Stefan Schnell*, Nils Norman Schiborr and Geoffrey Haig

Efficiency in discourse processing: Does morphosyntax adapt to accommodate new referents?

https://doi.org/10.1515/lingvan-2019-0064

Abstract: The introduction of new referents into discourse has traditionally been regarded as a major challenge to language processing, for which speakers deploy specific syntactic configurations, guided by the speaker's assessment of the recipient's state of mind ('recipient design'). In this paper we probe these assumptions against discourse data from nine languages. We find little evidence for specialized syntactic configurations accommodating new referents; the only notable exception is the association of new reference with direct objects, suggests that linking new referents to already established discourse frames through a transitive construction is preferable to isolating them in an intransitive one. Where specific intransitive predicates are indeed found to host new referents, we find this to be motivated primarily by semantic considerations. Contrary to long-held assumptions, we conclude that the cognitive challenge of referent introduction is only weakly reflected in morphosyntax; instead, discourse production is most efficient when new referents are integrated seamlessly with content-driven demands of the narration.

Keywords: efficiency, discourse processing, recipient design, information pressure, corpus-based typology

1 Introduction

There are in essence two sides to language processing, that of production and that of comprehension (see e.g. MacDonald 2013; McDaniel et al. 2015, among many others). In this study we are chiefly concerned with the former aspect of discourse processing as regards referent introduction. Discourse production should in principle follow considerations of efficiency, as speakers strive to expend minimal effort to achieve a desired communicative outcome (Hawkins 2004). This optimized state is complicated, however, by the needs of discourse planning as well as "recipient design". The basic idea of the latter is that speakers take their addressee's perspective into account when structuring utterances so as to ease comprehension on behalf of the addressee. While recipient design has been a cornerstone of usage-oriented linguistics (Blokpoel et al. 2012; Sacks et al. 1974), more recent psycholinguistic work has pointed out difficulties in distinguishing comprehension-related factors of recipient design from factors related to production and planning during language processing (Arnold 2008; MacDonald 2013), a point we will return to in the conclusions.

In traditional functionalist linguistic work, the same principle has been adapted to the area of production (see Arnold 2010 for an overview): while speakers follow production-related requirements, they are also assumed to monitor addressee's awareness states and adapt referent introduction accordingly (Ariel 2014 [1990]; Chafe 1994, 1987; Prince 1998, 1981). From a recipient design perspective, the introduction of new referents into the current discourse has been considered a major processing challenge because of the lack of a pre-established cognitive representation in the mind of the addressee. Introducing more than one new referent at a time is in fact seen as excessively challenging for interlocutors, who need to monitor and differentiate an increased set of referents. This view is reflected in Chafe's (1987: 32) famous 'one new concept at a time'

***Corresponding author: Stefan Schnell**, Department of General Linguistics, University of Bamberg, Bamberg, Germany; and ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia, E-mail: stefan.schnell@uni-bamberg.de, https://orcid.org/0000-0003-2036-2263

Nils Norman Schiborr and Geoffrey Haig, Department of General Linguistics, University of Bamberg, Bamberg, Germany, E-mail: nils-norman.schiborr@uni-bamberg.de (N.N. Schiborr), geoffrey.haig@uni-bamberg.de (G. Haig)

constraint. Du Bois (2003a: 38) thus refers to the "cognitively demanding task of introducing new or low-accessible information into the discourse."

A number of linguists have proposed that speakers deploy certain structures in response to these challenges: for instance, (Lambrecht 1994) formulates a general principle of separation of role and reference (PSRR), which states that referent information be established before elaborating on it to relevant states-of-affairs. Lambrecht identifies presentational and detachment constructions as implementations of the PSRR (see also Prince 1998 on the referent-introducing function of left-dislocations).

In this paper we take up the claim that morphosyntax is sensitive towards a higher processing load, with particular constellations preferentially adapted towards dealing with the increased cognitive effort associated with new information. Specifically, we re-evaluate the special role of intransitive subjects as a preferred entry point for new information (Du Bois 1987a, 2003a, 2003b, 2017) as well as specific constructions traditionally associated with referent introductions, such as presentationals (Abbott 1993; Birner and Ward 1998). In doing so, we draw on an unprecedented, richly annotated corpus of natural spoken discourse from nine languages (Haig and Schnell 2015). We analyse observable patterns of referent introduction in language production in regards to the considerations outlined above. We find limited support for the zero-sum approach to information processing, which leads us to propose alternative explanations better matching our corpus-based findings and a revised notion of efficiency.

2 Efficiency considerations in the introduction and retrieval of referents

Since Chafe's (1976) seminal work, the monitoring of referent information has been conceived of as an inherently scalar concept, whereby forms of reference reflect different degrees of referent accessibility at any given point in discourse, as assessed by the speaker on behalf of the addressee (Ariel 2014[1990]; Givón 1983). Thus the shortest possible, least informative forms (free and clitic pronouns, person affixes, zero anaphors) are used where referent retrieval can rely on contextual clues; longer, more informative forms conversely correspond to a lower likelihood of contextually driven referent retrieval. As such, choice of form is also a cue for the addressee to correctly retrieve the intended referent. A short form suggests a highly salient and activated referent, whereas a longer, more informative form suggests that the intended referent is not among the most accessible or activated ones. From the viewpoint of efficiency of production, speakers avoid longer, more informative forms (e.g. full lexical noun phrases) in order to minimize effort (cf. the minimize form principle in Hawkins 2004), relegating these to contexts where accessibility is deemed too low to use reduced forms. In sum, form–function pairing in terms of accessibility works both ways: on the production side it serves as a calculus for the least effort in referential choice, vis-à-vis recipient design requirements, whereas on the comprehension side it works as a retrieval script for discourse referents. Hence, opting for a more explicit form than appropriate in a given context is not merely less efficient and hence wasteful for the speaker, but can also be misleading and in this way increase processing costs for the recipient. Gordon and Chan (1994), for instance, find significantly longer processing times for "overexplicit" references, that is where a more explicit than necessary form was used in an experimental discourse processing task (see further Kwon et al. 2010).

