Language Documentation & Conservation Special Publication 25
Doing corpus-based typology with spoken language data: State of the art
edited by Geoffrey Haig, Stefan Schnell, and Frank Seifart, pp. 1 – 28
https://nflrc.hawaii.edu/ldc/sp25
https://hdl.handle.net/10125/74656

The role of language documentation in corpus-based typology

Stefan Schnell,[1] Geoffrey Haig,[2] Frank Seifart [3]

[1] *University of Zurich,* [2] *University of Bamberg,* [3] *Leibniz-Centre General Linguistics (ZAS)*

1 Introduction

The global decline in linguistic diversity was first brought to public attention some 30 years ago (Hale et al. 1992), and has since continued unabated (Seifart et al. 2018; Bromham et al. 2021). From the perspective of language communities, language loss means an irredeemable rupture of collective and individual memory; for the language sciences, each disappearing language shrinks our window on the range of variability in human language. The urgency of language documentation could scarcely be more obvious.[1]

For typology, a major utility of language documentation was the adequate representation of typological rara in linguistic theory: syntactic ergativity,

[1] This volume grew out of a workshop on *Corpus-based Typology: Spoken Language from a Cross-linguistic Perspective*, held at the Annual General Meeting of the German Linguistics Society (DGfS) at the University of Hamburg in March 2020. We, the editors, would like to express our thanks to the audience at the workshop for very stimulating discussion, to the external reviewers of the contributions of this volume for their critical feedback, and to Nils Schiborr for supervising the final volume production. The responsibility for any remaining errors is of course our own.

OSV constituent order, non-configurationality, and so on (Comrie 1981: 5–12). But more recently, the role of language documentation in typology has been re-defined by a shift towards a more usage-based approach to typology, incorporating methodological advances from corpus linguistics, variationist sociolinguistics, and cognitive sciences. Rather than extracting linguistic generalizations from pre-processed data as provided by descriptive grammars, the primary data created by language documentation itself, in the form of records of language usage, can provide the input for a usage-based approach to typology which seeks to go beyond the emphasis of traditional typology on construction types outside of context (e.g. Greenberg 1963), and incorporate patterns of variation and the impact of context.

Although it was not necessarily foreseen by earlier documentary linguistics, it transpires that the emphasis on maximally context-embedded documentation of language usage yields precisely the kind of data that is amenable for usage-based approaches to typology. This is particularly noteworthy in view of the critique of this kind of documentation as a "data graveyard" (Newman 2013). Data graveyards are now data goldmines, and typologists are just beginning to explore their potential (Seifart et al. 2012). In this volume we showcase a selection of state-of-the art research that embodies this usage- and corpus-based approach to typology, drawing extensively (but not exclusively) on primary linguistic data compiled and archived initially for the purpose of language documentation. In the remainder of this introduction, we outline in Section 2 a research agenda which we refer to as 'corpus-based typology' (CBT), outlining influential developments in corpus linguistics and typology, recent advances and remaining challenges. In Section 3 we summarize in what ways the contributions aggregated in this volume address some of these challenges, before concluding in Section 4.

2 Corpus-based typology

Corpus-based typology refers to a set of approaches that use corpora to conduct language typology. The following sections are devoted to key aspects of this research agenda: Section 2.1 addresses the concept of 'corpus' in a cross-linguistic context; Section 2.2 looks at the role of language-internal variation; Section 2.3 addresses different aspects of language usage. We refer the reader to Levshina (2021a) and Schnell & Schiborr (in press) for more extensive overviews of recent developments; here we merely summarize the main points before identifying remaining challenges and desiderata (Section 2.4), which the current volume intends to address.

2.1 Multilingual corpora

CBT deploys methods developed in neighbouring disciplines, such as corpus linguistics and variationist sociolinguistics (see below), to address a range of questions relevant for language typology. Hence, the research showcased in this volume all draw on corpora, understood here as follows (cf. Barth & Schnell 2021: 2–8): A corpus is a text or a collection of texts which are in turn defined as a succession of connected utterances that form a coherent whole. Texts in a corpus can be spoken, signed, or written. An important aspect of corpora is that individual linguistic structures and texts as a whole can be analysed with regards to its contextual determinants, although in practice the extent to which such analyses are possible depends heavily on the level of linguistic representation and their immediate context (e.g. sound segment within a phonological word or sentence within a text) and the detail and quality of metadata (for entire texts). These features differentiate a corpus from collections of elicited, context-free sentences of the kind frequently used in grammars, and from paradigms and word lists. A corpus in this sense permits the analysis of, for example, clause-combining, prosodic chunking, referent tracking, turn-taking, narrative structure, and so on. To serve as the input to corpus-based typology, a corpus in this sense should provide sufficient numbers of tokens relevant to a given research question to enable the identification of relevant patterns, possibly through meaningful statistical analysis.

Corpora in typology differ along a scale of content control, that is the degree to which the content of the corpus is pre-determined by the investigator. At the highest end of the scale are parallel texts: Corpora produced as translations from a single source text, for example the Parallel Bible Corpus (PBC) (Mayer & Cysouw 2014) and the *Universal Declaration of Human Rights*

(UDHR)[2] for which translations in about 1400 languages are now available (cf. Bentz & Ferrer-i-Cancho 2016). The VoxClamantis project (Salesky et al. 2020) is a corpus of Bible translations read aloud by native speakers, with first-pass phoneme-level alignments for more than 600 languages. A lower degree of content-control is found in so-called parallax corpora, where speakers produce texts in response to a specific non-verbal stimulus. The best-known examples are those based on re-tellings of the *Pear Story* video (Chafe 1980), or the *Frog Story* picture book (Mayer 1969; see Berman & Slobin 1994). See Barth & Evans (2017) and Barth & Evans (this volume) for examples of the parallax methodology. Finally, at the lowest level of content control are corpora produced without any specific pre-defined content constraints. These include, for example, life stories, traditional narratives, descriptions of activities, and so on, loosely referred to as "original text typology" (Haig et al. 2011). Texts of this type are among the most common outputs of language documentation, and are thus of particular relevance in the current connection. Degree of content control has a considerable impact on the nature of the typological questions that can be addressed: High content control is designed to elicit a high degree of semantic consistency across different corpora, and is thus ideal for probing, for instance, cross-linguistic differences and commonalities of specific event types in specific contexts (e.g. motion events, Wälchli & Sölling 2013). However, high content control comes at the cost of a lack of naturalness, and possible interference from the source texts.

