: Cascadilla Proceedings Project.

Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Seifart, Frank. 2021. Combining documentary linguistics and corpus phonetics to advance corpus-based typology. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 115–139. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74659).

Seifart, Frank & Paschen, Ludger & Stave, Matthew (eds.). 2022. *Language Documentation Reference Corpus (DoReCo) 1.0*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft and Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). (https://doi.org/10.34847/nkl.7cbfq779).

Seifart, Frank & Strunk, Jan & Danielsen, Swintha & Hartmann, Iren & Pakendorf, Brigitte & Wichmann, Søren & Witzlack-Makarevich, Alena & de Jong, Nivja H. & Bickel, Balthasar. 2018. Nouns slow down speech across structurally and

culturally diverse languages. *PNAS* 115(22). 5720–5725. (https://doi.org/10.1073/pnas.1800708115).

Shor, Leon. 2020. *Third person human reference in Israeli Hebrew Conversation*. Ph.D. dissertation, Tel Aviv University.

Siewierska, Anna. 2004. *Person*. Oxford: Oxford University Press.

Silva-Corvalán, Carmen. 2001. *Sociolingüística y pragmática del español* [Sociolinguistics and pragmatics of Spanish]. Washington, D.C.: Georgetown University Press.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Dixon, R. M. W. (ed.), *Grammatical categories in Australian languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Skirgård, Hedvig & Haynie, Hannah J. & Blasi, Damián E. & Hammarström, Harald & alii. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). (https://doi.org/{10.1126/sciadv.adg6175}).

Song, Sanghoun. 2017. *Modeling information structure in a cross-linguistic perspective*. Berlin: Language Science Press.

Sridhar, Shikaripur N. 1988. *Cognition and sentence production: A cross-linguistic study*. New York: Springer.

Stede, Manfred & Neumann, Arne. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In Calzolari, Nicoletta and Choukri, Khalid and Declerck, Thierry and Loftsson, Hrafn and Maegaard, Bente and Mariani, Joseph and Moreno, Asuncion and Odijk, Jan and Piperidis, Stelios (ed.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association.

Steinberger, Ralf & Pouliquen, Bruno & Widiger, Anna & Ignat, Camelia & Erjavec, Tomaž & Tufiş, Dan & Varga, Dániel. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy, 24–26 May 2006*.

Stevenson, Rosemary J. 1996. Mental models, propositions and the comprehension of pronouns. In Oakhill, Jane & Garnham, Alan (eds.), *Mental models in cognitive science*, 53–76. London: Psychology Press.

Stevenson, Rosemary J. & Crawley, Rosalind A. & Kleinman, David. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes* 9(4). 519–548. (https://doi.org/10.1080/01690969408402130).

Stevenson, Rosemary J. & Knott, Alistair & Oberlander, Jon & McDonald, Sharon. 2000. Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes* 15(3). 225–262. (https://doi.org/10.1080/016909600386048).

Stewart, Andrew J. & Pickering, Martin J. & Sanford, Anthony J. 2000. The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language* 42(3). 423–443. (https://doi.org/10.1006/jmla.1999.2691).

Stoll, Sabine & Bickel, Balthasar. 2009. How deep are differences in referential density? In Guo, Jiansheng & Lieven, Elena & Budwig, Nancy & Ervin-Tripp, Susan & Nakamura, Keiko & Özçalışkan, Şeyda (eds.), *Crosslinguistic approaches to the psychology of language*: *Research in the tradition of Dan Isaac Slobin*, 543–555. London: Psychology Press.

Stolz, Thomas. 2007. *Harry Potter* meets *Le Petit Prince*: On the usefulness of parallel corpora for crosslinguistic investigations. *STUF* 60(2). 100–117.

Strobl, Carolin & Malley, James & Tutz, Gerhard. 2009. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods* 14(4). 323–348.

Strube, Michael & Wolters, Maria. 2000. A probabilistic genre-independent model of pronominalization. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000)*. 18–25.

Sun, Chaofen. 2006. *Chinese: A linguistic introduction*. Cambridge: Cambridge University Press.

Taboada, Maite. 2008. Reference, centers and transitions in spoken Spanish. In Gundel, Jeanette K. & Hedberg, Nancy (eds.), *Reference and reference processing*, 176–215. Oxford: Oxford University Press.

Taboada, Maite & Hadic Zabala, Loreley. 2008. Deciding on units of analysis within Centering Theory. *Corpus Linguistics and Linguistic Theory* 4(1). (https://doi.org/10.1515/CLLT.2008.003).