There seems to exist wide agreement among linguists working in a functionalist tradition since Chafe (1976) that these principles of referent retrieval extend to referent introduction, which on this view constitutes merely an extreme case of low-accessibility reference. Although not all discourse-new referents pose difficulties for referent retrieval to the same extent (Prince 1981),[1] they are often jointly considered particularly

1 Interlocutors do not enter a communicative situation with a blank slate after all, and so world and cultural knowledge as well as the immediate experience of its physical and social environment play a significant role (see Ariel 2014[1990]: Ch. 1). Consider the differences in information status of NPs such as *the sun*, *the Prime Minister* (vs. *a woman/man*), or *a tree* in first-mention position. In this paper we basically consider all discourse-new referents *new*, but we also test for differences between 'brand-new' and 'evoked' ones in Section 3.2 below.

cognitively demanding for discourse processing, and therefore specifically monitored by interlocutors (Chafe 1987; Du Bois 1987a).

This verdict is echoed by research investigating discourse and information structure from various related perspectives. One such perspective is that of discourse cohesion: A sequence of states-of-affairs expressed by a succession of individual clauses forms a "[…] coherent whole […] recognized as such by speakers of a language" (Foley 2007: 354; see also Kehler 2002, 2004; Polanyi 1995; Polanyi and Scha 1988; Polanyi et al. 2003). Reference plays a significant role in the establishment of discourse cohesion together with specific morphosyntactic structures that explicitly mark coherence relations between sentences (Ward and Birner 2004: 153; Chafe 1994: 83), for example left or right dislocation, cleft constructions, inversion (Prince 1998; Ward and Birner 2004), or diatheses (Foley 2007). Hence, co-reference relations between discourse-given referents act implicitly towards discourse coherence, since they link incoming new information about states-of-affairs to already identifiable participants ("ties" in the terminology of Halliday and Hasan 1976).[2] Discourse-new referents disrupt discourse coherence since interlocutors need to register a new referent in their discourse model (Du Bois 2003b) in addition to processing a new state-of-affairs. However, most of the experimental research paradigm on discourse cohesion is concerned with the resolution of pronominal references, and has little to say on the processing of new referents (cf. the overview in Holler and Suckow 2016).

A more holistic model of these principles is Du Bois' hypothesis of preferred argument structure (PAS) (1987a, 2003a, 2003b, 2017), which considers not just special syntactic structures, but syntactic argument structure in general to work in service of new referent processing. Du Bois ascribes a dual function to argument positions, acting not only as syntactic representations of participant roles (Andrews 2007; Dixon 1994) but also as locations for specific operations in information processing (labelled "pragmatic linking" in Durie 2003).

Thus, while the syntactic function A ('subject of a transitive clause') is found to be avoided as a location for new referents, both the S ('subject of an intransitive clause') and P ('direct object') roles are open for introductions,[3] hence defining a "predictable locus for unpredictable work" (Du Bois 2003b: 47) where addressees may anticipate new referents to preferentially occur.[4]

Intransitive constructions are particularly relevant in this regard, since they do not contain further core arguments and hence allow a new referent to enter discourse detached from other referents, abiding by Lambrecht's above-mentioned PSRR and related constraints. Relevant intransitive predicates are also often regarded as "semantically bleached" (e.g. *come*, *arrive*, *appear*, etc.), yielding little more conceptual information than the (incurred) presence of a new entity (cf. Du Bois 1987a). This makes them convenient vehicles for offloading referent introductions, as they can be added more or less freely for this purpose: "Speakers need not say everything in one clause […] Facing cognitive constraints […], speakers can simply mobilize their planning capacity to organize a series of successive clauses." (Du Bois 2003a: 73).

In a similar vein, Lambrecht (1994: 176) points to presentational constructions, such as English *there is*, as serving to promote referents to topic status, including the introduction of new referents into discourse. Du Bois (1987a) points out that the extent to which specialized syntactic resources are mobilized for the accommodation of new referents may vary according to local context. Du Bois suggests they are most operative where information pressure (i.e. the "density" of discourse referents in a stretch of discourse) is particularly high.

In sum, if introducing new referents is an overall costly process, a zero-sum efficiency perspective of grammar would predict a trade-off, with processing costs offset by compensatory strategies in the way speakers formulate messages. In the following, we present findings from corpus-based studies on the

2 This roughly corresponds to the last of Givón's (1983: 7–8) discourse continuities in terms of theme, action, and participant/topic.

3 Du Bois adopts Dixon's (1987, 1994) conception of S, A, and O, whereas in this paper we follow Andrews (2007) definition for labelling S, A, and P (= O). On this latter account, A and P are respectively defined on the basis of the agent and patient argument of prototypical transitive verbs in transitive clauses, e.g. *kill* or *smash*; the labels A and P are then assigned to those NPs in a two-argument construction that are marked in the same way as agent and patient in a prototypical transitive clause. S is the sole core argument in an intransitive clause.

4 Many languages mark definiteness distinctions on the NP level, which would typically capture the given-new distinction. We neglect this aspect of NP syntax here, as is done in most of the PAS literature.

interaction of new information and syntax: first, we test the assumed association of specific syntactic roles (in particular, intransitive subjects) with new information; second, we consider the impact of information pressure. Lastly, we investigate the role of presentational constructions in referent introduction.