Corpora also vary across other dimensions, for example medium (spoken, signed, or written, further discussed in Section 2.4 below), single-participant (monologue) or multi-participant, register, and genre. This volume is devoted to spoken language corpora, which always minimally involve recording, and generally transcription, followed by further levels of annotation, which may exhibit various degrees of conceptual abstraction, from a simple, content-based translation to exhaustive morpheme-for-morpheme glossing, with indications of prosody, information structure, or semantic roles. See Seifart (this volume), Mettouchi & Vanhove (this volume) and Haig et al. (this volume) for different approaches to phonological, prosodic, morphosyntactic, and information structure annotation.

[2] https://www.unicode.org/udhr/

Common to all these annotation schemes is the commitment to maximal cross-linguistic applicability of the annotated categories. In this respect, CBT on spoken language corpora aligns with recent approaches to corpus- or token-based typology based on Universal Dependencies (UDs), a widely-used system of morpho-syntactic annotation that is primarily implemented semi-automatically on standardized written corpora (Zeman et al. 2021)[3]. Both compare corpora of different languages, both aim at deriving higher-level generalizations extracted directly from the corpora, rather than via the mediation of pre-formulated grammatical analysis (see Wälchli 2009 on the latter point), and both develop annotation schemes specifically designed to be applicable across a typologically diverse range of languages.

2.2 Variation centre stage

Typology in the tradition of Greenberg's seminal work focused on identifying limits of cross-language variation by classifying languages according to construction types (e.g. "prepositional", "SOV", etc.). Sampling procedures were developed to ensure that observed (dis)preferences were really universal to human language rather than stemming from shared histories of inheritance (i.e. from a genealogical bias), or from extensive long-term contact (i.e. from an areal bias). Explanations for (statistical) limits of variation were then sought in universal properties of human communication and language processing. For instance, the near-absence of OSV in the world's languages is explained in terms of violating two functional principles, "S before O" (new before given), and also "O should be adjacent to V", due to the close conceptual link between objects and verbs (e.g. Tomlin 1986; Hawkins 1994).

However, this "whole-language" typological approach glosses over a lot of the intra-linguistic variation (Bickel 2009). For instance, languages may show different word order rules in different sentence types (cf. independent and VO vs. subordinate clauses and OV in German) or different alignment systems (e.g. split in terms of tense, as in Iranian languages, or split along different encoding devices, e.g. ergative case marking and accusative agreement in Tukang Besi; Donohue 1999: 51–54). Zooming in even closer to intra-

[3] https://universaldependencies.org/

linguistic variation, Multivariate Typology (MVT, Bickel 2015) suggests that linguists should focus on low-level aspects of language, using what Levinson & Evans (2010: 2738–2739) call Low-Level Feature Metalanguage (LLFM). This low-level approach advocates late aggregation, where the investigation of low-level categories precedes the identification of higher-level concepts (Zakharko et al. 2017). In this bottom-up approach, typological and potentially universal trends are statistically significant clusters of properties. Crucially, variation is not glossed over in MVT, since observations that do not follow a given trend are still represented as statistical minorities or outliers.

The concern of CBT with variation builds on well-established foundations, most notably in variationist sociolinguistics in the tradition of William Labov (1972, 1994), anthropological linguistics (e.g. Hymes 1961, 1962; Duranti 1981, 1997), conversation analysis (Schegloff 2006; Sacks et al. 1974), research in language acquisition (MacWhinney 2000; Slobin 1985 and later works) as well as general corpus linguistics (McEnery & Wilson 2001 among many others). More recently, these sub-disciplines have moved beyond studying a small number of well-researched languages to investigating variation across increasingly diverse language samples; see the contributions in Stanford & Preston (2009), and Mansfield & Stanford (2017) on sociolinguistic studies of understudied languages; Meyerhoff (2009) and Torres Cacoullos & Travis (2018) on corpus-based approaches to null-subjects across languages; and Dingemanse et al. (2014) as well as Dingemanse et al. (2017) for pragmatics, including conversation analysis. Similarly, in more general corpus linguistics, studies in register and genre variation have also been conducted on lesser-studied languages (cf. Schnell & Barth 2018) and in comparative perspective (Biber 1995).

2.3 Language use

Despite its concern for intra-linguistic variation, MVT is primarily concerned with language systems and different structural subtypes. CBT research, on the other hand, aims at capturing intra-linguistic variation in language use and its relation to aspects of language systems. It may thus, for example, recast word order typology by refining characterizations such as “dominant OV” (Dryer 1992) as “73% OV” (Gerdes et al. 2021). Moreover, it can advance conceptions of linguistic categories as clusters of contextualized exemplars in language use. For instance, Cysouw (2014) induces semantic roles (such as Goal, Agent, or Patient), or even macro-roles (like S, A, and P) from the usage of case-like markers across 15 languages, based on parallel religious texts. In this section we give examples of research that seeks motivations for typological (dis)preferences in different levels of linguistics structure in patterns of language use. These examples illustrate what CBT's corpus-linguistic approach adds to the study of variability beyond the recognition of intra-language variation that is already registered within MVT and related approaches, which can, in principle, be applied to, for example, elicited or pre-processed data.

In syntactic typology, dependency length minimisation (DLM) has long been hypothesized to underlie the typological preference for harmonic word order patterns, for instance VO co-occurring with prepositions in head-initial languages (cf. Behaghel 1909; Dryer 1992; Hawkins 2014, 2004). How DLM shapes actual language use has now been directly investigated comparatively in corpora from over 50 languages (Futrell et al. 2020; Jing et al. 2021). Similarly, there are now corpus-based results on the avoidance of crossing dependencies (Blasi et al. 2019), as well as cross-linguistic differences in word order variability (more or less "free" word order) (Futrell et al. 2015; Levshina 2019) and the trade-off between word order variability and case marking (Koplenig et al. 2017; Levshina 2019, 2021b).