Taboada, Maite & Mann, William C. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies* 8(3). 423–459. (https://doi.org/10.1177/1461445606061881).

Tanenhaus, Michael K. & Carlson, Greg N. & Seidenberg, Mark S. 1985. Do listeners compute linguistic representations? In Dowty, David R. & Karttunen, Lauri & Zwicky, Arnold M. (eds.), *Natural language parsing*, 359–408. Cambridge: Cambridge University Press.

Tetreault, Joel R. 2005. Decomposing discourse. In Branco, António & McEnery, Tony & Mitkov, Ruslan (eds.), *Anaphora processing*: *Linguistic, cognitive and computational modelling*, 74–95. Amsterdam: John Benjamins.

Tetreault, Joel R. & Allen, James. 2003. An empirical evaluation of pronoun resolution and clausal structure. *Proceedings of the 2003 International Symposium on Reference Resolution and its Applications to Question Answering and Summarization*. 1–8.

Therneau, Terry & Atkinson, Beth. 2019. *rpart: Recursive partitioning and regression trees. R package version 4.1-15.* (http://CRAN.R-project.org/package=rpart) (Accessed 2020-09-16).

Thieberger, Nick. 1995. *South Efate (Vanuatu). Collection NT1 at PARADISEC*. (https://doi.org/10.4225/72/56E97595B6D0A).

Thieberger, Nick. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Austin, Peter (ed.), *Language documentation and description*, 169–178. London: Hans Rausing Endangered Languages Project, SOAS.

Thieberger, Nick. 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu: University of Hawaii Press. (http://hdl.handle.net/11343/31242) (Accessed 2019-07-31).

Thieberger, Nick & Brickell, Timothy. 2019. Multi-CAST Nafsan. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#nafsan) (Accessed 2019-07-31).

Thompson, Sandra A. 1987. 'Subordination' and narrative event structure. In Tomlin, Russell S. (ed.), *Coherence and grounding in discourse*, 435–454. Amsterdam: John Benjamins.

Thompson, Sandra A. & Couper-Kuhlen, Elizabeth. 2005. The clause as a locus of grammar and interaction. *Discourse Studies* 7(4–5). 481–505. (https://doi.org/10.1177/1461445605054403).

Tily, Harry & Piantadosi, Steven. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. In van Deemter, Kees & Gatt, Albert & van Gompel, Roger P. G. & Krahmer, Emiel J. (eds.), *Proceedings of the workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009)*. Amsterdam: University of Tilburg.

Tomlin, Russell. 1987a. Linguistic reflections on cognitive events. In Tomlin, Russell S. (ed.), *Coherence and grounding in discourse*, 455–479. Amsterdam: John Benjamins.

Tomlin, Russell S. (ed.). 1987b. *Coherence and grounding in discourse*. Amsterdam: John Benjamins.

Toole, Janine. 1996. The effect of genre on referential choice. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 263–290. Amsterdam: John Benjamins.

Torres Cacoullos, Rena & Travis, Catherine E. 2014. Prosody, priming and particular constructions: The pattern of English first-person singular subject expression in conversation. *Journal of Pragmatics* 63. 19–34.

Torres Cacoullos, Rena & Travis, Catherine E. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3). 653–692.

Trabasso, Tom & Rollins, Howard & Shaughnessy, Edward. 1971. Storage and verification stages in processing concepts. *Cognitive Psychology* 2(3). 239–289. (https://doi.org/10.1016/0010-0285(71)90014-4).

Travis, Catherine E. & Lindstrom, Amy M. 2016. Different registers, different grammars? Subject expression in English conversation and narrative. *Language Variation and Change* 28(1). 103–128. (https://doi.org/10.1017/S0954394515000174).

Travis, Catherine E. & Torres Cacoullos, Rena. 2018. Discovering structure: Person and accessibility. In Lapidus Shin, Naomi & Erker, Daniel (eds.), *Questioning theoretical primitives in linguistic inquiry*: *Paper in honor of Ricardo Otheguy*, 67–90. Amsterdam: John Benjamins.