3 Referent introduction across corpora

In this section we report on findings from our cross-corpus investigation of referent inroductions across nine spoken language corpora. We are specifically concerned with three issues here: first, we test the predictions of Du Bois' (1987a) preferred argument structure hypothesis, in particular whether the S role shows a generally higher proportion of new mentions, and whether it is specifically selected for the purpose of referent introduction. Second, we consider two factors that should relate to different degrees of cognitive challenges of introductions, namely that of information pressure and different information status of discourse-new mentions. The expectation here is that high information pressure as well as brand-newness of referents will incur higher processing costs of referent introductions and thus trigger the deployment of S as the preferred locus of introductions. Third, we take up the question whether such S roles are used particularly with specific predicates that are specialized for referent introductions to some extent, for instance presentational constructions or certain types of motion verbs that are often considered semantically bleached in the PAS literature, such as *come*, *approach* (Du Bois 2003a).

Our study is based on spoken corpora from nine languages from the Multi-CAST collection (Haig and Schnell 2015; version '2001' from January 2020).[5] Table 1 provides a summary of the sample.

Cypriot Greek	Indo-European, Greek	(Hadjidas and Vollmer 2015)
English	Indo-European, Germanic	(Schiborr 2015)
Mandarin	Sino-Tibetan, Sinitic	(Vollmer 2020)
Nafsan	Austronesian, Oceanic	(Thieberger and Brickell 2019)
Northern Kurdish	Indo-European, Iranian	(Haig et al. 2019)
Sanzhi Dargwa	Nakh-Daghest., Dargin	(Forker and Schiborr 2019)
Teop	Austronesian, Oceanic	(Mosel and Schnell 2015)
Tulil	Papuan, Taulil-Butam	(Meng 2019)
Vera'a	Austronesian, Oceanic	(Schnell 2015)

The texts in Multi-CAST are original, unelicited narratives, and predominantly monologic. All corpora have been annotated for the form, syntactic function, and semantic properties of core and oblique arguments (with the GRAID scheme, Haig and Schnell 2014), as well as for referent identity and the information status of new referents (with the RefIND scheme, Schiborr et al. 2018). The annotation practices are described in detail in the guidelines for the two schemes, both available from the Multi-CAST website. A concise summary of information on the languages, our methodology, and the quantitative analyses presented below can also be found in the supplementary material published alongside this article. The combination of these different levels of annotation allows us to determine the semantic properties of every discourse referent and the formal properties of any of its mentions, as well as locate its introduction into discourse.

In the following, a referent is deemed 'new' on its first verbalization in a discourse, and 'given' on all subsequent mentions. While we treat all referents as equally new upon their first mention in a given text, we do examine possible effects of differences in information status among such discourse-new referents in Section 3.2. Additionally, we only consider expressions (i) that clearly evoke an identifiable referent, (ii) whose underlying referent is mentioned at least twice in a text (i.e. its introduction and one further reference), (iii) that

5 Accessible online at https://multicast.aspra.uni-bamberg.de/. See Schnell and Schiborr (2018) for an overview, and Schiborr (2018) for a companion R package.

Table 1: Statistical overview of the referents and mentions in the corpus data.

				Referents		Mentions	
Corpus	**Texts**	**Spkrs.**	**Clauses**	**All**	**Human**	**All**	**Human**
Cypriot Greek	3	1	1071	160	70	1209	853
English	4	3	4184	897	258	3790	1398
Mandarin	3	3	1194	195	82	1547	1068
Nafsan	9	4	1012	161	66	1306	819
North Kurdish	2	1	1359	176	55	1615	1018
Sanzhi Dargwa	8	4	1066	170	76	1011	665
Teop	4	4	1302	133	55	1433	1016
Tulil	6	5	1264	226	70	1683	755
Vera'a	10	10	3608	423	144	4770	3377
Total	49	35	16060	2541	876	18364	10969

are in the third person (i.e. excluding first and second person). Clausal references such as headless relative clauses are excluded, as are complement clauses.

3.1 The loci of new referents

Figure 1 shows the cross-corpus distribution of the proportion of discourse-new mentions among all mentions of referents in each corpus (i.e. what proportion of mentions in a given role is new introductions?).[6] In order to control for the evidently considerable degree of cross-corpus (and cross-text) variation, we have fitted a generalized log-linear mixed-effects model to the data, with frequency as response, role (S + A vs. rest; $\beta = 0.119$, $p < 0.001$), newness ($\beta = -1.490$, $p < 0.001$), and role × newness ($\beta = -1.157$, $p < 0.001$) as fixed effects, and corpus ($\sigma = 0.620$) and text ($\sigma = 0.485$) as random effects. The model indicates that referential mentions are not evenly distributed across roles and newness; specifically, the P role and all other non-core argument positions in our corpus data show higher proportions of new mentions compared to the S and A roles. The goodness of fit for the entire model (as per Nakagawa and Schielzeth 2012) is $R^2_C = 0.994$, excluding random effects $R^2_M = 0.657$; this and the standard deviations for the random effects show that while cross-corpus and cross-text variation does play a role, we can nevertheless identify an underlying cross-linguistic tendency in the distribution of new mentions across roles. A second model with the same parameters contrasting only the two subject roles (S vs. A: $\beta = 1.457$, $p < 0.001$; newness: $\beta = -2.891$, $p < 0.001$; S vs. A: × newness: $\beta = 1.098$, $p < 0.001$; corpus: $\sigma = 0.619$, text: $\sigma = 0.485$; $R^2_C = 0.996$, $R^2_M = 0.774$) shows that the S role has a higher proportion of new mentions than A, but the difference between subjects and non-subjects seen in the first model is still noticeably greater than the difference between S and A in the second. The A role in particular is extremely unlikely to host new referents (cross-corpus mean $m = 4.1\%$, standard deviation $\sigma = 2.5\%$); S shows only slightly higher proportions of referent introductions ($m = 8.3\%$, $\sigma = 3.8\%$), though with a somewhat wider spread of values across corpora (3–16% vs. 2–9%). Conversely, non-subject arguments are more likely and approximately equally likely to represent new referents ($m = 19.9\%$, $\sigma = 6.5\%$). This suggests that, with the exception of A and S, no specific argument positions are systematically associated with particular rates of new mentions. While there is thus little evidence in support of the specialization of certain roles for new information, casting doubt on Du Bois' notion of "predictable loci", the data do suggest a broad subject versus non-subject distinction, at least from a general cross-linguistic perspective. This is in line with the generally

6 A brief explanation of Tukey boxplots: the solid grey box shows the interquartile range (ranging from the first to the third quartile); the horizontal line within it indicates the median of the data (i.e. the second quartile); the vertical 'whiskers' extending from either end of the box include the most extreme datum not deemed an outlier; finally, the X marks show outliers from the central distribution, here defined as data greater (less) than the third (first) quartile plus (minus) 1.5 times the interquartile range.

accepted association between the subject role and given information. The evidence for intransitives assuming a special role, however, is comparatively weak.