Another classic topic in syntactic typology is zero reference, discussed in the theoretical and typological literature under the heading of "pro-drop" (Rizzi 1982; Jaeggli & Safir 1989; Roberts & Holmberg 2009; cf. Dryer 2013). Bickel (2003) and Stoll & Bickel (2009) investigate the rate of zero expression of syntactic arguments across a set of three languages from the Himalayas, and Russian. Factors impacting on referential density, that is the rate of overt to covert forms of reference, are identified as patterns of clause combining as well as ethnolinguistic considerations of discourse production and reception. While these authors focus on possible realizations of all arguments in a particular corpus, Torres Cacoullos & Travis (2019) focus on the specific conditions of zero reference for subjects, finding, for instance, that the same factors are involved in both English and Spanish texts, but differ in their degree of magnitude. Vollmer (2019) applies classification tree methodologies to investigate referential choice in a diverse sample of 8 languages, likewise identifying underlying commonalities across diverse corpora. These works thus open up comparative and quantitative ways to address systematic differences in discourse production that can be correlated with other linguistic features and/or cultural differences.

In morphological typology, marking asymmetries such as the presence versus absence of morphological marking on word pairs like Spanish *romper* '(causative) break' and *romper-se* '(inchoative) break', have long been discussed (e.g. Greenberg 1966). Haspelmath et al. (2014; see also Haspelmath 2021) found that, across corpora from various languages, overt morphological marking corresponds to lower mention frequencies, compared to the unmarked member. This provides a motivation for marking asymmetries in human language based on information-theoretic accounts in terms of efficiency going back to Zipf (1935; cf. Gibson et al. 2019 for an overview). Regarding word length more generally, recent work in CBT has found support for Zipf's Law of Abbreviation, which describes the relation between the length of words and their text frequencies in corpora from about 1 400 languages (Bentz & Ferrer-i-Cancho 2016). Other recent CBT research found that word length is even more closely related to a higher predictability of words in context, compared to the raw text frequency of a word (Piantadosi et al. 2011; Gibson et al. 2019). Similarly, albeit more directed towards system (lexicon-related) properties, is Cohen Priva's (2017) study of form reduction and functional load that essentially finds that reduction is more likely with elements that have a low functional load.

In pioneering work at the interface of prosody and morphological typology, Himmelmann (2014) studied patterns of hesitations and disfluencies observed in spoken corpora of English, German, and Tagalog, providing an alternative explanation for the suffixing preference in the world's languages (Cutler et al. 1985; Bybee et al. 1990). Word class-specific hesitation patterns were studied in Seifart et al. (2018) across nine typologically diverse languages. Both studies intend to link typological preferences in the morphology of the world's languages to speech and sentence planning by taking speech rate reduction and pauses, as observable in spoken corpora, as a proxy indicating planning difficulties.

Focussing on potentially universal characteristics of information transmission in spoken language production, Pimentel et al. (2021) study the duration of phones as a function of surprisal across 600 languages, based on the VoxClamantis corpus, finding evidence for a surprisal-duration trade-off, and thus for the Uniform Information Density hypothesis (Frank & Jaeger 2008). Comparing 17 languages, Coupé et al. (2019) find that higher syllable complexity (which induces higher informativity of each syllable) correlates with slower articulation rate, resulting in a pattern whereby information transmission rates across languages are surprisingly comparable. Regarding the temporal chunking of speech, Inbar et al. (2020) measured the regularity of sequences of intonation units across six spoken corpora, with results suggesting that they closely match brain waves at 1Hz.

2.4 Remaining challenges

Significant results are already emerging from CBT studies that connect preferences in usage and underlying processing biases much more immediately to typological distributions of linguistic structures. Remaining challenges relate to, firstly, reliance on written modality, secondly, relatively small and biased samples in terms of languages and genres, and, thirdly, the comparability of the corpora regarding contents and annotation.

The large collections of Universal Dependency-annotated texts, and the large parallel text corpora such as the PBC consist predominantly or exclusively of published written texts. These typically involve multiple passes of editing (potentially by multiple parties). UD corpora and the PBC are thus fairly remote from the primary, spoken or signed, modes of human speech production and reception. A recent example highlighting these issues is Just & Čéplö's (to appear) corpus-based investigation of bound object indexing in Maltese, an apparently widespread phenomenon in spoken Maltese. Their initial attempt to work with UD corpora of Maltese had to be abandoned because of the scarcity of the relevant constructions in the written corpora. Though non-comparative, Just & Čéplö's study suggests that considerations of mode will likely be highly relevant for comparative work in CBT as well. Further initial support for this comes from Schnell & Schiborr's (in press) comparison of (predominantly written) texts in UDs and the entirely spoken texts from Multi-CAST (Haig & Schnell 2021) which suggests that considerations of DLM may be hardly relevant in the latter corpus simply because it lacks critical proportions of complex, long NPs. Relatedly, greater tolerance for more, and more complex NPs in written texts has been explained by the greater ease of written text comprehension. Thus, written text production represented in UDs does not directly reflect relevant constraints on processing themselves, although effects of such constraints may be carried over from spoken or signed text production. Generally, spoken or signed text production is of primary interest to CBT since only here do we find 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.

In addition to modes of production and reception, current efforts in CBT are fairly constrained in the groups of language users they represent. On a global scale, this concerns the number and diversity of languages: While the PBC and VoxClamantis collections include subcorpora from a number of understudied languages (Bible translations are often prepared as part of missionary linguistic work in small language communities), UD corpora show a massive imbalance towards better-studied, often standard written languages of western Europe, largely corresponding to speaker populations identified as WEIRD and non-representative of humanity more generally by Henrich et al. (2010). And both the European and the few non-European languages represented in UD corpora are mostly those that Dahl (2015) labels LOL languages ("Literary tradition, Official language status, Large number of speakers"), likewise a tiny minority among the world's languages, and demonstrably not representative of these.

Likewise, intralinguistic variation in terms of, for example, speakers, genres, and registers is not properly reflected in current CBT research and cross-linguistic corpora: An extreme example are the PBC and VoxClamantis, which both represent a single text for each language community that is relevant for an extremely limited social register (practice of a religious cult), often excluding large proportions of any given community who do not read their language fluently. Though variation within understudied languages has only recently gained more attention (cf. Hildebrandt et al. 2017; Mansfield & Stanford 2017 and references therein), it seems obvious from the long tradition of sociolinguistics that inter-user variation is universal across languages and needs to be included in CBT (cf. Barth et al., this volume). Concerning variation in terms of text types, Schnell & Barth's (2018) study of object pronoun use in the Oceanic language Vera'a points to the relevance of considering cross-register variation for their finding of discourse topicality (rather than animacy) as the most relevant factor of pronoun use. Biber (1995: 359) finds differences across registers to be remarkably similar across diverse languages (English, Somali, Korean, Nukulaelae Tuvaluan),[4] pointing to the high relevance of considering cross-register variation in cross-linguistic research of language use.