Tutin, Agnès & Viegas, Evelyne. 2000. Generating coreferential anaphoric definite NPs. In Botley, Simon P. & McEnery, Tony (eds.), *Corpus-based and computational approaches to discourse anaphora*, 227–248. Amsterdam: John Benjamins.

van Deemter, Kees & Gatt, Albert & van Gompel, Roger P. G. & Krahmer, Emiel J. (eds.). 2009. *Proceedings of the workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009)*. Amsterdam: University of Tilburg.

van Deemter, Kees & Gatt, Albert & van Gompel, Roger P. G. & Krahmer, Emiel J. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science* 4(2). 166–183. (https://doi.org/10.1111/j.1756-8765.2012.01187.x).

van Deemter, Kees & Kibble, Rodger. 1999. What is coreference, and what should coreference annotation be? *Proceedings of the ACL'99 Workshop on Coreference and its Applications*. 90–96.

van Gijn, Rik & Hammond, Jeremy & Matić, Dejan & van Putten, Saskia & Vilacy Galucio, Ana (eds.). 2014. *Information structure and reference tracking in complex sentences*. Amsterdam: John Benjamins.

Van Valin, Robert D., Jr. 2005. *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press.

Van Valin, Robert D., Jr. & LaPolla, Randy J. 1997. *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press.

Vollmer, Maria. 2020. Multi-CAST Mandarin. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#mandarin) (Accessed 2020-01-03).

Vonk, Wietske & Hustinx, Lettica G. M. M. & Simons, Wim H. G. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes* 7(3/4). 301–333.

Wälchli, Bernhard. 2009. *Motion events in parallel texts: A study in primary data typology*. Unpublished Habilitationsschrift, University of Bern.

Walker, James A. & Dunn, Michael & Markussen-Daval, Aymeric & Meyerhoff, Miriam. 2015. *Modelling the speech community through multiple variables: Trees, networks and clades*. Poster presented at New Ways of Analyzing Variation (NWAV 44), Toronto, Canada, 22–25 October 2015.

Walker, Marilyn A. & Joshi, Aravind K. & Prince, Ellen F. (eds.). 1998. *Centering Theory in discourse*. Oxford: Clarendon.

Walker, Marilyn & Iida, Masayo & Cote, Sharon. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20(2). 193–232.

Wang, Luming & Schlesewsky, Matthias & Bickel, Balthasar & Bornkessel-Schlesewsky, Ina. 2009. Exploring the nature of the 'subject'-preference: Evidence from online comprehension of simple sentences in Mandarin Chinese. *Language and Cognitive Processes* 24(7–8). 1180–1226. (https://doi.org/10.1080/01690960802159937).

Ward, Gregory & Birner, Betty J. 2004. Information structure and non-canonical syntax. In Horn, Laurence & Ward, Gregory (eds.), *Handbook of pragmatics*, 153–174. Malden, MA: Blackwell.

Watson, Duane & Gibson, Edward. 2004. The relationship between intonational phrasing and syntactic structures in language production. *Language and Cognitive Processes* 19(6). 713–755. (https://doi.org/10.1080/01690960444000070).

Webber, Bonnie L. 1988. Discourse deixis and discourse processing. *Technical reports of the University of Pennsylvania*.

Weiss, Gary M. 2004. Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsletter* 6(1). (https://doi.org/10.1145/1007730.1007734).

Weiss, Gary M. & Provost, Foster. 2001. *The effect of class distribution on classifier learning: An empirical study*. Technical report, ML-TR-44, Department of Computer Science, Rutgers University, New Jersey. (https://doi.org/10.7282/t3-v9kt-9510).

Werth, Alexander. 2020. Referenzkoordination: Namengrammatik im Dienste des Rezipientendesigns [Reference coordination: The grammar of proper names in the service of recipient design]. In Kempf, Luise & Nübling, Damaris & Schmuck, Mirjam (eds.), *Die Linguistik der Eigennamen*, 259–284. Berlin: Mouton de Gruyter.

Wolf, Florian & Gibson, Edward & Fisher, Amy & Knight, Meredith. 2005. *Discourse Graphbank (LDC2005T08)*. Philadelpha, PA: Linguistic Data Consortium. (https://doi.org/10.35111/7snd-y3976).

Yamamoto, Mutsumi. 1999. *Animacy and reference: A cognitive approach to corpus linguistics*. Amsterdam: John Benjamins. (https://doi.org/10.1075/slcs.46).

Yoshida, Etsuko. 2011. *Referring expressions in English and Japanese: Patterns of use in dialogue processing* (Pragmatics & Beyond 208). Amsterdam: John Benjamins.

Yule, George. 1981. New, current and displaced entity reference. *Lingua* 55(1). 41–52.

Zeman, Daniel & Nivre, Joakim & Abrams, Mitchell & alii. 2020. *Universal Dependencies 2.6*. Prague: Universal Dependencies Consortium.

