Language Documentation & Conservation Special Publication 25 141
Doing corpus-based typology with spoken language data: State of the art
edited by Geoffrey Haig, Stefan Schnell, and Frank Seifart, pp. 141 – 177
https://nflrc.hawaii.edu/ldc/sp25
https://hdl.handle.net/10125/74660

Universals of reference in discourse and grammar: Evidence from the Multi-CAST collection of spoken corpora

Geoffrey Haig,[1] Stefan Schnell,[2] Nils Norman Schiborr [1]

[1] *University of Bamberg,* [2] *University of Zurich*

Abstract

Data from under-researched languages are now available in sufficient quantity and quality to feed into corpus-based approaches to language typology. In this paper we present Multi-CAST (*Multilingual Corpus of Annotated Spoken Texts*), a project designed to facilitate cross-linguistic comparison of naturalistic discourse across typologically diverse languages, which implements a purpose-built shared annotation scheme. After sketching the rationale and architecture of Multi-CAST, we illustrate the efficacy of the method with two case-studies: The first one investigates the rates of lexical (as opposed to pronominal and zero) realization of arguments in discourse across a sample of 15 typologically diverse languages. Our results reveal a remarkable and hitherto unnoticed uniformity in the density of lexical references, despite the lack of content control in the corpora. The second addresses the question of whether cross-linguistically attested regularities in morphosyntax can meaningfully be related to frequency effects in discourse. We find some support for frequency-based explanations, but our data also show that

the frequency accounts leave several key questions unanswered. Overall, our findings underscore that research based on language documentation-derived corpus data, and in particular spoken language data, is not only possible, but in fact crucially necessary for testing frequency-based explanations, because these data stem from spoken language and typologically diverse languages. We also identify a number of epistemological and methodological shortcomings with our approach, and discuss some of the requirements for further innovation in areas of corpus building, corpus annotation, and typological comparability.

Keywords: corpus-based typology, universals of language use, discourse structure, referential choice, marking asymmetries

1 Introduction

Since the 1970s, a number of researchers have focused on the interface of discourse and grammar, exploring how grammatical elements of language systems are used to produce coherent discourse, and how in turn language usage shapes (and has shaped) the diachronic development of morphosyntax. A major focus has been on universal aspects of discourse, which, with varying degrees of success, have been related to considerations of communicative functionality and language processing. Systematic differences in linguistic behaviour between speech communities and cultures have been of lesser importance here than in other linguistic disciplines like anthropological linguistics and the ethnography of speaking (Hymes 1961, 1962), which likewise aim to identify commonalities and differences in patterns of language use across speech communities.

Our own research is guided by the conviction that tackling the significant questions of discourse and grammar, and their connections with language processing and language structure, requires data from spoken (or signed) discourse from typologically diverse languages (cf. Schnell et al., this volume). Data of this nature are generally hard to come by, and their processing is

extremely labour-intensive.[1] The situation has been improving since the advent of modern language documentation after Himmelmann (1998), and usage data in unprecedented amounts and quality are now available in digital archives, but their potential has only recently begun to be exploited by linguists working within a typological research framework (see Schnell & Schiborr, in press; Schnell et al., this volume; and below for exemplification).

In this paper, we introduce the rationale and design of a corpus development project Multi-CAST ('*Multilingual Corpus of Annotated Spoken Texts*', Haig & Schnell 2021[2015]),[2] designed to address foundational issues within linguistics, and in typology in particular (Section 2). In Section 3 we present case-studies based on Multi-CAST addressing two topics in typology: First, referential density, or more generally, the cross-linguistic commonalities and differences in choices for referential expressions (pronouns, zero, or full noun phrases) in Section 3.1; and second, split marking of core arguments (e.g. differential argument marking) in Section 3.2. We show that even with the current modest sample size of Multi-CAST (15 language corpora with at least 1 000 clause units per language; see Figure 1 and Table 1 below), it is possible to make significant advances over existing research in both topics, and to identify hitherto unnoticed generalizations on the cross-linguistic uniformity of discourse. In Section 4, we summarize the main findings and consider the broader relevance of Multi-CAST within the context of typologically-informed research on grammar and discourse.

1 An illustrative example are the comparatively small proportions of spoken language in the large corpora of English, such as the COCA (Davies 2008) or the BNC (BNC Consortium 2008). The ICE corpora (Hundt et al. 2016) are a notable exception, but in turn draw on substantial labour forces unattainable for smaller-scale projects.

2 https://multicast.aspra.uni-bamberg.de/

2 The Multi-CAST initiative

Multi-CAST has been developed over the past decade with the primary aim of establishing a fully accessible database of spoken language use from diverse languages in order to promote cross-linguistic research in the field of discourse and grammar. This takes up a research programme established in the 1970s and 1980s (e.g. Chafe 1976; Givón 1979, 1983; Prince 1981; Du Bois 1985; among many others) that explored the impact of regularities in discourse on the shape and development of grammar(s). Like this earlier research, our interests are explicitly cross-linguistic in scope, and target universal patterns of discourse organization. However, our initiative also seeks to raise the bar in terms of data accessibility and accountability, breadth of typological coverage, and statistical methodologies. The next section first outlines main issues of corpus design, then turns to our annotation conventions; for a more comprehensive description of the collection and its design see Schiborr (2018).