Figure 2 shows the distribution of the proportion of different roles among all new and all given mentions in each corpus (i.e. what proportion of given and new mentions is in which role?). This perspective on the data corresponds to the idea that speakers produce certain clause constructions with the intent of placing new referents in certain argument structure positions; whether this is in fact a realistic model of discourse production is an open question which we will take up again in Section 4 (but see Haig and Schnell 2016 for relevant remarks). What we find here is that it is P that harbours the largest shares of new mentions (m = 35.0%, σ = 7.9%), followed by adjuncts ('other' m = 23.8%, σ = 6.5%) and S (m = 21.0%, σ = 6.1%), whereas the proportion of the A function (m = 5.6%, σ = 1.5%) as well as that of non-core arguments (m = 7.3%, σ = 4.0%) are particularly low. While these findings meet the predictions for A, they do not square with the predictions for the non-core argument roles, which are assumed to play a bigger role in referent introduction according to PAS theory (Du Bois 1987a: 831). The S role occupies an intermediate position between P on the one hand and A and non-core arguments on the other.

Comparing these distributions with those of given referents, what is particularly remarkable is that S has a much higher share among given referents than among new ones (m = 33.1%, σ = 6.4%; almost 1.6 times larger), which suggests that, if anything, S is "selected" more for referent tracking rather than for introduction. More accurately, however, the observed patterns are simply the result of the S function being significantly much more frequent than other functions, with intransitive constructions typically accounting for roughly two thirds of clauses in any of our corpora (m = 63.7%, σ = 5.6%).

In sum, we find that the only two roles that stick out from the perspective of referent introduction are P and adjuncts. We will present a tentative explanation for this pattern later in Section 4, in contrast to the perceived view of recipient design and efficiency.

3.2 Information pressure and brand-newness

Earlier we mentioned two circumstances under which referent introduction could be regarded as particularly demanding. One of these is high information pressure, that is a context in which a high number of discourse

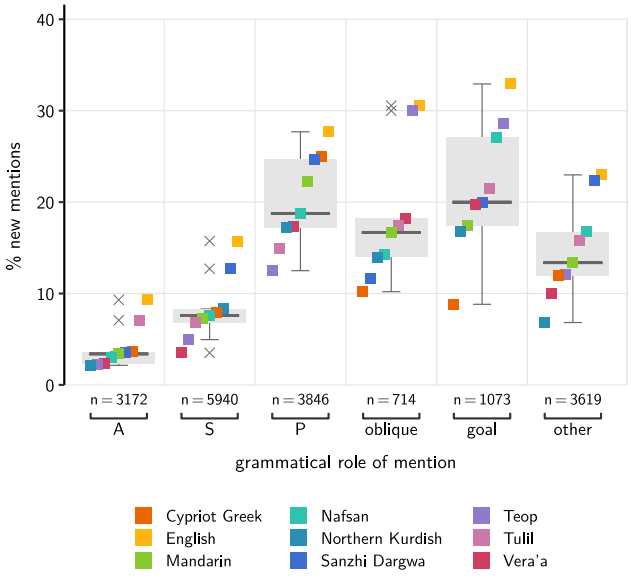

Figure 1: Proportions of new mentions among all mentions in six grammatical roles across nine languages. 'A', 'S', and 'P' are, respectively, the subjects of transitive and intransitive clauses, and direct objects. The 'other' category subsumes adjuncts and various other positions of marginal frequency, such as noun phrase-internal possessives and nominal predicates.

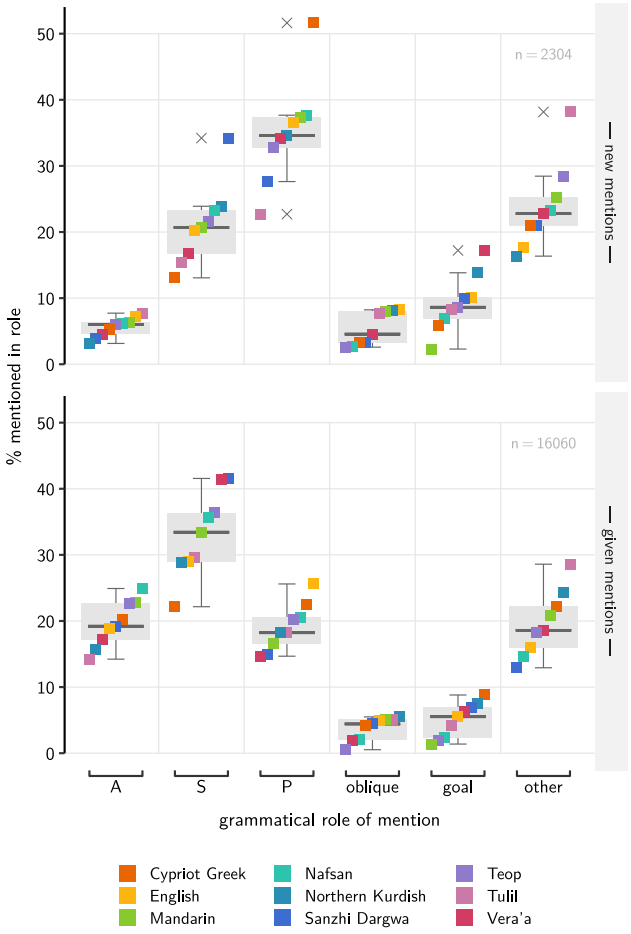

Figure 2: Proportions of grammatical roles among all new (top) and all given (bottom) mentions across nine languages.

referents are co-present in a given stretch of discourse, hence complicating the task of keeping a new referent distinct from previously established referents. The prediction is that increased information pressure would correlate with an increased deployment of new information in the S role. Since Du Bois (1987b), information pressure has been measured as a ratio of discourse referents against text length, the latter typically measured in number of clause units. We provide associated figures of this kind among the supplementary materials. Here, we instead take a more localized perspective, which we believe to be more revealing.