Ziemski, Michał & Junczys-Downmunt, Marcin & Pouliquen, Bruno. 2016. The United Nations Parallel Corpus v1.0. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'2016), Portorož, Slovenia, 23–28 May 2016*.

Zimmermann, Malte & Féry, Caroline (eds.). 2010. *Information structure: Theoretical, typological, and experimental perspectives*. Berlin: Language Science Press.

Zipf, George K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Zulaica Hernández, Iker. 2009. Referential distance, demonstrative anaphors and the current focus of attention. In van Deemter, Kees & Gatt, Albert & van Gompel, Roger P. G. & Krahmer, Emiel J. (eds.), *Proceedings of the workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009)*. Amsterdam: University of Tilburg.

Appendices

A | Lists of symbols

A.1 | Abbreviations

A	subject of a transitive clause
GBM	gradient boosting machine (Section 7.2)
GLMR	generalized linear mixed-effects regression
GRAID	Grammatical Relations and Animacy in Discourse (Haig & Schnell 2014, Section 3.3.1)
Multi-CAST	Multilingual Corpus of Annotated Spoken Texts (Haig & Schnell 2015, Section 3.2)
NLP	natural language processing
NP	noun phrase
P	direct object
PR	precision-recall curve
RD	referential density (Noonan 2003; Bickel 2003)
RefIND	Referent Indexing in Natural-language Discourse (Schiborr et al. 2018, Section 3.3.2)
ROC	receiver operating characteristic
S	subject of an intransitive clause
VP	verb phrase

acc	accuracy
AIC	Akaike information criterion
d.f.	degrees of freedom
e^{β}	log odds
lr	learning rate
M	median
MCC	Matthews correlation coefficient

N	frequency
P	proportion
p	probability
prc	precision
R^2	Nakagawa's R^2
r	Pearson's correlation coefficient
r_{pb}	point-biserial correlation coefficient
SE	standard error
sns	sensitivity / recall
spc	specificity
tc	interaction depth
β	regression coefficient
η	correlation ratio
ρ	Spearman's rank correlation coefficient
σ	standard deviation
φ	Pearson's φ-coefficient
χ^2	χ^2 statistic

A.2 | Morphological categories

Listed here are only those abbreviations that are used in the examples in this book. For an exhaustive list of abbreviated morphological categories employed in the description of the analyzed languages, please refer to the lists included at the end of the annotation notes for each Multi-CAST corpus.

1	first person
2	second person
3	third person
ABL	ablative
ABS	absolutive
ACC	accusative
ADD	additive (Tabasaran)
ADV	adverbial
ADVZ	adverbializer
ANTE	anterior case
AOR	aorist
APUD	apudessive case
ART	article
ART2	article 2 (Teop)
ASP	aspect marker
ATTR	attributive
CAUS	causative
CIT	quotative (Tabasaran)
CL	classifier
CS	construct suffix (Vera'a)
CVB	converb
DAT	dative
DEF	definite
DEL	delimitative aktionsart (Vera'a)
DEM1	basic demonstrative 1 (Vera'a)
DEM2	basic demonstrative 2 (Vera'a)
DET	determiner
DIM	diminutive
DISC	discourse particle (Vera'a)
DIST	distal

DL	dual
ELAT	elative
EMPH	emphatic; emphatic particle (Vera'a)
ERG	ergative
EZ	ezafe (Northern Kurdish)
F	feminine
FUT	future
GEN	genitive
GEN.NEG1	general negation 1 (Vera'a)
GEN.NEG2	general negation 2 (Vera'a)
HPL	human plural (Sanzhi Dargwa)
ICVB	imperfective converb
IMP	imperative
IN	inessive case
INAL	inalienable possession (Tulil)
IND	indicative
INDF	indefinite
INF	infinitive
IPFV	imperfective
LAT	lative case
LOC	locative; locative preposition (Vera'a)
M	masculine
MAN	manner (Vera'a)
MASC	masculine class (Tulil)
MED	medial
MOD	modifier particle (Mandarin)
MP	modal particle (Mandarin)
N	neuter
NEG	negation
NOM	nominative
NPL	neuter plural (Sanzhi Dargwa)
NSG	neuter singular (Tabasaran)
OBL	oblique case; oblique stem (Sanzhi Dargwa)
PAT	patientive
PCVB	perfective converb (Tabasaran)