2.1 Multi-CAST corpus design

Multi-CAST draws on data collected during various language documentation and fieldwork projects on diverse languages. All data are freely available online under a Creative Commons licence (CC-BY 4.0). The current version (from August 2021) contains data from 15 languages, see Figure 1 and Table 1. Further languages are continuously added to the collection as they become available, so that Multi-CAST is progressively expanded on a rolling basis. The selection of languages is primarily opportunistic, and is dependent on the collaboration of fellow linguists who are experts on individual languages and willing to engage in the arduous process of data processing and annotation. At this stage, the sample is still areally biased towards Western Asia and the Pacific region, reflecting the respective geographic foci of the two editors. Each individual language subcorpus has a minimal size of 1000 clause units, and comprises a number of different texts (between three and twelve), with the size of individual texts varying across the sub-corpora.[3]

3 For related corpus-building enterprises that focus on specific language families and areas, see the INEL project (‘*Grammatical Descriptions, Corpora and Language Technology for In-*

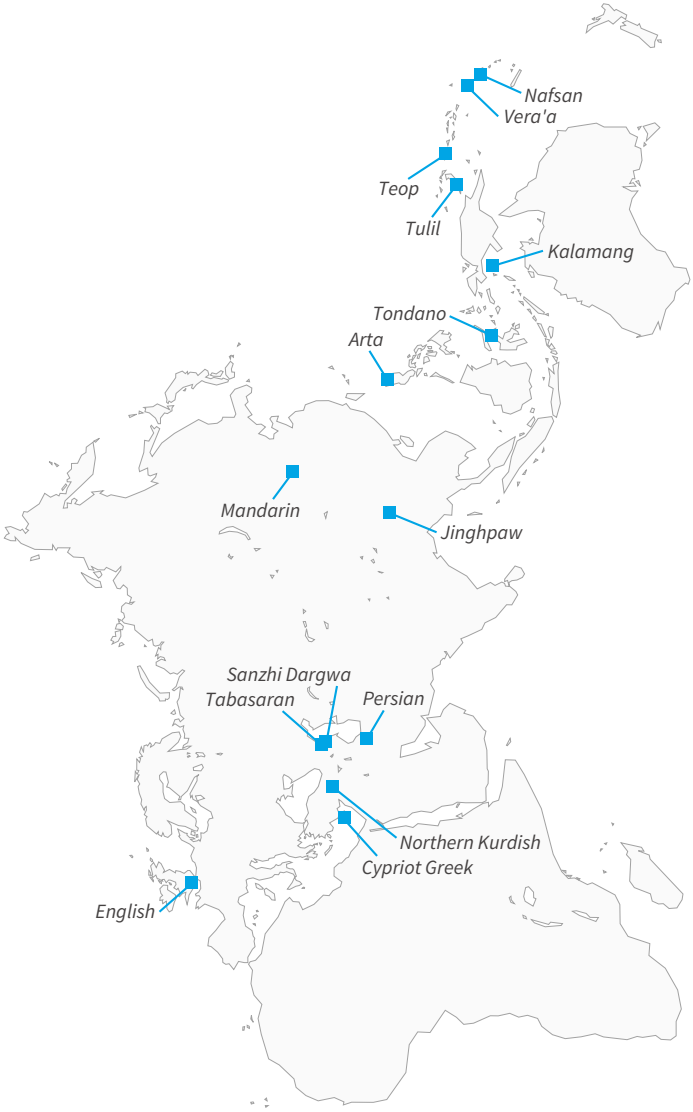

Figure 1 The Multi-CAST languages (as of version 2108).

corpus	affiliation	citation	texts	clause units
Arta	Austronesian, Malayo-Polynesian	Kimoto 2019	11	1030
Cypriot Greek	Indo-European, Greek	Hadjidas & Vollmer 2015	3	1070
English	Indo-European, Germanic	Schiborr 2015	5	5649
Jinghpaw	Sino-Tibetan, Tibeto-Burman	Kurabe 2021	11	1278
Kalamang	Papuan, West Bomberai	Visser 2021	6	1051
Mandarin	Sino-Tibetan, Sinitic	Vollmer 2020	3	1194
Nafsan	Austronesian, Oceanic	Thieberger & Brickell 2019	9	1012
Northern Kurdish	Indo-European, Iranian	Haig et al. 2019	3	1841
Persian	Indo-European, Iranian	Adibifar 2016	29	1418
Sanzhi Dargwa	Nakh-Daghestanian, Dargwic	Forker & Schiborr 2019	8	1066
Tabasaran	Nakh-Daghestanian, Lezgic	Bogomolova et al. 2021	5	1383
Teop	Austronesian, Oceanic	Mosel & Schnell 2015	4	1303
Tondano	Austronesian, Malayo-Polynesian	Brickell 2016	8	1085
Tulil	Papuan, Taulil-Butam	Meng 2019	6	1264
Vera'a	Austronesian, Oceanic	Schnell 2015	10	3608
totals			121	25252

Table 1 Overview of the Multi-CAST corpora (as of version 2108).

With its comparatively small number of languages, areal bias, and relatively short texts, Multi-CAST can hardly claim to be representative of global language use. Moreover, the texts are not controlled for content or structure, so that the various subcorpora are not comparable in a "parallax" sense (Barth & Evans 2017; Barth et al., this volume); Multi-CAST is in this sense comparable to Universal Dependency (UD) corpora (Zeman et al. 2021), which mostly contain texts not collected for the specific purposes of cross-linguistic comparative research. Finally, our commitment to spoken language data imposes restrictions on corpus size due to the immensely labour-intensive process of manual transcription and annotation. This constitutes a major difference to Universal Dependencies and related research, which mainly draw on existing pre-digitalized written language (e.g. Futrell et al. 2015, 2020; Levshina 2019). Although many UD corpora are much larger both in terms of individual corpus sizes and number of languages included, these approaches nevertheless suffer from a bias towards Eurasia (as noted in Levshina 2019), and reliance on generally standardized written varieties with official status as national languages, akin to Dahl's (2015) "LOL" languages ("Literate, Official, with Lots of users"). Our aim is thus to counteract the existing bias in corpus-based typology towards written forms of well-researched and generally standardized languages.

Corpus-based typology (CBT) involves statistical comparisons across usage samples from different languages. This in turn implies that we have identified the relevant quantifiable units in a uniform manner across different languages, that is, that we are comparing items that are both conceptually and practicably "comparable" across languages. At this point the comparability problem raises its head, as it does in any approach to typology; see Evans (2020) and other contributions to the recent special issue on comparability in the journal *Linguistic Typology* (2020, 24:3). For corpus-based typology, these issues need to be addressed at the annotation stage, and inevitably reflect the state of a particular research tradition at a particular point in time. Our solu-

digenous Northern Eurasian Languages'; `https://inel.corpora.uni-hamburg.de/`), the CorpAfroAs ('*Corpus of Spoken AfroAsiatic Languages*', Mettouchi et al. 2015; Mettouchi & Vanhove, this volume), as well as the CorTypo project (`https://cortypo.huma-num.fr/`; Mettouchi & Vanhove, this volume).

tions were initially guided by our research focus on the interface of grammar and discourse, and we hence needed to develop annotation schemes that capture certain features of morphosyntax and discourse structure in a manner that would be applicable to a maximally diverse set of languages; we introduce the basic scheme in the next section.