Finally, CBT has to confront the problem of the relative comparability of corpora: Bickel (2003) makes a strong case for the use of content-controlled data in a study of referential density, given the high content-sensitivity of the parameter under study. Given the lesser representativeness of such experimentally elicited texts, an alternative approach is to determine pragmatically defined usage contexts that can be considered broadly comparable, as advanced in pragmatic typology (e.g. Dingemanse et al. 2014). Which of these two approaches is best suited for a given research agenda will depend heavily on the respective projects.

3 CBT and language documentation: Case studies

In the following sections we identify four areas where CBT and language documentation are already yielding a fertile union, addressing some of the challenges in areas of cross-linguistic and typological research identified in the previous section. We illustrate each of these areas with reference to the contributions to this volume. Section 3.1 looks at language acquisition in language documentation, Section 3.2 at the interface between prosody and morpho-syntax, Section 3.3 at cross-linguistic approaches to interaction, information packaging, animacy, and morpho-syntax, while Section 3.4 reports on cross-linguistic variation vis-à-vis intra-speaker variability in social cognition. Lastly, Section 3.5 discusses the challenges of corpus processing, annotation, and usability.

[4] Biber (1995) focuses on a comparative study of English and Somali registers, yet the assessment encompasses all four languages.

3.1 Acquisition in language documentation (Hellwig et al.)

In research on language acquisition corpus-based work has a long tradition (MacWhinney 2000). But availability of corpora of children's language production and child-directed speech are to date still limited to much less than 2% of the world's languages (Stoll & Bickel 2013: 198), and biased towards WEIRD societies and LOL languages. Efforts to broaden the scope of acquisition corpora have recently gained traction, for instance within the ACQ(uisition)DIV(ersity) research group led by Sabine Stoll at the University of Zurich (Moran et al. 2016). Hellwig et al. (this volume) outline the Sketch Acquisition project, which addresses the challenges of including a range of underdocumented and understudied languages in corpus building for language acquisition research while still compiling sufficient data for robust findings on acquisition. They stress the importance of an open corpus design for this kind of research matching the spirit of language documentation by keeping corpora maximally amenable to a range of research questions and projects. This also means that corpus annotations should be kept at a relatively general level, capturing what is relevant for a specific agenda, but leaving open the pursuit for other types of research. The authors report on a range of first-hand novel observations of children's verbal behaviour during language acquisition in the Papuan language Qaqet. For example, in their Section 2.2.1 the authors note that in this speech community, children's verbal interactions are mostly with other children, and these child-to-child interactions are characterized by a high frequency of exact repetitions. These findings serve as an important corrective to conventional views on the nature of the acquisition process, largely based on data from a small number of WEIRD speech communities, with a focus on adult-to-child transmission.

3.2 The interface of prosody and morpho-syntax: Towards a typologically informed and corpus-based approach (Seifart; Mettouchi & Vanhove)

As mentioned in Section 2.4 above, the study of specific properties of spoken language discourse production is a central concern for CBT, as this is the actual arena where discourse processing, prosodic chunking, and the formation of linguistic units plays out. Seifart (this volume) reports on research into durational prosodic properties of spoken text from diverse languages based on the multilingual corpus DoReCo ("Language DOcumentation REference COrpus"). DoReCo's goal is to significantly enhance the coverage of global linguistic diversity in spoken language corpus data, aiming at inclusion of 50 language corpora in the first release planned for mid-2022. For that purpose, DoReCo selects and processes subsets of language documentation collections, consisting mostly of traditional and personal narratives. Its particular value for linguistic analysis is the phonetic segment-level time-alignment of the transcription and dependent annotations, including morphological glossing for 35 languages.

Vanhove and Mettouchi's (this volume) contribution showcases two collections of multilingual corpora, collectively compiled and annotated: CorpAfroAs (2007–2012) and CorTypo (2013–2017). The first focuses on thirteen Afro-Asiatic languages, while the latter includes twelve languages from a range of phyla and interfaces the corpora with a searchable typological database. Both are characterized by an exceptionally rich annotation scheme, implemented in a purpose-built variant of the ELAN software (ELAN developers 2020; Chanard 2015) and combining morphological, syntactic, and prosodic segmentation. Their morphosyntactic annotation draws on those categories and features which are demonstrably relevant in the actual data, through "slow empirical building of established language-internal categories as a basis for further comparison, in a bottom-up perspective" (Mettouchi & Vanhove, this volume; see Section 2 above). Another crucial feature is the systematic segmentation of the corpora into intonation units (Izre'el & Mettouchi 2015). The result is an extremely fine-grained glossing and the possibility of detailed cross-corpus queries using the CorpAfroAs and CorTypo websites. Several case studies are presented illustrating the potential of this approach. One of these finds hitherto unnoticed interactions between word order and the prosodic integration of reported speech across a sample of four languages (Beja, Zaar [Chadic], Juba Arabic [Creole], and Modern Hebrew [Semitic]): In SOV languages, the onset of the speech report is systematically set off from the previous intonation unit through a clear prosodic cue whereas in SVO languages, it is the end of the speech report which is set off from the next intonation unit.

3.3 Cross-linguistic approaches to interaction, information structure, animacy, and morphosyntax (Ozerov; Haig et al.)

Another area of discourse production concerns the structuring and delivery of information: Producers typically have to navigate between their own planning and production efforts while considering the needs of their addressees and how to draw their attention to what is of most interest at any given point in discourse. Studies that address such questions are often confined to single languages and have been relying to a major extent on macro-categories like topic and focus, which upon closer inspection have been found to be not readily associable with specific means of expression across languages (Matić & Wedgwood 2013; Ozerov This volume, among others).

An alternative approach is taken by Ozerov (This volume), who conceives of information packaging as a level of structuring that is derivative of the clustering of distinct interactional moves in discourse production. Within this framework, Ozerov (this volume) investigates question formation and so-called detachment constructions together with a range of concomitant structural aspects, including prosodic aspects. Ozerov demonstrates that individual structures are more readily associated with specific interactional moves (e.g. getting addressee's attention through pitch) and/or addressee-oriented effects (e.g. pausing after mentioning an entity) that may be interpreted by addressees in certain ways (e.g. this entity is relevant for the immediately subsequent discourse). Comparing two unrelated languages, Anal Naga (Tibeto-Burman) and Ivrit (also known as Modern Israeli Hebrew), this study suggests that the low-level categories of individual interactional moves find similar expressions across different languages and are also a suitable basis for cross-linguistic comparisons.