PERS	personal (Vera'a)
PFV	perfective
PL	plural
POSS	possessive
PRET	preterite
PRON	pronominal (Teop)
PROX	proximal
PRS	present
PRT	particle
PST	past
PTCP	participle
PURP	purposive
QUOT	quotative
RED	reduplication
REFL	reflexive
RS	realis subject (Nafsan)
SG	singular
SIM	simultaneity (Vera'a)
SP	specific (Vera'a)
STAT	stative tense-aspect-mood marker (Vera'a)
SUPER	superlative case
TAM1	tense-aspect-mood-polarity marker 1 (Vera'a)
TAM2	tense-aspect-mood-polarity marker 2 (Vera'a)
TAM3	tense-aspect-mood marker 3 (Teop)
TOP	topic marker
TR	transitivizer (Nafsan)

A.3 | GRAID symbols

Listed here are only those GRAID annotation symbols (cf. Haig & Schnell 2014 that are used in the examples in this book. For an explanation of how form, animacy/person, and function symbols combine in the GRAID system, see Section 3.3.1. In the list below, language-specific symbols that extend the core symbol inventory of GRAID are marked with [+].

A.3.1 | Form symbols

⟨0⟩	paradigmatic zero
⟨f0⟩	structurally enforced zero
⟨pro⟩	free definite pronoun
⟨np⟩	lexical noun phrase
⟨refl⟩	reflexive
⟨v⟩	verb or verb complex
⟨vother⟩	non-canonical verb form
⟨other⟩	other forms, not further specified

A.3.2 | Person/animacy symbols

⟨.1⟩	first person
⟨.2⟩	second person
⟨.h⟩	third person, human
⟨.d⟩	third person, anthropomorphized
ø	third person, non-human (not annotated)

A.3.3 | Function symbols

⟨:s⟩	subject of an intransitive clause
⟨:ncs⟩	non-canonical subject
⟨:a⟩	subject of a transitive clause
⟨:p⟩	object of a transitive clause
⟨:p2⟩	secondary object of a ditransitive clause
⟨:obl⟩	oblique argument
⟨:g⟩	goal, recipient, addressee (subtype of ⟨:obl⟩)
⟨:l⟩	static location (subtype of ⟨:obl⟩)

⟨:lvc⟩	complement of a light verb
⟨:dt⟩	dislocated topic, may combine with other function symbols
⟨:voc⟩	vocative
⟨:poss⟩	possessive
⟨:appos⟩	apposition
⟨:pred⟩	predicate
⟨:predex⟩	predicate of existential/presentational constructions

A.3.4 | Form and function specifiers

⟨pn_⟩	proper name; attaches to ⟨np⟩
⟨dem_ ⟩	[+] demonstrative pronoun; attaches to ⟨pro⟩, ⟨other⟩
⟨nc_⟩	not further classified; attaches to most form symbols
⟨_ds⟩	subject of a verb of speech; attaches to ⟨:s⟩, ⟨:ncs⟩, ⟨:a⟩
⟨_dem⟩	[+] demonstrative determiner; attaches to ⟨ln⟩, ⟨rn⟩
⟨_det⟩	[+] definite determiner; attaches to ⟨ln⟩, ⟨rn⟩
⟨_deti⟩	[+] indefinite determiner; attaches to ⟨ln⟩, ⟨rn⟩
⟨_num⟩	[+] numeral; attaches to ⟨ln⟩, ⟨rn⟩
⟨_aux⟩	auxiliary; attaches to ⟨lv⟩, ⟨rv⟩
⟨:pred⟩	predicate
⟨:predex⟩	predicate of existential/presentational constructions

A.3.5 | Other symbols

⟨ln⟩	NP-level subconstituent, left of head
⟨rn⟩	NP-level subconstituent, right of head
⟨lv⟩	VP-level subconstituent, left of head
⟨rv⟩	VP-level subconstituent, right of head
⟨adp⟩	adposition
⟨cop⟩	copula
⟨nc⟩	not classified
⟨=⟩	marks clitics
⟨-⟩	marks affixes

A.3.6 | Clause boundary symbols

⟨##⟩	left-edge boundary of an independent clause
⟨#⟩	left-edge boundary of a dependent clause
⟨%⟩	right-edge boundary of an embedded clause
⟨(_)ds⟩	direct speech; attaches to ⟨##⟩, ⟨#⟩
⟨(_)cc⟩	complement clause; attaches to ⟨#⟩
⟨(_)ac⟩	adverbial clause; attaches to ⟨#⟩
⟨(_)rc⟩	relative clause; attaches to ⟨#⟩
⟨(_)cv⟩	[+] converb clause; attaches to ⟨#⟩
⟨nc⟩	segment not classified; attaches to ⟨#⟩
⟨(.)neg⟩	negated clause; attaches to ⟨##⟩, ⟨#⟩