2.2 Corpus annotation with GRAID and RefIND

The corpora in Multi-CAST include transcriptions, conventional morpheme-for-morpheme glossings, and an idiomatic English translation, all of which are supplied by the respective language experts. In addition to these standard annotation tiers, Multi-CAST adds an overlay of (currently) two additional types of annotation, namely GRAID ('Grammatical Relations and Animacy in Discourse', Haig & Schnell 2014) and RefIND ('Referent Indexing in Natural-language Discourse', Schiborr et al. 2018). GRAID and RefIND each provide a set of tags which are implemented manually, and which primarily target those referential expressions which introduce and track discourse referents (cf. Karttunen 1976). GRAID assigns to each referential expression a set of values for syntactic function, animacy, and person. In addition, the GRAID tier also identifies predicative items, clause boundaries, and certain kinds of clausal operators such as negation or subordination. RefIND, on the other hand, captures co-reference relations across the referring expressions that are tagged with GRAID. RefIND annotations are actually spread over two tiers, labelled here RefIND and ISNRef ('Information Status of New Referents'), respectively. ISNRef only identifies discourse-new (rather than given) referents, hence many of the cells in this tier remain empty. The following examples from English (1) and Vera'a (2) illustrate the annotations with GRAID, RefIND, and ISNRef.[4]

4 Morphological glossing follows the Leipzig Glossing Rules. Abbreviations: 1 – first person; 3 – third person; ABS – absolutive; ACC – accusative; ART – common article (in Vera'a); CM – conjugation marker; DAT – dative; DEF – definite; DU – dual; ERG – ergative; IRR – irrealis; LOC – locative preposition (in Vera'a); M – masculine; NOM – nominative; OBL – oblique case (in English); PRF – perfect; PST – past; SG – singular; TRANS – transitive.

(1) English

transcription		*I*	*went*	*in*	*the*	*stillroom*	*with*	*the*	*milk*
morph. gloss		1SG	go.PST	in	DEF	stillroom	with	DEF	milk
GRAID	##	pro.1:s	v:pred	adp	ln	np:g	adp	ln	np:obl
RefIND		0000				0092			0091
ISNRef									

transcription		*and*	*the*	*stillroom maid*		*give*	*me*	*a*	*shilling*
morph. gloss		and	DEF	stillroom_maid		give.PST	1SG.OBL	a	shilling
GRAID	##	other	ln	np.h:a		v:pred	pro.1:p	ln	np:p2
RefIND				0093			0000		0094
ISNRef				bridging					new

(mc_english_devon01_0040, Schiborr 2015)

(2) Vera'a

transcription		[zero]	*'ō*	*duru*	*lē*	*=n*	*Wērēsurō*
morph. gloss			take	3DU	LOC	=ART	W.
GRAID	##	0.d:a	v:pred	pro.h:p	adp	=ln	np:g
RefIND		0002		0008			0026
ISNRef							

transcription		[zero]	*le*	*mē*	*duru*	*=n*	*gengen*
morph. gloss			give	DAT	3DU	=ART	food
GRAID	##	0.d:a	v:pred	adp	pro.h:g	=ln	np:p
RefIND		0002			0008		0003
ISNRef							new

'(It) took them to Wērēsurō, (it) gave them food.'

(mc_veraa_mvbw_0089-0090, Schnell 2015)

As these examples demonstrate, many details of constituent and/or dependency structure or linear precedence are only coarsely indicated. For example, since GRAID targets phrasal constituents, the annotation expressions align with the lexical heads of these constituents, as exemplified by the annotation

⟨np:obl⟩ in the first line of (1).[5] Additional subconstituents of phrases are only minimally analyzed via symbols that indicate their phrase membership and position relative to the head, as with ⟨ln⟩ in (1), which indicates that this item 'belongs to a NP, and is to the left of the lexical head of that NP'.

The annotation practices implemented in Multi-CAST were developed in response to an early research focus on grammatical relations in discourse (Du Bois 2003), so for this reason we have incorporated into GRAID the concept of "core arguments" following Andrews (2007), distinguishing intransitive subject (S, annotated ⟨:s⟩), transitive subjects (A, ⟨:a⟩), and direct object (P, ⟨:p⟩). On Andrews's approach, A and P in a given language can be identified through reference to the specific morphosyntactic properties associated with the two arguments of a prototypical transitive verb in that language, such as *kill* or *smash.* Essentially, clauses exhibiting the same formal features that are identified for these reference verbs are also considered to include A and P. For example, the two core arguments of the English verb *see*, though it is not a prototypical transitive verb, are nevertheless identified as A and P respectively, since in terms of case marking and word order, they do not differ from the arguments of *kill*. S, conversely, is taken for the syntactically privileged argument of clauses that do not match the transitivity prototype. While this approach is not without its drawbacks, it has proved a viable compromise solution for the demands of a cross-linguistically applicable syntactic coding system.

Other functions recognized include ⟨:l⟩ for locatives and ⟨:g⟩ for goals, recipients, as well as addressees, in cases where these are oblique rather than core arguments; that is, where a P argument bears, for instance, the semantic role of recipient, it is nonetheless annotated as ⟨:p⟩. In addition to these basic function glosses, values for animacy (human vs. non-human) and person are included.[6] This system has evolved over the years in response to the chal-

5 We use the convention of encasing annotation symbols in angular brackets.

6 ⟨.1⟩ and ⟨.2⟩ for first and second person (assumed human), ⟨.h⟩ for human third person; non-human third person is not explicitly annotated. Additionally, ⟨.d⟩ may optionally be used for anthropomorphized third-person references as in (2), where the zero subject refers to the spirit of a reef.

lenges raised by different languages, but is now largely standardized.[7] Each individual corpus is further accompanied by what we call "annotation notes", a document that includes a list of all annotation symbols used in that corpus and details of language-specific coding decisions.