In Figure 3, the number of other referents co-present in a local context of three clauses preceding the point of introduction of a new referent is juxtaposed with the role of the newly introduced referent, with each boxplot showing the distribution across the nine corpora. While the S role shows a considerably higher newness level in contexts where no or only one other referent is co-present, its share decreases as soon as more than one other referent is present (Spearman's rank correlation coefficient $\rho = -0.47$), sharply contradicting the idea that S serves as a specialized locus for referent introduction under higher information pressure. Instead, it is the P role that hosts an increasing proportion of new referents as the number of co-present referents rises ($\rho = 0.19$). Moreover, non-core roles (including obliques, goals, and other roles from previous figures) show by far the steepest gains ($\rho = 0.20$), especially in very-high-pressure contexts. The data thus speak against the notion that the S role, and intransitive clauses more generally, serve as a figurative "escape valve" against high information pressure and excessive processing demands (cf. Durie 2003).

Turning now to the information status of discourse-new referents, we apply a rough distinction between (i) those referents that are either in some way evoked by the discourse context, including frame-semantic

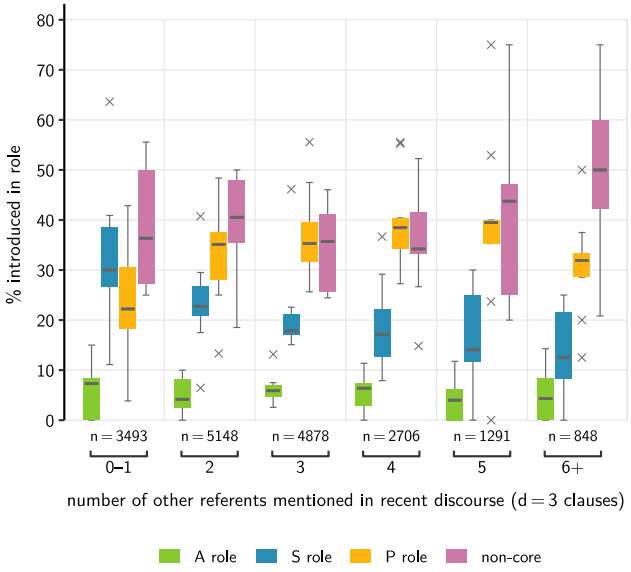

Figure 3: Proportions of new referents in A, S, P, and non-core roles by the number of other referents mentioned in the three preceding clauses, across nine languages.

considerations, or are uniquely identifiable based on world knowledge, which we conjointly label 'bridging (anaphora)', and (ii) those which are not inferable in such a way, labelled 'brand new'. Assuming that the latter type of referent is more processing-demanding due to the lack of contextual clues, we would expect a higher frequency of brand-new referents in syntactic positions that are deemed to be specialized for processing new information.

As can be seen from Figure 4, which shows the proportion of brand new (vs. bridging) introductions among all introductions across corpora (i.e. what proportion of introductions in a given role is brand new?), this expectation is not borne out, as we find no appreciable difference between the two types of

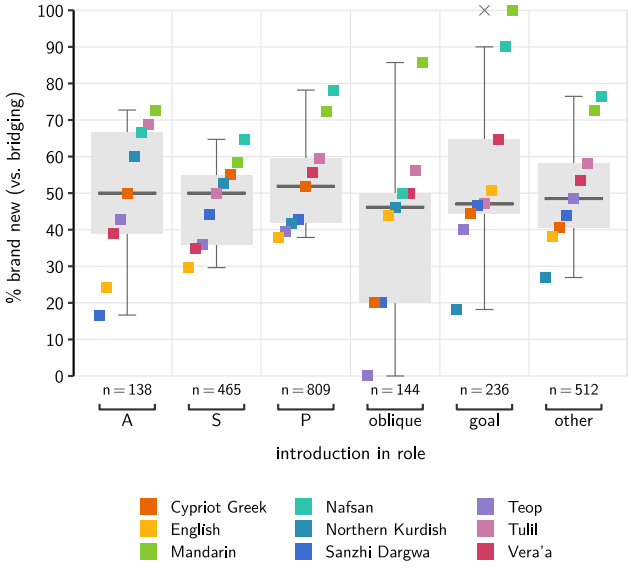

Figure 4: Proportions of brand new (i.e. non-inferable) referents among all discourse-new referents in six grammatical roles across nine languages.

introduction on a cross-linguistic scale (cross-corpus mean of means for each role m_m = 49.6%, σ = 5.0%), although inter-linguistic differences do exist (mean standard deviation within each role m_σ = 18.8%, σ = 5.6%). In principle, these findings could be taken to indicate that differences in information status among discourse-new referents are either irrelevant, or that argument positions are simply not sensitive to such differences.

3.3 The role of more specific predicates and presentational constructions

So far, we have been concerned entirely with the role of different argument positions. These correspond to fairly general semantic roles, such as 'agent-like' for the A function, 'patient-like' for the P-function, and so on. The S role in particular is in fact open to a wide range of semantic roles, and we next turn to the question whether referent introductions are associated with specific types of predicates and clause constructions. This includes presentational constructions such as English *there is*, which is often regarded as specialized for referent introductions to a considerable extent.