B | Corpus metadata

B.1 | List of texts

						referents		mentions	
corpus	text	type	speaker	rec'd	clauses	all	sampl.	subj.	obj.
C. Greek	*jitros*	TN	CG01	1960	271	89	32	119	57
C. Greek	*minaes*	TN	CG01	1960	359	110	45	115	72
C. Greek	*psarin*	TN	CG01	1964	440	97	42	207	106
English	*devon01*	AN	EN02	1980	590	327	101	142	62
English	*kent01*	AN	EN01	1975	622	311	103	237	123
English	*kent02*	AN	EN01	1975	1637	670	230	544	330
English	*kent03*	AN	EN03	1976	1335	625	182	432	208
Mandarin	*hml*	TN	MD01	2015	301	191	48	196	59
Mandarin	*jgz*	TN	MD02	2015	711	186	71	404	126
Mandarin	*lzh*	TN	MD03	2015	182	89	21	119	18
Nafsan	*kori*	TN	NF01	1998	284	33	20	207	57
Nafsan	*lelep*	TN	NF01	1998	129	34	21	62	19
Nafsan	*lisau*	TN	NF02	1998	58	22	15	41	21
Nafsan	*litog*	TN	NF02	1998	86	24	10	77	25
Nafsan	*maal*	TN	NF03	1997	52	15	9	31	21
Nafsan	*nmatu*	TN	NF03	1996	88	28	10	62	32
Nafsan	*ntwam*	TN	NF04	1996	186	51	23	126	31
Nafsan	*taapes*	TN	NF02	1998	67	12	5	57	14
Nafsan	*tafra*	TN	NF03	1997	62	43	13	35	8
N. Kurdish	*muserz01*	TN	NK01	2000	627	152	58	346	107
N. Kurdish	*muserz03*	TN	NK01	2002	732	145	59	297	150
S. Dargwa	*asabali*	AN	SD01	2012	142	73	19	32	11
S. Dargwa	*bazhuk*	TN	SD02	2013	99	21	10	74	14
S. Dargwa	*dragon*	TN	SD02	2013	121	30	13	96	18
S. Dargwa	*kurban*	AN	SD03	2011	164	47	15	68	18
S. Dargwa	*mill*	TN	SD01	2013	130	38	18	79	24
S. Dargwa	*patima*	TN	SD02	2013	133	31	13	83	16
S. Dargwa	*ramazan*	AN	SD04	2012	209	99	24	65	21
S. Dargwa	*tape*	AN	SD03	2011	68	20	9	27	3
Tabasaran	*belt*	TN	TS01	2010	170	33	18	96	32
Tabasaran	*horse*	TN	TS02	2010	422	111	44	272	80

						referents		mentions	
corpus	text	type	speaker	rec'd	clauses	all	sampl.	subj.	obj.
Tabasaran	*horse*	TN	TS02	2010	422	111	44	272	80
Tabasaran	*naz*	TN	TS01	2010	118	48	14	57	14
Tabasaran	*nuradin*	AN	TS01	2010	150	90	19	87	18
Tabasaran	*work*	TN	TS01	2010	523	116	39	274	100
Teop	*iar*	TN	TP01	2003	348	83	29	190	76
Teop	*mat*	TN	TP02	2004	207	38	21	123	37
Teop	*sii*	TN	TP03	2004	590	116	30	326	100
Teop	*viv*	TN	TP04	2004	158	30	13	112	42
Tulil	*all1*	TN	TL01	2012	93	33	12	66	24
Tulil	*alrm*	AN	TL01	2014	407	118	44	210	93
Tulil	*jkpp*	AN	TL02	2014	414	163	62	112	55
Tulil	*lnsl*	TN	TL03	2014	92	14	10	52	12
Tulil	*lrdw*	TN	TL04	2007	157	25	9	97	29
Tulil	*sves*	TN	TL05	2014	101	30	12	83	24
Vera'a	*anv*	TN	VR01	2007	182	43	24	143	41
Vera'a	*as1*	TN	VR02	2007	213	34	20	151	45
Vera'a	*gabg*	TN	VR03	2007	174	31	15	87	15
Vera'a	*gaqg*	TN	VR04	2007	226	43	20	137	37
Vera'a	*hhak*	TN	VR05	2007	432	77	37	313	74
Vera'a	*isam*	TN	VR06	2007	238	59	17	148	30
Vera'a	*iswm*	TN	VR07	2007	576	149	64	418	109
Vera'a	*jjq*	TN	VR08	2007	880	114	54	539	137
Vera'a	*mvbw*	TN	VR09	2007	307	50	29	201	53
Vera'a	*pala*	TN	VR10	2007	380	56	34	293	73

Table B.1 | List of the texts in the data used for this study, all taken from the Multi-CAST collection (Haig & Schnell 2015). See Schiborr (2019: Appx. A) for a complete list of texts in Multi-CAST.