It is important to bear in mind that the GRAID annotations co-exist with the conventional morphological glossing, as shown in (2), so that finer categorical distinctions not captured by GRAID are still available in the data. But GRAID is not simply a translation of conventional glossing; it renders a level of grammatical relations that is not immediately recoverable from morpheme-for-morpheme glossing. For example, we systematically indicate the presence of a zero exponent of a referential expression, and, as mentioned above, tag the subject of a transitive verb (A) distinctly from that of an intransitive clause (S), something that most other transcription systems do not do. Crucially, neither can be read off the morphological glossing in sufficiently reliable ways.

The core of the GRAID annotation system is thus kept fairly general so as to enable immediate quantitative comparison with respect to the relevant categories. For example, it is a very straightforward matter to extract the distribution of pronouns (as opposed to, e.g., zero), across transitive clauses (as opposed to intransitive clauses). In this sense, a basic set of comparative concepts is built into corpus annotations, rather than leaving comparison to calculations of equivalence based on language-specific annotations (as in CorpAfroAs and CorTypo; cf. Mettouchi & Vanhove, this volume). Finer-grained language-specific distinctions can be added to the core set of GRAID annotation tags, for instance ⟨dem_pro⟩ for a pronominally used demonstrative, extending the core symbol ⟨pro⟩ for free definite pronouns, or ⟨:a_cv⟩ for a 'transitive subject' in a converb construction specific to some languages. For a concise overview of the main tags used in GRAID, see the *GRAID manual* (Haig & Schnell 2014: 54).

7 More details on the meaning and functions of the GRAID and RefIND annotations can be found in the annotation manuals (Haig & Schnell 2014; Schiborr et al. 2018). Furthermore, Schnell and Schiborr (2018) outline how GRAID and RefIND annotation together can be analyzed in the context of research on discourse and grammar.

The seemingly coarse and relatively simplistic annotations in GRAID derive their true power from the combination with information on other tiers. It is possible to construct complex queries that combine information on, for example, syntactic functions, referent indices (e.g. identifying all expressions with the same referent), animacy features, clause boundaries, and, if necessary, language-specific morphological tags from the morphological glossing. As an example of the latter, it is possible to design a query that captures all those referential expressions on the GRAID tier which align with an item carrying plural marking in the morphological glossing tier. In other words, even though GRAID itself does not reserve tags for distinguishing plural and singular, this feature could be recovered from the existing morphological glossing. Distance measures and word-order features can also be derived from the corpus, see for instance (Haig 2020) for an application to word-order variation, and (Schiborr 2021) to anaphoric distance. Another example is Schnell's (2018) study of demonstratives in Vera'a, which relies on GRAID annotations in combination with morphological glossing to determine adnominal and pronominal uses of demonstratives.

Corpus annotations are undertaken manually by a language expert in collaboration with the Multi-CAST team. Each corpus raises its own annotation challenges, which are resolved collectively and documented in the annotation notes that accompany each corpus. For referent indexing, manual annotation practices are well established in corpus linguistics (cf. e.g. Mitkov 2000; Garside 1993). Considerations of manual coding procedures also motivate our decision to keep referent annotation limited to simple identifiers, rather than annotating a detailed set of properties, as is done in Riester and Baumann's (2017) RefLex scheme.[8] As mentioned above, quite complex analyses are nevertheless possible through combining information across multiple tiers. In order to facilitate this, data sets are available in a table format

8 For example, with RefIND distances between referring expressions can be automatically calculated in terms of utterances, clause units, word forms, elapsed time, and so on, whereas RefLex requires annotators to hard-code distance values of a certain pre-defined type (such as clauses) into the annotations. This constrains analytical possibilities, since it limits relevant categories from the start. The minimal approach to annotation adopted in the RefIND tier is also conceptually motivated by the desire to avoid top-down imposition of controversial categories such as "topic"; see e.g. Ozerov (2018) for a discussion.

(in addition to ELAN and XML files) for cross-corpus analysis with statistical software; for R, we also provide a companion package called *multicastR* (Schiborr 2019) to simplify this process. We will not explain the details of these analyses here, but instead point out that our annotations yield a grid-like structure with explicit information on paradigmatic and syntagmatic relationships that enable a range of queries whose relevance will be obvious to most readers. For a more detailed description, see Schiborr (2018).

Finally, for the sake of general corpus linguistic interest, it is worth noting that various corpus-building projects in the past have developed morphosyntactic annotation schemes and have attempted to make these amenable to research on diverse languages, centering around part-of-speech tagging and treebanks. Despite certain overlaps with existing tagging schemes, GRAID takes a more typologically-oriented approach from square one, taking its cue from the typological tradition of combining both semantics and morphosyntax in defining relevant categories, rather than attempting to extend the applicability of tagsets originally developed for a single language (English), via other European languages (German, Greek, etc.) to other Indo-European languages (e.g. Urdu; Hardie 2003), and so forth (cf. the EAGLES standards on tagging, Leech et al. 1996).

3 Case studies in discourse and grammar based on Multi-CAST

It is increasingly recognized that many of the foundational research issues in traditional grammar-based typology are also amenable to corpus-based approaches (see Bresnan et al. 2001 for early applications, Cysouw 2009 and 2014 for a theoretical justification, and Levshina 2019 for recent work on written corpora). In this section we illustrate the utility of spoken language corpora for two central issues in mainstream typology: (i) referential density and anaphora (Section 3.1), and (ii) splits in the case-marking of core arguments (differential or split argument coding, Section 3.2). The first case-study builds on work by Schiborr (2021), while the second is an exploratory application of Multi-CAST that tests the predictions of efficiency-based accounts of argument splits (Haspelmath 2021b).