Since the comparison of lexical predicates requires the occurrence of similar types, we do not rely on our free narrative corpora, but use instead two Pear story corpora, an English one from the appendix of Chafe (1980; 20 stories), and a Persian one (Adibifar 2016; 29 stories). We specifically focus on the five main human participants in the film, namely the (i) pear picker, (ii) the man leading the goat, (iii) the boy who steals a basket of pears, (iv) the girl on the bike, and (v) the three boys who help the first boy pick up the spilled pears. Note that not all characters are mentioned in all retellings of the film.

Figure 5 shows that the pear picker is the only referent to be introduced to a greater degree by way of a presentational construction (37.9% in Persian, 60.0% in English). Other than that, presentationals play some role only in the introduction of the three boys, and only in the English stories (35.0%). The other three referents are introduced mainly as S arguments of motion predicates (e.g. *a boy comes along*) (mean m = 68.3% in Persian, m = 56.9% in English) While motion predicates are regarded as semantically bleached in much of the PAS literature, they seem to be restricted here to those contexts where a character in fact enters a scene of the film through some motion, be it riding on a bike or walking towards the pear tree. In other words, the use of motion predicates seems to be motivated primarily by semantic considerations of event content rather than strict information management.[7] This means that referent introduction follows primarily local semantic considerations, so that when the goatherd is introduced as a subject of *come*, it is because he is moving towards the scene from the perspective of the viewer. The girl presents a mixed case, showing a more even share of S in motion events and P, which seems to suggest that narrators exercise greater freedom to present her either as coming towards the boy on her bike or as seen or met by the boy who had been established as a focal referent before.

In sum, the findings from the two Pear story corpora suggest that presentational constructions are the preferred option only at the outset of the narrative and/or where the characters in question are not primarily engaged in an activity, and hence merely appear in the scene, typically after a cut, as is the case for the three boys. Presentationals, then, could be considered interactional, attention-directing devices for signalling a new scene; the association with new referents would thus be epiphenomenal of a broader interactional function, rather than their primary purpose.[8]

7 It should be noted that half of all introductions (50.0%, 111 out of 222) are elaborated in some way (e.g. *there was a man picking pears*; *a boy with a bike comes along*), and that elaboration is especially common with motion predicates (71.0%, 71 out of 100). Note that elaboration of this kind is not expected under the hypothesis that such motion predicates are semantically bleached and serve primarily referent introduction.

8 In this regard, it is worth noting that presentational constructions are also frequently found with given referents, including in the Pear stories, where the three baskets are often taken up by way of a *there are* constructions (cf. Abbott 1993, 1997).

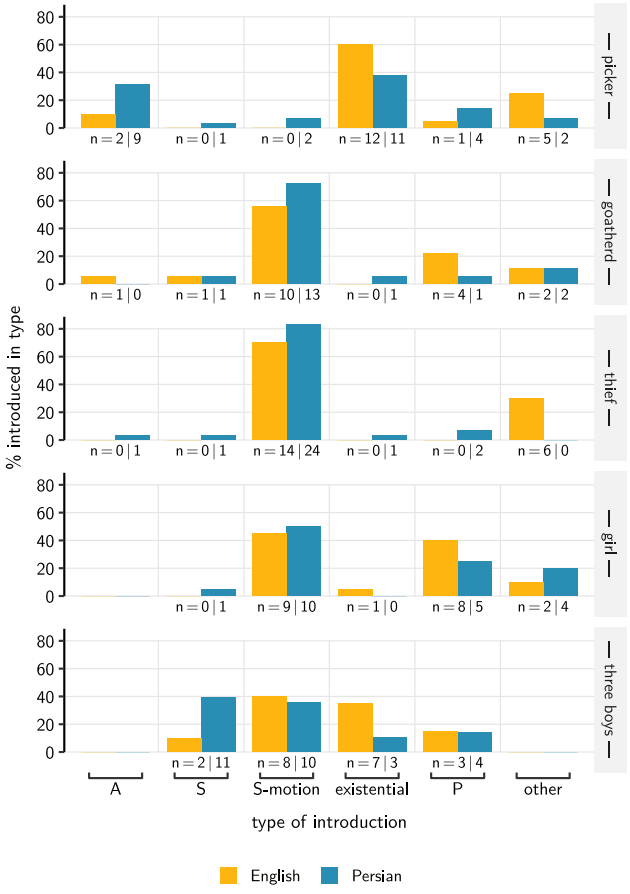

Figure 5: Proportions of different introduction strategies for the five human protagonists in 20 English and 29 Persian Pear Stories. Intransitive subjects are split into existentials, motion predicates ('S-motion'), and all other S arguments ('S'). The category labelled 'other' subsumes all introductions in non-core roles.

4 Conclusions

Overall, we do not find clear indication that syntactic argument structure is sensitive to newness as a specific referential feature, vis-à-vis referential choice and accessibility.[9] The only syntactic functions that show a consistently high proportion of new referents are direct objects (P) and various oblique arguments. The noted contrast between subjects and non-subjects stems mainly from the marked association of the former with continuing topics, and hence with given referents (cf. Chafe's 1994 light subject constraint), as speakers, above all, aim to observe coherence principles. The association of non-subjects with referent introductions is conversely quite open. However, we find that the range of non-subject argument functions hosting similar proportions of new referents is simply too broad for any specific construction to signal to the addressee to expect a new discourse referent. Rather, non-subject roles in general carry the potential of new information chiefly because subjects generally do not. Moreover, the lack of a distinct profile for P relative to other non-core

9 We do not, however, deny that newness can have various effects on morphosyntax. The NP-internal definiteness marking found in many languages has already been mentioned. Moreover, newness has been found to bear on constituent order within individual clauses, so that typically new information follows given information. In languages with overall fixed word order, this may involve specific constructional variants, like inversions in English (Arnold et al. 2000; Birner 2016; Birner and Ward 1998). An analysis of the relative order of constituents representing new versus given information within clauses in our corpora is a topic for future research.

syntactic functions casts serious doubt on any particular association of newness with any of the core grammatical relations S, A, or P, and hence on the claimed connection to ergative morphosyntax (Du Bois 2017).