'Type' here refers to text type, with 'TN' being traditional narratives and 'AN' autobiographical narratives. 'Rec'd' is the year the text was recorded. The four rightmost columns list, from left to right, the total number of unique referents in the text, the number of referents whose mentions are part of the effective sample, and the number of mentions in subject and object position in the sample.

B.2 | List of speakers

corpus	speaker	gender	age	born	rec'd	texts	clauses
C. Greek	CG01	female	73 77	1887	1960 1964	*jitros* *minaes* *psarin*	1070
English	EN02	male	c80	c1900	1980	*devon01*	590
English	EN01	male	85	1890	1975	*kent01* *kent02*	2259
English	EN03	male	87	1889	1976	*kent03*	1335
Mandarin	MD01	male	23	1992	2015	*hml*	301
Mandarin	MD02	male	23	1992	2015	*jgz*	711
Mandarin	MD03	male	22	1993	2015	*lzh*	182
Nafsan	NF01	male	65	1933	1998	*kori* *lelep*	413
Nafsan	NF02	female	67	1931	1998	*lisau* *litog* *taapes*	211
Nafsan	NF03	male	85	1912	1997 1996	*maal* *nmatu* *tafra*	202
Nafsan	NF04	male	45	1951	1996	*ntwam*	186
N. Kurdish	NK01	male	c50 c60	c1950	2000 2002	*muserz01* *muserz03*	1359
S. Dargwa	SD01	male	76 77	1935	2012 2013	*asabali* *mill*	272
S. Dargwa	SD02	male	51	1963	2013	*bazhuk* *dragon* *patima*	353
S. Dargwa	SD03	male	60	1951	2011	*kurban* *tape*	232
S. Dargwa	SD04	male	58	1954	2012	*ramazan*	209
Tabasaran	TS01	male	52	1958	2010	*belt* *naz* *nuradin* *work*	961
Tabasaran	TS02	male	64	1946	2010	*horse*	422
Teop	TP01	female	c70	c1930	2003	*iar*	348
Teop	TP02	female	c30	c1970	2004	*mat*	207
Teop	TP03	female	c60	c1940	2004	*sii*	590
Teop	TP04	female	c30	c1970	2004	*viv*	158

corpus	speaker	gender	age	born	rec'd	texts	clauses
Teop	TP04	female	c30	c1970	2004	*viv*	158
Tulil	TL01	male	53 55	1959	2012 2014	*all1* *alrm*	500
Tulil	TL02	male	74	1940	2014	*jkpp*	414
Tulil	TL03	male	c55	c1960	2014	*lnsl*	92
Tulil	TL04	male	77	1930	2007	*lrdw*	157
Tulil	TL05	female	c80	c1930	2014	*sves*	101
Vera'a	VR01	female	c20	c1985	2007	*anv*	182
Vera'a	VR02	male	c40	c1965	2007	*as1*	213
Vera'a	VR03	male	c40	c1965	2007	*gabg*	174
Vera'a	VR04	male	c40	c1965	2007	*gaqg*	226
Vera'a	VR05	male	c20	c1985	2007	*hhak*	432
Vera'a	VR06	male	c60	c1950	2007	*isam*	238
Vera'a	VR07	male	c60	c1950	2007	*iswm*	576
Vera'a	VR08	male	c60	c1950	2007	*jjq*	880
Vera'a	VR09	male	c30	c1975	2007	*mvbw*	307
Vera'a	VR10	female	c40	c1965	2007	*pala*	380

Table B.2 | List of the speakers in the data used for this study, all taken from the Multi-CAST collection (Haig & Schnell 2015). See Schiborr (2019: Appx. B) for a complete list of speakers in Multi-CAST.

'Born' is the year the speaker was born, and 'rec'd' is the year the text was recorded. Approximate values are prefixed with 'c'.