3.1 Investigating the density of lexical referential expressions cross-linguistically

Speakers of any language exercise a certain degree of freedom in their choice of referential expressions. Within typology, a broad distinction is generally drawn between three options: zero, pronouns, and lexical (or "nominal") NPs. Even when reduced to just these three options, the complexities behind such choices are formidable; the second sentence in example (3) provides a broad illustration of the kinds of choices available:

(3) a. *[The teacher]$_i$ asked [a student]$_j$ to hand in [his assignment]$_k$.*

b. *[The student/he/ø]$_j$ handed [the assignment/it/ø]$_k$ to [the teacher/her/ø]$_i$ and [the teacher/she/ø]$_i$ left the classroom.*

These choices are heavily constrained by language-specific preferences: Speakers of English, for example, would not permit zero in any of these slots (at least on the co-reference interpretations provided). Other languages are more tolerant of zero, and might avoid a pronoun for the object of *handed*, and so on. These cross-linguistic differences have given rise to a number of typological classifications, most famously Perlmutter's (1971) distinction between pro-drop and non-pro-drop languages, while more recent approaches consider languages in terms of higher or lower overall levels of overt referential expressions. Bickel (2003: 710), adopting terminology originally from McLuhan (1964), refers to "hot" and "cool" languages, with the latter being characterized by sparse information density. In a similar vein, Huang (2000: 261–277) refers to "pragmatic languages" such as Mandarin, which apparently require speakers to infer intended reference from contextual cues to a greater extent than, for example, English. A more nuanced approach is Bisang (2015), who develops the concept of "hidden complexity", an example of which is the radical pro-drop characteristic of many "East and mainland Southeast Asian (EMSEA) languages" (Bisang 2015: 180). Regardless of the terminology, much of this research has focused on the respective levels of pronouns and zero in discourse, and here indeed very considerable cross-language differences obtain (cf. Schiborr 2021: Fig. 5.1).

Conversely, the overall rates of lexical expressions (*the teacher*, *the assignment*, etc. above), have seldom been investigated cross-linguistically. Figure 2

shows the percentage of lexical NPs (vs. reduced forms, i.e. pronouns or zero) in various argument positions (subjects, objects, obliques), extracted from the 15 Multi-CAST corpora. The figures are for third person forms only, as it is only here that speakers may exercise a choice between a lexical and a reduced form. Despite the uncontrolled nature of the texts in terms of content and the relatively small size of the respective samples, these data cluster fairly consistently around a mean rate of about 44% (standard deviation SD = 4.6%). Interestingly, it is Mandarin that shows the highest rate of lexical expressions, thus being the "most explicit" corpus in terms of lexical reference, a finding that is not expected according to the aforementioned view of Mandarin discourse as characterized by low informational density and high levels of contextual inferencing by interlocutors (Huang 2000; re-assessed in Vollmer 2019).

The cross-linguistic uniformity of rates of lexical expression is a surprising finding, which runs counter to remarks in the literature that languages may exhibit drastic differences when it comes to the use of lexical NPs, what Stoll and Bickel (2009) term "lexical referential density".[9] While individual texts in some corpora may stray considerably from the 44% mean rate, this is most likely a consequence of the small absolute size of the texts concerned; the central tendency asserts itself when sufficient text data is examined.

Schiborr (2021) finds that the selection of lexical anaphors responds to very similar factors across languages, primarily those related to low discourse coherence (Givón 1983; Kehler 2002, 2004), but that there are also notable differences between languages as regards the impact of certain factors such as animacy. The pattern seen in Figure 2 is chiefly determined by the regularity of subject anaphors in particular, which are predominantly non-lexical (cross-corpus mean 30%, SD = 6.4%). This can be tied to the role of subject as a pragmatic pivot (in terms of Foley & Van Valin 1984) and hence as a vector of referential continuity. Subjects very commonly form anaphoric chains that repeat the reference of the previous clause's subject; in this highly topical con-

9 In fact, the difference in mean lexical referential density between the two corpora compared in Stoll and Bickel (2009), from Russian and Belhare, is only 15%, which falls well into the range we find in the Multi-CAST data. This underlines the importance of data breadth when evaluating cross-linguistic differences.

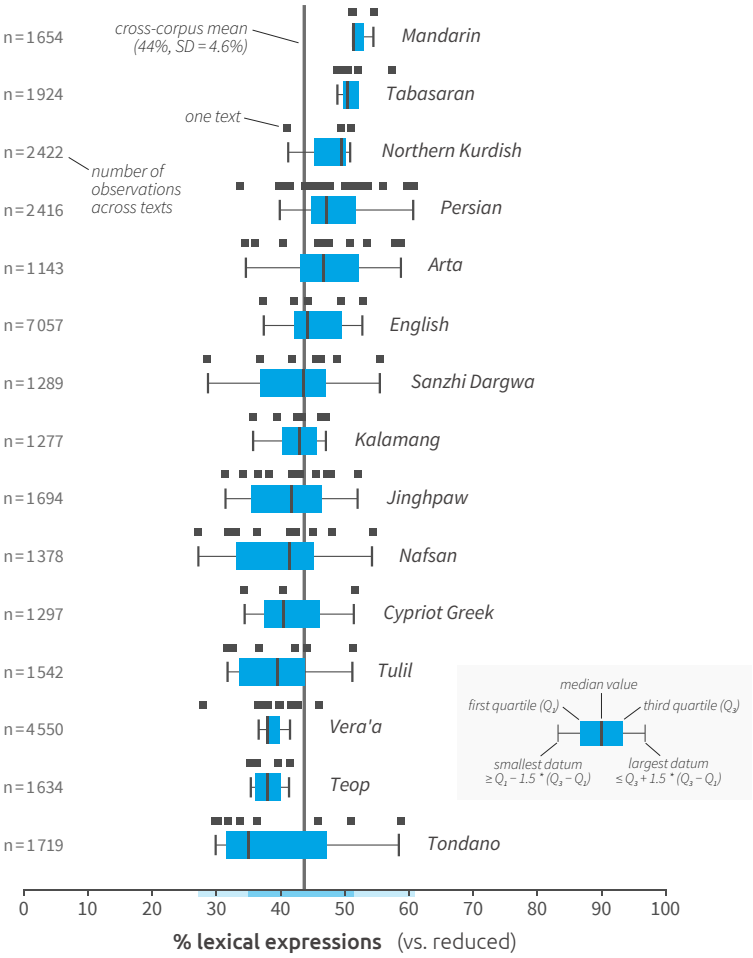

Figure 2 Proportions of full lexical NPs (vs. reduced forms) in argument positions (subject, object, obliques) across 15 Multi-CAST corpora (Haig & Schnell 2021, version 2108), third-person expressions only.