A further aspect of information packaging relates to argument marking and argument expression, that is the alternation between full NPs versus pronouns and zero reference. Haig et al. (this volume) investigate the rates of full NPs across corpora from 15 languages, finding astonishingly stable rates of full lexical NPs across all corpora. This is noteworthy given that the corpus texts have not been controlled for content, unlike Bickel's (2003) or Stoll and Bickel's (2009) parallax corpora. Furthermore, this finding suggests an overall consistent response by discourse producers to motivations of informativeness and tellability on the one hand (pushing up the rate of lexical NPs) and to discourse coherence and referential continuity on the other hand (dragging down the rate of NPs).

The authors also investigate the related question of how full NPs are marked for their syntactic role, specifically the differential marking of NPs for different syntactic functions as an example of marking asymmetries, focusing here on the functions A and P in transitive clauses. The authors assess earlier accounts of differential argument marking in terms of frequencies of verbalization of participant role with different semantic properties (person, animacy, definiteness; cf. Haspelmath 2021). Their study suggests that while attested corpus frequencies do support the overall trend to overtly mark the less frequent member of a particular opposition, the results do not square up with the attested frequencies for different systems of differential marking of A and P in the world's languages, in particular the comparative rarity of differential A marking. For this, considerations of ambiguity in connection with efficiency, as explicated by Piantadosi et al. (2012), may be more relevant than corpus frequency.

3.4 Investigating social cognition: Purpose-built corpora (Barth et al.)

An aspect of spoken and signed discourse production that goes beyond specific structures of the texts themselves is how language users navigate the social reality they communicate in and about. This dimension of social cognition is relevant for any discourse in any language; yet of particular interest here is a comparison of languages and their use in diverse cultures, to find out how different cultural backgrounds (concepts, values) influence the way language users verbalize states of affairs in discourse. Also, matters of social cognition are mostly relevant in face-to-face interactions.

These issues are addressed by the Social Cognition project that Barth et al. (this volume) report on. Corpus-based research into social cognition across diverse, mostly only spoken and signed languages runs into the problem of finding sufficient data points that are comparable across languages, and are relevant for the research questions at hand. For instance, free narrative corpora normally do not contain many lexical NPs referring to humans since these are often salient discourse referents that tend to be expressed by non-lexical forms. Additionally, available cross-linguistic corpora rarely properly cover potential variation across language users (as mentioned in Section 2.4). Hence, to investigate lexical choices in reference to humans across diverse languages, Social Cognition Parallax Interview Corpus (SCOPIC) texts are elicited with the help of a picture task stimulus that is designed so as to elicit instances of human references, and likewise lexical choices in other domains. This optimally enables comparisons across languages and across individual users, since contexts are kept stable. Barth et al. (this volume) present four case studies on intra-language and intra-language user variation in a sample of thirteen languages using these data. Overall, the authors find substantial variation across users of a single language, which actually exceeds the variability between languages, especially in the case of research questions in semantic typology. These findings highlight the necessity for closer monitoring of community-level representativeness in CBT and for investigating methods that allow researchers to assess contribution of individual language users.

3.5 Corpus processing, annotation, and usability

Our focus on spoken-language material from understudied languages comes with the attendant burden of exceptionally labour-intensive processing and annotation of data. Such data are typically recorded and annotated together with native speakers during fieldwork, often in the absence of standardized orthographies, reference sources such as dictionaries and grammars, or other luxuries taken for granted in most of corpus linguistics. This is quite obviously the major reason why corpora of spoken- or signed-language texts from underdescribed languages remain relatively small. Moreover, only a small number of research questions, such as studies of word length, can be undertaken on the basis of transcriptions alone, while investigations of most prosodic, morphosyntactic, and other structures usually necessitate further annotation. All contributions in this volume detail the kind of corpus processing and annotation involved in the corpus development projects they build on, and we will not go into more detail here (cf. also Barth & Schnell 2021: Chapters 6 and 7).

A major issue in the further development of CBT is the balance between diversity and standardization of annotation schemes across a scientific community with very diverse theoretical backgrounds working on very diverse languages. Even Universal Dependency annotations, which have become a fairly widely-accepted standard in syntactic parsing of written corpora, have spawned a range of "dialectal" variants and offshoots. For example, Gerdes et al. (2021) use the variant annotation system SUD (Surface-Syntactic Universal Dependencies), which implements the opposite direction of dependency between adpositions and complements when compared to standard UD. The GRAID system (Haig & Schnell 2014) posits hierarchical levels of annotation, utilizing top-level categories such as ⟨`pro`⟩ ('free definite pronoun') as a required minimum, which can be (optionally) further specified, for instance as ⟨`rel_pro`⟩ 'relative pronoun'. The Hambam Corpus (Hamedan-Bamberg Corpus of Contemporary Spoken Persian) implements a simplified version of GRAID, GRAID-L,[5] which is largely interoperable with standard GRAID, and these examples could be multiplied.

Like corpus development more generally (cf. Hellwig et al., this volume), the development of corpus annotation schemes can profit from adapting the spirit of language documentation of keeping conventions as open, transparently documented, and flexible as possible to be amenable to further research agendas. While standards for phonemic transcriptions and the *Leipzig Glossing Rules* for morphemic annotation seem to be firmly established, it is still unclear whether in future developments of CBT further gold standards of annotation will evolve, especially for higher structural levels.

Similar issues arise in relation to the comparability of texts and corpora across languages and individual users. As discussed above, for some research agendas more immediate comparability is key, for instance when it comes to comparison of rates of zero reference or reference to humans. As with standards of corpus annotation, there is no gold standard regarding comparability of this order. In fact, most studies in this volume draw on texts that have originally been selected for recording for documentary reasons, including upon request from community members, showing that such uncontrolled data is amenable to a range of research questions (in the same way as likewise uncontrolled UD corpora are). While language documentation as conceived by Himmelmann (1998) does encompass the collection of more controlled data for specific descriptive and analytical purposes (though this has often fallen

[5] `https://multicast.aspra.uni-bamberg.de/resources/hambam/`

by the wayside in the reception and citation of this seminal article), there are good reasons from a collaborative documentary point of view to give precedence to communities' preferences, often for traditional narratives, and the contributions in this volume bear witness to the fact that such data are most useful for CBT. This is corroborated by a general corpus-linguistic view that seeks to determine various conditions on language use, so that confinement to experimentally elicited data would yield too narrow a scope of situational factors.