Deep in the jungle, in a lonely village, the woman was born. As a child, she learnt the ways of the villagers. How they celebrated, how they loved, how they worked, how they died. But when she was old enough to understand what she wanted out of her own life, she knew such an existence was not for her. She dediced then that she would achieve more than small victories and fleeting joys. She would achieve greatness. She walked into the jungle alone, and began to build the *monument*.

The villagers wondered what she was doing, and why. They said to her, "You are a young girl. You should be spending your time being free and happy, not building useless buildings." She told them it was a monument to *happiness*, and it would stand for the happiness of all people, in all times. The villagers scoffed at her foolishness, but they could not convince her to return to the village, so they left her be.

Seasons passed, and still she built. She was now a woman, and the base of the monument was taking shape. The villagers came again and told her that a woman her age should not concern herself building monuments to happiness. She should be looking for love, like the other village women. She told them that she was building a monument to *love*, and it would be standing long after the love in the hearts of the village women was gone. They scoffed again at her foolishness, but still they could not dissuade her, so they left her be.

The summer heat came on, and the mosquitoes with it, but she carried on. The monument was rising into the sky, a testament to her labour. The villagers returned. They pleaded with her to stop this foolishness. They told her that a woman her age should be raising children and preparing them for the future. She told them that she was building a monument to the *future*. It should stand as a symbol of hope, that things to come may be better than what has been. The villagers could only shake their heads again, and leave her be.

The winters were the hardest. The cold made the stone hard to work, but she carried on still. The villagers came to plead with her one last time. They said that she was now an old woman, and she should be resting. But she would not stop now. Her old hands rolled the last stones into place. The *monument* was complete. She looked at it in its glory, knowing now that despite its magnificence, it could not stand for the *happiness*, *love*, or *future* of all humanity. Nothing could, for no person has the authority to set in stone the purpose of others' lives. It was a monument to only her. A monument to her life.

She died perfectly at peace, knowing that she had completed the project of her life. But whenever the villagers passed it by, they shook their heads. For them, it stood as a monument only to the woman's *folly* at having wasted her life. But soon the villagers died, and their children's children saw it differently. They looked at it with wonder, unable to believe a single person could achieve such a thing.

As time passed the village too disappeared. All those who stumbled upon the monument wondered at what great people might have lived there long ago. They wondered what it was for, and why it was built. It seemed to serve no practical purpose, nor be dedicated to any god. It had only one small inscription on the base.

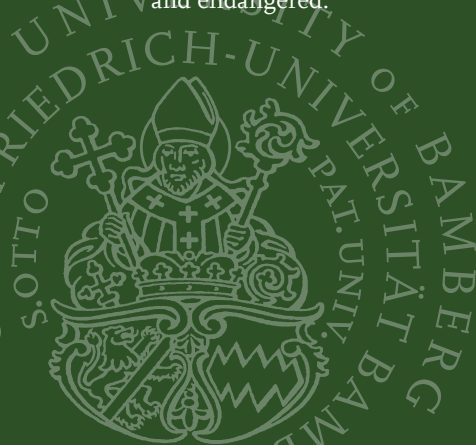
"Each of us can only live as we choose."

| Corey Mohler, 'The Monument'

University
of Bamberg
Press

Whenever speakers refer to entities or events, they are tasked with making a series of decisions: which entity to refer to, how to embed it into syntactic structures, and which type of expression to use for the reference. This study concerns itself with the last of these decisions, the selection of the linguistic exponents of reference, or referential choice. While it is well known that languages differ greatly in their preference for either pronominal *(she, this)* or zero anaphora (ellipsis) to refer back to previously mentioned discourse referents, the overall rates of occurrence of lexically-headed anaphora *(the woman, Jane)* in monological discourse turn out to be remarkably stable across languages.

This study examines the circumstances in which speakers opt for the more informative but less econominal choice of lexical references over reduced alternatives. It does so from a typological and comparative angle, charting the cross-linguistic stability of certain classes of explanatory factors and the parametricization of others across languages. Rather than adopting a specific theoretical framework in its approach to the question, it instead explores a number of empirical bottom-up approaches, deriving complex analytical categories from multiple levels of relatively basic annotations. Where earlier research on referential choice has been focused predominantly based on written data, or else on small data sets of spoken language from English and other overrepresented languages, this study addresses issues of typological representativity by employing spoken corpora from a diverse set of ten languages, many of which are understudied and endangered.



ISBN: 978-3-86309-939-8

www.uni-bamberg.de/ubp