As such, counter to the predictions of PAS (Du Bois, 1987a), speakers do not preferentially opt for the S role for the introduction of referents, especially not for human ones, and our findings from the two Pear story corpora strongly suggest that the motivations for introductions in S are either semantic (e.g. motion) or restricted to very local discourse contexts, such as introductions of characters at the outset of a narrative or a major scene transition. Conversely, circumstances that could be regarded as increasing the processing difficulty of introductions (e.g. high local information pressure; a paucity of contextual clues) appear to have little to no effect.

Similarly, the more common choice of P arguments for introductions seems to be primarily motivated by semantic and general conceptual-psychological considerations. Human agents typically encounter or interact with an ever expanding array of new inanimate objects in their environment, which is reflected by their introduction into narratives. Hence, in the P role, new referents are anchored to a state-of-affairs with another, already established referent. In some cases, this reflects a naturalistic introduction pattern through perception (*see*, *hear*) or a spatial encounter (*come across*). Anchoring new referents in such a way could be considered preferable for discourse processing when compared to isolating new referents from the rest of the discourse into the position of intransitive S arguments, contrary to what has been suggested under the perceived view on referent introductions.

In sum, our findings suggest considerations of recipient design and efficiency are much less relevant in accounting for how speakers accommodate new referents than has previously been assumed. Speakers seem to debut new referents in the argument role that best suits their participatory role in the event they are first mentioned in, and addressees are fully capable of registering them as new discourse entities irrespective of other considerations. That recipient design should play a rather limited role in discourse processing has also been found by psycholinguistic research in this area: as regards referent processing, Arnold (2008); Arnold et al. (2003) find that speakers are unlikely to actively attempt facilitation of addressees' processing efforts, but rather prefer to focus on issues of planning. This latter aspect, however, may result in an implicit learning effect for newness, as argued by Arnold (2008), Arnold et al. (2003), and MacDonald (2013).

In this view, efficiency in the introduction of new referents is achieved not by the partitioning of information management from content advancement, but by seamlessly integrating the former into the latter. Particularly salient introductory sequences in coherent chunks of discourse may of course be signalled by specific constructions, but our findings, based on the full range of new introductions in connected discourse samples, suggests that most of the work involved in introducuing new referents is accommodated within semantically appropriate predications, with little evidence for specialization.

Acknowledgment: The research reported here was made possible through the following grants: a DFG grant (Sachbeihilfe, DFG project no. 323627599), Schnell's post-doctoral position at the University of Melbourne within the Australian Research Council's Centre of Excellence for the Dynamics of Language (CE140100041), and Schnell's Australian Research Council DECRA grant (DE120102017). For acknowledgments relating to individual corpus compilation and annotation projects, please refer to the MultiCAST website.[10] We thank the audiences at the CoEDL Seminar, Australian National University, 11 November 2018, and the *Workshop Comparative Corpus Linguistics: New Perspectives and Applications*, Tallinn, 31 August to 1 September 2018, as well as one anonymous reviewer and the two editors of this special issue, Natalia Levshina and Steve Moran, for their constructive feedback. Their advice has greatly improved our work on the questions examined here. All remaining errors are our own responsibility.

10 multicast.aspra.uni-bamberg.de/.

Research funding: The research reported here was made possible through the following grants: a DFG grant (Sachbeihilfe, DFG project no. 323627599), Schnell's post-doctoral position at the University of Melbourne within the Australian Research Council's Centre of Excellence for the Dynamics of Language (CE140100041), and Schnell's Australian Research Council DECRA grant (DE120102017).