Finally, corpora need to be sustainably and accessibly archived with clear instructions how they can be used, in line with best practices of documentary linguistics (Gippert et al. 2006; Thieberger & Berez 1963) and of open and reproducible science (Wilkinson et al. 2016; Berez-Kroeker et al. 2018). Nearly all contributors to this special publication are in the process of building web-accessible corpora and hence contribute to the enterprise of open science and collaborative research that CBT relies on much more than other fields in linguistics.

4 Conclusions

Its initial success story notwithstanding, CBT is still in its infancy and unified research agendas and standards are still emerging. Considerable advances have already been made by researchers working primarily on digital corpora of written standardized languages. This volume broadens the scope of CBT by connecting it with language documentation, enabling a shift towards comparison based on context-embedded, naturally variable spoken and signed language usage, where the social and cognitive-articulatory factors that motivate typological generalizations are actually operative. As coverage of the world's languages in spoken and signed corpora grows, CBT will also be able to feed into a more holistic approach to areal and diachronic typology by investigating the conditioned use of linguistic structures by different users with different demographics and across social settings and communities with different cultural backgrounds. This development will incur major challenges in data collection and processing, including in particular linguistic annotations of various kinds for comparative purposes. While this amounts to an enormous undertaking, evoking a "sea change" in linguistics propagated by

Levinson & Evans (2010), we believe that the confluence of current advances in corpus-based typology and language documentation will bring us a significant step further in this direction.

References

Barth, Danielle & Evans, Nicholas. 2017. SCOPIC design and overview. In Barth, Danielle & Evans, Nicholas (eds.), *The Social Cognition Parallax Interview Corpus (SCOPIC)*: *A cross-linguistic resource* (*Language Documentation & Conservation* special publication 12), 1–23. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/24742).

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. This volume. 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).

Barth, Danielle & Schnell, Stefan. 2021. *Understanding corpus linguistics*. London: Routledge.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142. (https://doi.org/10.1515/9783110242652.110).

Bentz, Christian & Ferrer-i-Cancho, Ramon. 2016. Zipf's law of abbreviation as a language universal. In Bentz, Christian & Jäger, Gerhard & Yanovich, Igot (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tübingen: University of Tubingen. (https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558).

Berez-Kroeker, Andrea L. & Andreassen, Helene N. & Gawne, Lauren & Holton, Gary & Kung, Susan Smythe & Pulsifer, Peter & Collister, Lauren B. & The Data Citation and Attribution in Linguistics Group & The Linguistics Data Interest Group. 2018. *The Austin Principles of Data Citation in Linguistics (Version 1.0): Introduction and guidelines for annotators (Version 7.0)*. (http://site.uit.no/linguisticsdatacitation/austinprinciples/).

Berman, Ruth A. & Slobin, Dan I. 1994. *Relating events in narrative: A cross-linguistic developmental study*. Mahwah, NJ: Erlbaum.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Bickel, Balthasar. 2009. *Typological patterns and hidden diversity*. Plenary talk delivered at the 8th Biannual Meeting of the Association for Linguistic Typology, Berkeley, United States of America, 24 July 2013. (https://www.comparativelinguistics.uzh.ch/dam/jcr:00000000-774a-e877-0000-0000407727a0/alt2009bickel-plenary.pdf).

Bickel, Balthasar. 2015. Distributional typology. In Heine, Bernd & Narrog, Heiko (eds.), *The Oxford handbook of linguistic analysis*, 2nd edn. Oxford: Oxford University Press. (https://doi.org/10.1093/oxfordhb/9780199677078.013.0046).

Blasi, Damian & Cotterell, Ryan & Wolf-Sonkin, Lawrence & Stoll, Sabine & Bickel, Balthasar & Baroni, Marco. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In Korhonen, Anna & Traum, David & Màrquez, Lluís (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3938–3943. Florence: Association for Computational Linguistics.

Bromham, Lindell & Dinnage, Russell & Skirgård, Hedvig & Ritchie, Andrew & Cardillo, Marcel & Meakins, Felicity & Greenhill, Simon & Hua, Xia. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*. (https://doi.org/10.1038/s41559-021-01604-y).

Bybee, Joan L. & Pagliuca, William & Perkins, Revere D. 1990. On the asymmetries in the affixation of grammatical material. In Croft, William & Denning, Keith & Kemmer, Suzanne (eds.), *Studies in typology and diachrony*: *Papers presented to Joseph H. Greenberg on his 75th birthday*, 1–42. Amsterdam: John Benjamins.

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Chanard, Christian. 2015. ELAN-CorpA: Lexicon-aided annotation in ELAN. In Mettouchi, Amina & Chanard, Christian (eds.), *CorpAfroAs*: *The CorpAfroAs corpus of spoken AfroAsiatic languages*. (https://doi.org/10.1075/scl.68.website).

Cohen Priva, Uriel. 2017. Informativity and the actuation of lenition. *Language* 93(3). 569–597.

Comrie, Bernard. 1981. *Language universals and linguistic typology*. London: Blackwell.

Coupé, Christophe & Oh, Yoon & Dediu, Dan & Pellegrino, François. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5(9). eaaw2594. (https://doi.org/10.1126/sciadv.aaw2594).

Cutler, Ann & Hawkins, John A. & Gilligan, Gary. 1985. The suffixing preference: A processing explanation. *Linguistics* 23(5). 723–758. (https://doi.org/10.1515/ling.1985.23.5.723).

Cysouw, Michael. 2014. Inducing semantic roles. In Luraghi, Silvia & Narrog, Heiko (eds.), *Perspectives on Semantic Roles*, 23–68. Berlin: Mouton de Gruyter. (https://doi.org/10.1075/tsl.106.02cys).

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. (https://www.eva.mpg.de/fileadmin/content_files/linguistics/conferences/2015-diversity-linguistics/Dahl_slides.pdf).

Dingemanse, Mark & Blythe, Joe & Dirksmeyer, Tyko. 2014. Formats for other-initiation of repair across languages: An exercise in pragmatic typology. *Studies in Language* 38(1). 5–43. (https://doi.org/10.1075/sl.38.1.01din).

Dingemanse, Mark & Rossi, Giovanni & Floyd, Simeon. 2017. Place reference in story beginnings: A cross-linguistic study of narrative and interactional affordances. *Language in Society* 46(2). 129–158. (https://doi.org/10.1017/S0047404516001019).