text, lexical expressions are virtually absent (mean 11% lexical, SD = 4.1%).[10] It is only where reference changes that full NPs become common (mean 38%, SD = 8.9%). Quite notably, the rate at which these topic shifts occur is remarkably stable across the corpora (accounting for a mean 43% of subject anaphors, SD = 8.2%), which Schiborr (2021: 397–409) explains in terms of an optimal information density of subject references: A discourse with too little differentiation among topics would not be worth telling (cf. Labov & Waletzky 1967; Engelhardt et al. 2006), whereas a discourse that switches topics too frequently would risk becoming incoherent. Of course, speakers do not mechanically switch topics every few clauses, and there may be local extrema in the rate of topic shifts, but over longer stretches of discourse, the rates – and consequently the proportion of lexical expressions – level out, regardless of language.

This strong cross-linguistic stability is largely limited to subject anaphors, however; when differentiating the overall picture in Figure 2 by syntactic functions (subjects, objects, etc.), greater cross-linguistic differences emerge (Schiborr 2021: 157). Lexicality rates for objects, for example, fluctuate considerably more across different languages than they do for subjects, a fact presumably related to the less consistent association of the object role with topicality (Schiborr 2021: 410). The cross-linguistic differences that obtain here have seldom been explored, and offer a fruitful field for future inquiry.

3.2 Referential properties of core arguments: Mapping frequency to typology

In this section we turn our attention to a central topic of traditional typology, namely marking asymmetries in argument encoding (so-called differential or split argument coding). It is widely acknowledged that these are co-conditioned by extra-syntactic factors, for example contextual factors such as givenness, or semantic factors including animacy, yielding precisely the kinds of variable patterns which invite corpus-based inquiry. In Section 3.2.1

10 These and the following figures exclude the data from the Arta, Persian, and Tondano corpora, for which the referent indexing with RefIND (see Section 2.2 above), which is required for the analysis of clause chains, are not yet available at the time of writing.

we summarize the traditional observations of differential argument encoding and accounts thereof, as well as a recent efficiency-based account proposed by Haspelmath (2021a; b). In Section 3.2.2 we present findings pertaining to this hypothesis from Multi-CAST before turning to a number of conclusions relevant to the role of corpus-based approaches to typology.

3.2.1 DOM and DAM: Haspelmath's efficiency-based account

Differential argument coding systems are widely attested across the world's languages; their global distribution, and the nature of the conditioning factors, are a focus of ongoing research (see Fauconnier & Verstraete 2014; Sinnemäki 2014; Haspelmath 2021b; among many others). Here we focus on the two most widely researched kinds of argument split, namely differential object marking (DOM) and differential A marking (DAM). DOM is the more familiar as well as the more widely attested split; the following example from Amharic (Afro-Asiatic/Ethio-Semitic, Ethiopia) illustrates a typical DOM system:

(4) Amharic (Ethio-Semitic, Afro-Asiatic)

a. *ləmma and t'ərmus səbbər-ə*
Lemma one bottle break.PRF-3M
'Lemma broke one bottle.'

b. *ləmma t'ərmus-u-n səbbər-ə*
Lemma bottle-DEF-ACC break.PRF-3M
'Lemma broke the bottle.' (Amberber 2008: 4)

The direct object in (4a) is indefinite, and unmarked for case, whereas the direct object in (4b) is definite, and consequently receives overt case marking. In this study we restrict ourselves to splits with one marked and one unmarked member, leaving aside splits with two marked members, for instance dative-accusative splits in DOM (e.g. in Punjabi).

According to Sinnemäki (2014), DOM is actually more frequently attested than systems with consistent marking of direct objects. Furthermore, DOM is relatively evenly distributed across the languages of the world (Sinnemäki 2014: 293). Differential A marking (DAM), on the other hand, is widely

considered less frequent (Fauconnier & Verstraete 2014: 5; Levshina 2021: 2), though we are unaware of a systematic comparison of global frequencies.[11] An example of DAM from Ngiyambaa, a language of Australia, is provided in (5):

(5) Ngiyambaa (Pama-Nyungan, New South Wales)

a. *miri-**gu**=na* *bura:y-ø* *gadhiyi*
dog-ERG=3ABS child-ABS bite.PST
'The dog bit the child.' (adapted from Donaldson 1980: 128)

b. *ŋadhu=na* *bura:y-ø* *yada* *bun-ma-l-aga*
1SG.NOM=3ABS child-ABS well change-TRANS-CM-IRR
'I will make the child well.' (adapted from Donaldson 1980: 220)

In (5a), the A argument is ergative marked, while in in (5b) the first person A is in the nominative case, which is formally unmarked (the clitic *=na* cross-references another argument; it is irrelevant for case marking of the A). The conditioning factor in Ngiyambaa is person: First and second person A arguments are in the unmarked nominative case, while third person arguments, including pronouns, take the marked ergative case.[12]

In accounting for the distribution and nature of argument splits in the world's languages, scholars have regularly appealed to universal cognitive and communicative constraints operative in language usage, which over time engender broadly similar grammatical systems across multiple languages

11 The outcome of such a survey would depend heavily on how one defines DAM. If we broaden the definition to include various kinds of differential subject marking (e.g. non-canonical marking of experiencers and possessors), it would be considerably more frequent cross-linguistically. However, these phenomena are heavily dependent on the lexical semantics of the predicate (typically predicates of cognition, perception, experience, and possession), rather than properties of the argument itself. We adopt a narrower view of DAM here, restricted to transitive verbs and triggered primarily by properties of the A argument – in effect to so-called split ergativity.