References

Abbott, Barbara. 1993. A pragmatic account of the definiteness effect in existential sentences. *Journal of Pragmatics* 19(1). 39–55.
Abbott, Barbara. 1997. Definiteness and existentials. *Language* 73(1). 103–108.
Adibifar, Shirin. 2016. Multi-CAST Persian. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#persian.
Andrews, Avery. 2007. The major functions of the noun phrase. In Timothy Shopen (ed.), *Language typology and syntactic description*, vol. 1: Clause structure, 132–223. Cambridge: Cambridge University Press.
Ariel, Mira. 2014. *Accessing noun-phrase antecedents*. New York: Routledge.
Arnold, Jennifer E. 2008. Reference production. *Language & Cognitive Processes* 23(4). 495–527.
Arnold, Jennifer E. 2010. How speakers refer. *Language & Linguistics Compass* 4(4). 187–203.
Arnold, Jennifer E., Maria Fagnano & Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research* 32(1). 25–36.
Arnold, Jennifer E., Anthony Losongco, Thomas Wasow & Ryan Ginstrom. 2000. Heaviness versus newness. *Language* 76(1). 28–55.
Birner, Betty J. 2016. English inversions as constructional alloforms. *Proceedings of the Linguistic Society of America* 1(19). 1–9.
Birner, Betty J. & Gregory Ward. 1998. *Information status and noncanonical word order in English*. Amsterdam: John Benjamins.
Blokpoel, Mark, Marlieke van Kesteren, Arjen Stolk, Pim Haselager, Ivan Toni & Iris van Rooij. 2012. Recipient design in human communication. *Frontiers in Human Neuroscience* 6. https://doi.org/10.3389/fnhum.2012.00253.
Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 25–55. New York: Academic Press.
Chafe, Wallace (ed.). 1980. *The pear stories*. Norwood, NJ: Ablex.
Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*, 21–51. Amsterdam: John Benjamins.
Chafe, Wallace. 1994. *Discourse, consciousness, and time*. Chicago: The University of Chicago Press.
Dixon, Robert M. W. 1987. Studies in ergativity. *Lingua* 71(1). 1–16.
Dixon, Robert M. W. 1994. Adjectives. In *The encyclopedia of language and linguistics*, vol. 1, 28–35. Oxford: Pergamon Press.
Du Bois, John. 1987a. Absolutive zero. *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. In John Du Bois, Lorraine Kumpf & William J. Ashby (eds.), *Preferred argument structure*, 11–60. Amsterdam: John Benjamins.
Du Bois, John. 2003b. Discourse and grammar. In Michael Tomasello (ed.), *The new psychology of language*, vol. 2, 47–88. Mahwah, NJ: Erlbaum.
Du Bois, John. 2017. Ergativity in discourse and grammar. In Jessica Coon, Diane Massam & Lisa D. Travis (eds.), *The Oxford handbook of ergativity*, 23–57. Oxford: Oxford University Press.
Durie, Mark. 2003. New light on information pressure. In John Du Bois, Lorraine Kumpf & William J. Ashby (eds.), *Preferred argument structure*, 159–196. Amsterdam: John Benjamins.
Foley, William A. 2007. A typology of information packaging in the clause. In Timothy Shopen (ed.), *Language typology and syntactic description*, 362–446. Cambridge: Cambridge University Press.
Forker, Diana & Nils N. Schiborr. 2019. Multi-CAST Sanzhi Dargwa. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#sanzhi.
Givón, Talmy (ed.). 1983. *Topic continuity in discourse* (Typological studies in language 3). Amsterdam: John Benjamins.
Gordon, Peter C. & Davina Chan. 1994. Pronouns, passives, and discourse coherence. *Journal of Memory & Language* 34(2). 216–231.
Hadjidas, Harris & Maria C. Vollmer. 2015. Multi-CAST Cypriot Greek. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#cypgreek.
Haig, Geoffrey & Stefan Schnell. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse)*. Available at: https://multicast.aspra.uni-bamberg.de/#annotations.
Haig, Geoffrey & Stefan Schnell (eds.). 2015. *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/.
Haig, Geoffrey & Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618.
Haig, Geoffrey, Maria C. Vollmer & Hanna Thiele. 2019. Multi-CAST Northern Kurdish. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#nkurd.

Halliday, Michael Alexander K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
Holler, Anke & Katja Suckow. 2016. Introduction. In Anke Holler & Katja Suckow (eds.), *Empirical perspectives on anaphora resolution*, 1–10. Berlin: de Gruyter.
Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar*. Stanford: CSLI Publications.
Kehler, Andrew. 2004. Discourse coherence. In Laurence Horn & Gregory Ward (eds.), *Handbook of pragmatics*, 241–265. Malden, MA: Blackwell.
Kwon, Nayoung, Yoonhyoung Lee, Peter C. Gordon, Robert Kluender & Maria Polinsky. 2010. Cognitive and linguistic factors affecting subject/object asymmetry. *Language* 86(3). 546–582. https://www.jstor.org/stable/40961691.
Lambrecht, Knud. 1994. *Information structure and sentence form*. 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.
McDaniel, Dana, Cecile McKee, Wayne Cowart & Merrill F. Garret. 2015. The role of the language production system in shaping grammars. *Language* 91(2). 415–441.
Meng, Chenxi. 2019. Multi-CAST Tulil. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#tulil.
Mosel, Ulrike & Stefan Schnell. 2015. Multi-CAST Teop. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#teop.
Nakagawa, Shinichi & Holger Schielzeth. 2012. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology & Evolution* 4(2). 133–142.
Polanyi, Livia. 1995. *The linguistics structure of discourse*. Stanford: CSLI Publications.
Polanyi, Livia & Remko J. H. Scha. 1988. An augmented context free grammar of discourse. In *Proceedings of COLING-88*. Budapest, Hungary: Association for Computational Linguistics.
Polanyi, Livia, Martin van den Berg & David D. Ahn. 2003. Discourse structure and sentential information structure. *Journal of Logic, Language & Information* 12(3). 337–350.
Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical pragmatics*, 223–255. New York: Academic Press.
Prince, Ellen F. 1998. On the limits of syntax, with reference to left-dislocation and topicalization. In Peter W. Culicover & Louise McNally (eds.), *Syntax and semantics 29*, 281–302. San Diego: Academic Press.
Sacks, Harvey, Emanuel A. Schegloff & Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4). 696–735.
Schiborr, Nils N. 2015. Multi-CAST English. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#english.
Schiborr, Nils N. 2018. multicastR. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://cran.r-project.org/package=multicastR.
Schiborr, Nils N., Stefan Schnell & Hanna Thiele. 2018. RefIND — Referent Indexing in Natural-language Discourse. Bamberg/Melbourne: University of Bamberg. Available at: https://multicast.aspra.uni-bamberg.de/#annotations.
Schnell, Stefan. 2015. Multi-CAST Vera'a. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#veraa.
Schnell, Stefan & Nils N. Schiborr. 2018. Corpus-based typological research in discourse and grammar. *Asian and African Languages & Linguistics* 12. 1–16. Available at: http://hdl.handle.net/10108/91145.
Thieberger, Nick & Timothy Brickell. 2019. Multi-CAST Nafsan. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#nafsan.
Vollmer, Maria. 2020. Multi-CAST Mandarin. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*. Available at: https://multicast.aspra.uni-bamberg.de/#mandarin.
Ward, Gregory & Betty J. Birner. 2004. Information structure and non-canonical syntax. In Laurence Horn & Gregory Ward (eds.), *Handbook of pragmatics*, 153–174. Malden, MA: Blackwell.

Supplementary material: The online version of this article offers supplementary material (https://doi.org/10.1515/lingvan-2019-0064).