Donohue, Mark. 1999. *A grammar of Tukang Besi*. Berlin: Mouton de Gruyter.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68(1). 81–138.

Dryer, Matthew S. 2013. Feature 101A: Expression of pronominal subjects. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/feature/101A).

Duranti, Alessandro. 1981. *The Samoan fono: A sociolinguistic study*. Canberra: Australian National University.

Duranti, Alessandro. 1997. *Linguistic anthropology*. Cambridge: Cambridge University Press.

ELAN developers. 2020. *ELAN (Version 6.2)*. Nijmegen: Max Planck Institute for Psycholinguistics. (https://archive.mpi.nl/tla/elan).

Frank, Austin F. & Jaeger, Florian. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci-08), Washington D.C., United States of America, 23–26 July 2008*. 939–944.

Futrell, Richard & Levy, Roger P. & Gibson, Edward. 2020. 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.

Gerdes, Kim & Kahane, Sylvain & Chen, Xinying. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa* 6(1). 1–31. (https://doi.org/{10.5334/gjgl.764}).

Gibson, Edward & Futrell, Richard & Piantadosi, Steven T. & Dautriche, Isabelle & Mahowald, Kyle & Bergen, Leon & Levy, Roger. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23(12). 1087. (https://doi.org/10.1016/j.tics.2019.09.005).

Gippert, Jost & Himmelmann, Nikolaus P. & Mosel, Ulrike (eds.). 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, Joseph H. (ed.), *Universals of language*, 73–113. Cambridge, MA: MIT Press.

Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

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).

Haig, Geoffrey & Schnell, Stefan (eds.). 2021. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. Version 2108. (https://multicast.aspra.uni-bamberg.de/).

Haig, Geoffrey & Schnell, Stefan & Schiborr, Nils N. This volume. 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. 2011. 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. (https://doi.org/10.1515/9783110260021.55).

Hale, Ken & Krauss, Michael & Watahomigie, Lucille J. & Yamamoto, Akira Y. & Craig, Colette & Masayesva Jeanne, LaVerne & England, Nora C. 1992. Endangered languages. *Language* 68(1). 1–42. (https://doi.org/10.1353/lan.1992.0052).

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. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Hellwig, Birgit & Defina, Rebecca & Kidd, Evan & Allen, Shanley & Davidson, Lucy & Kelly, Barbara F. This volume. Child language documentation: The sketch acquisition project. 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), 29–58. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74657).

Henrich, Joseph & Heine, Steven J. & Norenzayan, Ara. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3). 61–83. (https://doi.org/10.1017/S0140525X0999152X).

Hildebrandt, Kristine A. & Jany, Carmen & Silva, Wilson (eds.). 2017. *Documenting variation in endangered languages* (*Language Documentation & Conservation* special publication 13). Honolulu, HI: University of Hawai'i Press. (http://hdl.handle.net/10125/24754).

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(2). 161–195.

Himmelmann, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960.

Hymes, Dell H. 1961. Functions of speech: An evolutionary approach. In Gruber, Frederick C. (ed.), *Anthropology and education*, 55–83. Philadelphia, PA: University of Philadelphia Press.

Hymes, Dell H. 1962. The ethnography of speaking. In Gladwin, Thomas & Sturtevant, William C. (eds.), *Anthropology and human behaviour*, 13–53. Washington, D.C.: Anthropological Society of Washington.

Inbar, Maya & Grossman, Eitan & Landau, Ayelet N. 2020. Sequences of intonation units form a 1 Hz rhythm. *Scientific Reports* 10(1). 15846. (https://doi.org/10.1038/s41598-020-72739-4).

Izre'el, Shlomo & Mettouchi, Amina. 2015. Representation of speech in CorpAfroAs: Transcriptional strategies and prosodic units. In Mettouchi, Amina & Vanhove, Martine & Caubet, Dominique (eds.), *Corpus-based studies of lesser-described languages: The CorpAfroAs corpus of spoken AfroAsiatic*, 13–41. Amsterdam: John Benjamins.

Jaeggli, Osvaldo & Safir, Kenneth J. 1989. The null subject parameter and parametric theory. In Jaeggli, Osvaldo & Safir, Kenneth J. (eds.), *The null subject parameter*, 1–44. Dordrecht: Kluwer.

Jing, Yingqi & Widmer, Paul & Bickel, Balthasar. 2021. Word order variation is partially constrained by syntactic complexity. *Cognitive Science* 45(11). e13056. (https://doi.org/10.1111/cogs.13056).

Just, Erika & Čéplö, Slavomír. To appear. Differential object indexing in Maltese — a corpus based pilot study. In Turek, Przemysław & Nintemann, Julia (eds.), *Maltese: Contemporary changes and historical innovations*. Mouton de Gruyter: Berlin.

Koplenig, Alexander & Meyer, Peter & Wolfer, Sascha & Müller-Spitzer, Carolin. 2017. The statistical trade-off between word order and word structure — Large-scale evidence for the principle of least effort. *PLoS ONE* 12(3). e0173614. (https://doi.org/10.1371/journal.pone.0173614).

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, William. 1994. *Principles of linguistic change: Internal factors*. Malden, MA: Blackwell.

Levinson, Stephen C. & Evans, Nicholas. 2010. Time for a sea change in linguistics: Response to comments on 'The myth of language universals'. *Lingua* 12. 2733–2758. (https://doi.org/10.1016/j.lingua.2010.08.001).

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. 2021a. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*. (https://doi.org/10.1515/lingty-2020-0118).

Levshina, Natalia. 2021b. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations: A reverse-engineering approach. *Frontiers in Psychology* 12. 648200. (https://doi.org/10.3389/fpsyg.2021.648200).

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analysing talk*. Mahwah, NJ: Erlbaum.

Mansfield, John & Stanford, James N. 2017. Documenting sociolinguistic variation in lesser-studied indigenous communities: Challenges and practical solutions. In Hildebrandt, Kristine A. & Jany, Carmen & Silva, Wilson (eds.), *Documenting variation in endangered languages* (*Language Documentation & Conservation* special publication 13), 116–136. Honolulu, HI: University of Hawai'i Press. (`http://hdl.handle.net/10125/24751`).

Matić, Dejan & Wedgwood, Daniel. 2013. The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis. *Journal of Linguistics* 49(1). 127–163. (`https://doi.org/10.1017/S0022226712000345`).