12 A reviewer points out that (5) exhibits similarities to a person-based inverse system. Although it is undeniably true that a similar set of factors is involved in both inverse and DOM/DAM, for the sake of brevity, we continue to focus on DOM and DAM. We return to the relevance of inverse systems in Section 3.2.3.

(Haspelmath 2021b).[13] At the core of these explanations lies the postulation of some notion of natural transitive event, which supposedly determines the properties of the typical, or "unmarked" A and P. As Comrie (1979: 19) put it more than 40 years ago, in a natural transitive event, "subjects [A] tend to be definite, animate, and topic (thematic); while direct objects [P] tend to be indefinite, inanimate, and rhematic." Where A and P diverge from these expectations, they are said to be "marked", and may (depending on the language) require additional phonological material. Those exemplars of A and P that comply with the expectations, on the other hand, have less overt marking. The result is a case marking asymmetry, DOM or DAM.[14]

Haspelmath (2021b), in line with a long tradition in functional linguistics, pursues this approach, though he eschews the notion of "markedness" (see below). Among the factors he considers relevant, we focus here on the following three, data for which can be readily extracted from the Multi-CAST corpora: newness (given vs. new in the sense of Chafe 1976), animacy, and person. In line with most recent research, Haspelmath (2021b) sets up hierarchies (termed "scales of referential prominence"), contrasting more and less prominent values for each feature, shown in (6):

(6)		high prominence		low prominence
	newness scale	discourse-given	>	discourse-new
	animacy scale	human	>	non-human
	person scale	first/second person	>	third person

13 Bickel et al. (2015), however, conclude that there is little evidence for universal factors driving various kinds of argument split; but see Schmidtke-Bode and Levshina (2018) for a reassessment of the data that comes closer to the traditional view presented here. For an acquisitional perspective on DOM, see Mardale and Montrul (2020).

14 A reviewer points out that there are other strategies for handling atypical A arguments, for example through demotion of an indefinite or inanimate A in a passive construction. This is a valid point, but the existence of alternative strategies would not impinge on the way differential case marking is distributed over A and P in active transitive clauses. It could, however, affect the relative frequencies of such atypical core arguments, because passivized clauses are generally considered intransitive, and would therefore lie outside of the frequency data presented in Section 3.2.2 below. This is a potentially important confounding factor, though we should note that with the exception of English, none of the languages in our sample make very widespread use of passivization.

According to Haspelmath, Comrie's vague notion of "tend to be" and the appeal to markedness regularly made in the literature should be replaced by the more objective measure of corpus frequency: "[...] [W]henever I say, for example, that 'A shows a greater tendency to be definite than P', I mean that we find more definite A-arguments than definite P-arguments in all representative texts in all languages" (Haspelmath 2021b: 126). The focus on frequency permits a reformulation of Comrie's insight as (7), which is based on Haspelmath (2021b: 129):

(7) An A argument is more frequently associated with the high-prominence values, while a P argument is more frequently associated with the low-prominence values in (6).

With regard to the actual marking (or flagging) of A and P, it will be recalled that DOM and DAM generally involve a variation between a heavier marked variant and a lighter (or phonologically unmarked) variant. And for both DOM and DAM, it is the heavier variant that is associated with the less frequently attested prominence values. For example, in Amharic, the phonologically heavier variant of the direct object (P) is the definite one in (4b), which carries an additional suffix. According to (7), we expect a P argument to have a low-prominence value with regard to newness, hence be discourse-new. When it is discourse-given (definite), it violates this expectation, and a phonologically heavier form is predicted. In the Ngiyambaa example for DAM, the heavier variant is the third person A in (5a), a low-prominence variant, again violating the predictions of (7).

The association of heavier form with less frequent variant in both DAM and DOM can be interpreted as confirmation of a more general principle, according to which less frequent, and hence less expectable, variants are coded more heavily, for example singular versus plural marking of nouns and many other types of coding asymmetries found in grammar. According to Haspelmath (2021a: 1), coding asymmetries of this type reflect "[...] a cross-linguistic pattern in which the less frequent member of the opposition gets special coding, unless the coding is uniformly explicit or uniformly zero." From an efficiency angle, heavier forms require greater production effort, so considerations of economy predict the heavier form to occur in the least fre-

quent context. The lighter variant, on the other hand, occurs in the most frequent context types with more predictable values. Frequency, so the general idea, primes speakers' expectations, so that highly frequent forms are more routinely and efficiently processed, hence requiring less overt phonological marking.[15] The appeal of this line of explanation is that DOM and DAM emerge as simply sub-cases of a more general principle, dubbed by Haspelmath (2021b: 125) "Universal 1":

(8) **The role–reference association universal:**
Deviations from usual associations of role rank and referential prominence tend to be coded by longer grammatical forms if the coding is asymmetric.

(8) suggests that frequency asymmetries in usage are reflected in the architecture of grammars cross-linguistically: The emergence of longer forms for less frequent variants. If one takes this approach seriously, then it follows that the larger the relative frequency asymmetry between related variants, the more likely it is to be reflected in asymmetrical coding.

Turning to DAM and DOM, it was noted above that they are not distributed evenly across the languages of the world: DOM is much more widespread, though exact figures are not available for reasons discussed above. Furthermore, DOM and DAM are not sensitive to the same set of factors; DOM is most commonly associated with definiteness/newness and animacy, whereas DAM is conditioned by focus, not definiteness/newness (Fauconnier & Verstraete 2014). Person (first/second vs. third) is widely attested as conditioning the split marking of A in ergative languages of Australia – see (5) above – but not newness (Coon & Preminger 2017: 245). Of course the feature of person is intertwined with newness/definiteness (first and second person arguments are often just considered the extreme pole of definiteness), but

15 We are summarizing Haspelmath's line of argument here, which assumes that frequency is the ultimate cause. However, alternative information-theoretical approaches also merit consideration, particularly those that consider some measure of informativity as the primary factor in shaping processes of reduction in language; see Cohen Priva (2017: 576–578) for discussion related to phonology, Piantadosi et al. (2011), and the review article Gibson et al. (2019) for alternative approaches to frequency and efficiency, and on ambiguity as an efficiency property of human languages.

here we take definiteness to imply some differentiation among different types of full NP arguments on the basis of discourse recoverability. Pure person-based marking asymmetries are nonetheless attested for both DOM and DAM (e.g. in Taleshi, Indo-European/Northwest Iranian, Iran; Haig 2017: 495).