Mayer, Mercer. 1969. *Frog, where are you?* New York: Dial Books for Young Readers.

Mayer, Thomas & Cysouw, Michael. 2014. Creating a massively parallel Bible corpus. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014.*

McEnery, Tony & Wilson, Andrew. 2001. *Corpus linguistics: An introduction.* Edinburgh: Edinburgh University Press.

Mettouchi, Amina & Vanhove, Martine. This volume. Prosodic segmentation and cross-linguistic comparison in CorpAfroAs and CorTypo: Corpus-driven and corpus-based approaches. 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), 59–113. Honolulu, HI: University of Hawai'i Press. (`https://hdl.handle.net/10125/74658`).

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.

Moran, Steven & Schikowski, Robert & Pajović, Danica & Hysi, Cazim & Stoll, Sabine. 2016. The ACQDIV database: Min(d)ing the ambient language. *Proceedings of the 10th Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2020.* 4423–4429.

Newman, Paul. 2013. *The law of unintended consequences: How the endangered languages movement undermines field linguistics as a scientific enterprise.* Seminar talk delivered at the School of Orential and African Studies, University of London, London, United Kingdom, 15 October 2013.

Ozerov, Pavel. This volume. This research topic of yours – is it a research topic at all? Using comparative interactional data for a fine-grained reanalysis of traditional concepts. 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), 233–280. Honolulu, HI: University of Hawai'i Press. (`https://hdl.handle.net/10125/74662`).

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).

Piantadosi, Steven T. & Tily, Harry J. & Gibson, Edward. 2012. The communicative function of ambiguity in language. *Cognition* 122(3). 1280–1291.

Pimentel, Tiago & Meister, Clara & Salesky, Elizabeth & Teufel, Simone & Blasi, Damián & Cotterell, Ryan. 2021. A surprisal–duration trade-off across and within the world's languages. *Preprint published on arXiv* 2109.15000. (http://arxiv.org/abs/2109.15000).

Rizzi, Luigi. 1982. *Issues in Italian syntax*. Dordrecht: Foris.

Roberts, Ian & Holmberg, Anders. 2009. Introduction: Parameters in minimalist theory. In Biberauer, Theresa & Holmberg, Anders & Roberts, Ian & Sheehan, Michelle (eds.), *Parametric variation*: *Null subjects in minimalist theory*, 1–57. Cambridge: Cambridge University Press.

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).

Salesky, Elizabeth & Chodroff, Eleanor & Pimentel, Tiago & Wiesner, Matthew & Cotterell, Ryan & Black, Alan W. & Eisner, Jason. 2020. A corpus for large-scale phonetic typology. *Proceedings of the 58nd Annual Meeting of the Association for Computational Linguistics (ACL'20)*. 4526–4546. (https://doi.org/10.18653/v1/2020.acl-main.415).

Schegloff, Emanuel A. 2006. *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.

Schnell, Stefan & Barth, Danielle. 2018. Discourse motivations for pronominal and zero objects across registers in Vera'a. *Language Variation and Change* 30(1). 51–81.

Schnell, Stefan & Schiborr, Nils N. In press. Cross-linguistic corpus studies in linguistic typology. *Annual Review of Linguistics*.

Seifart, Frank. This volume. 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 & Evans, Nicholas & Hammarström, Harald & Levinson, Stephen C. 2018. Language documentation 25 years on. *Language* 94(4). e324–e345. (https://doi.org/10.1353/lan.2018.0070).

Seifart, Frank & Haig, Geoffrey & Himmelmann, Nikolaus P. & Jung, Dagmar & Margetts, Anne & Trilsbeek, Paul (eds.). 2012. *Potentials of language documentation: Methods, analyses, and utilization* (*Language Documentation & Conservation* special publication 3). Honolulu, HI: University of Hawai'i Press. (http://hdl.handle.net/10125/4510).

Slobin, Dan I. (ed.). 1985. *The crosslinguistic study of language acquisition, Volume 1: The data.* Mahwah, NJ: Erlbaum.

Stanford, James N. & Preston, Dennis R. (eds.). 2009. *Variation in indigenous and minority languages.* Amsterdam: John Benjamins.

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.

Stoll, Sabine & Bickel, Balthasar. 2013. Capturing diversity in language acquisition research. In Bickel, Balthasar & Grenoble, Lenore A. & Peterson, David A. & Timberlake, Alan (eds.), *Language typology and historical contingency*, 195–216. Amsterdam: Benjamins.

Thieberger, Nicholas & Berez, Andrea L. 1963. Linguistic data management. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.

Tomlin, Russell. 1986. *Basic word order: Functional principles.* London: Routledge.

Torres Cacoullos, Rena & Travis, Catherine E. 2018. *Bilingualism in the community: Code-switching and grammars in contact.* Cambridge: Cambridge University Press. (https://doi.org/10.1017/9781108235259).

Torres Cacoullos, Rena & Travis, Catherine E. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3). 653–692.

Vollmer, Maria C. 2019. *How radical is pro-drop in Mandarin? A quantitative corpus study on referential choice in Mandarin Chinese.* MA thesis, University of Bamberg.

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77–94.

Wälchli, Bernhard & Sölling, Arnd. 2013. The encoding of motion events: Building typology bottom-up from text data in many languages. In Goschler, Juliana & Stefanowitsch, Anatol (eds.), *Variation and change in the encoding of motion events*, 102–125. Amsterdam: John Benjamins.

Wilkinson, Mark D. & Dumontier, Michel & Aalbersberg, IJsbrand J. & Appleton, Gabrielle & Axton, Myles & Baak, Arie & Blomberg, Niklas & alii. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1). 1–9. (`https://doi.org/10.1038/sdata.2016.18`).

Zakharko, Taras & Witzlack-Makarevich, Alena & Nichols, Johanna & Bickel, Balthasar. 2017. *Late aggregation as a design principle for typological databases*. Paper presented at the ALT Workshop on Design Principles of Typological Databases, Canberra, Australia, 15 December 2017.

Zeman, Daniel & Nivre, Joakim & Abrams, Mitchell & alii. 2021. *Universal Dependencies 2.8*. Prague: Universal Dependencies Consortium. (`https://hdl.handle.net/11234/1-3687`).

Zipf, George K. 1935. *The psycho-biology of Language: An introduction to dynamic philology*. Cambridge, MA: MIT.