If we take the view that coding asymmetries in grammar are related to frequency differences in usage, then the differences between DAM and DOM sketched in the preceding paragraph would presumably be reflected in some kind of frequency differences in usage in the relevant forms. For example, we might assume the predominance of DOM over DAM reflects a greater degrees of frequency asymmetry in the relevant forms of O than of A, and so on. In the following section, we investigate some of these questions using the spoken language data from Multi-CAST. Recently, Levshina (2021) has undertaken a similar investigation on five spoken language corpora, which we will refer to at various points in the discussion, though we are unable to do it full justice here. Note that Levshina's corpus is considerably smaller than the one currently used (a total of 975 transitive clauses across five corpora, mean size of 195 per corpus), while the present investigation considers around 9000 transitive clauses in 15 corpora.

3.2.2 Frequency data from Multi-CAST: Animacy, newness, and person

We structure our investigation as follows. First we provide data from Multi-CAST illustrating the frequencies for A and P across the three prominence scales set out in (5): newness, animacy, and person. We then discuss the data from each of three prominence scales, addressing the question of whether the (considerable) differences between them can be related to the way DOM and DAM systems are distributed globally. Section 3.2.3 summarizes our main findings and considers more general questions of methodology and the limitations of purely frequency-based approaches.

The primary findings from Multi-CAST are summed up Figure 3. Each point represents a corpus from a particular language in the sample, and the y-axis provides the respective percentages of the high-prominence value among all referential arguments in the respective roles. The horizontal lines show the baseline rates for the high-prominence value across all NPs in the corpora, regardless of role (i.e. A, P, and all other referential expressions in-

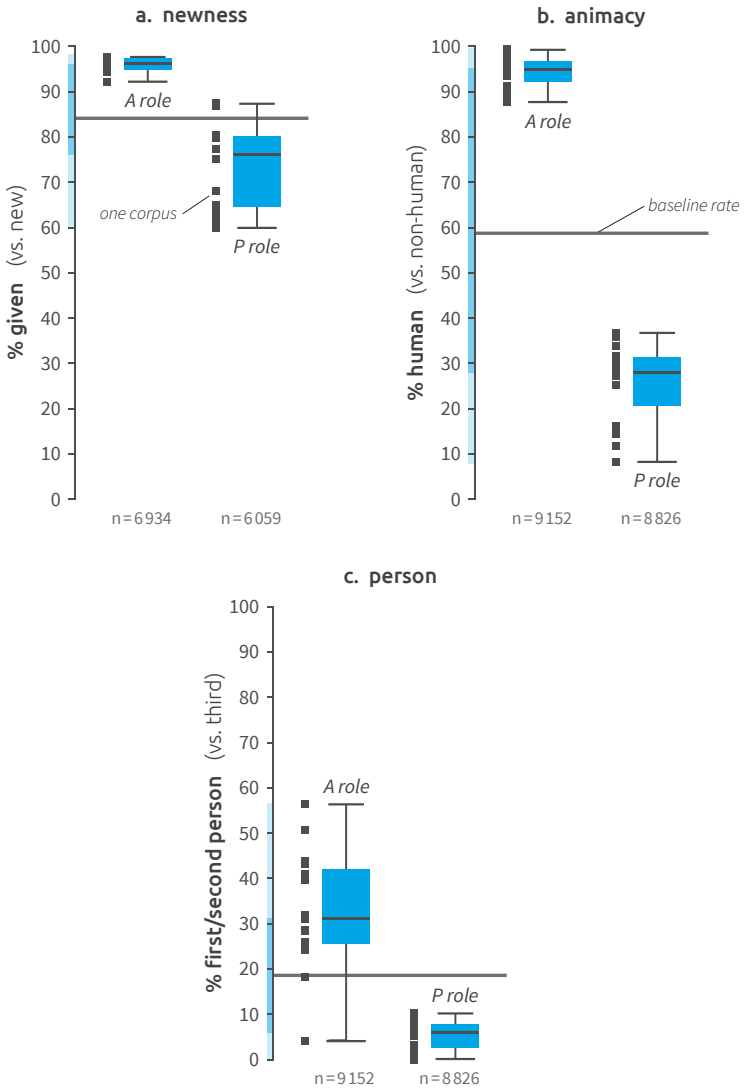

Figure 3 Proportion of high-prominence values for A and P arguments in 15 (twelve for newness) Multi-CAST corpora (Haig & Schnell 2021, version 2108).

cluding non-core arguments).[16] In other words, it tells us that for the feature of newness, the mean high-prominence value (i.e. 'given') for all NPs in all corpora is around 85%. This is an important metric, which is generally not provided in the literature; see the discussion in Section 3.2.3. The data for animacy and person include 15 languages (i.e. all those listed in Table 1) while the data for newness excludes three corpora (Arta, Persian, and Tondano) for which no referent indexing is available yet, which we employ to determine newness.

With regard to the degree of cross-language variability, the newness and animacy (i.e. humanness) values for A and P across corpora are approximately normally distributed, with a narrow range for A and a comparatively greater dispersion for P. Overall, the high-prominence values for A are consistently higher than for P across all three scales, which does provide overall support for the prominence asymmetry predicted in (7).

It is nevertheless evident that there are considerable differences in the patterns that emerge for our three factors. The obvious question is: To what extent do the attested frequency differences square up with what is known about the distribution of DOM and DAM cross-linguistically? As noted in Section 3.2.1, it is generally acknowledged that DOM is more frequently attested than DAM. From the frequency perspective, this would suggest that the greatest frequency asymmetries should be associated with the object role (P in our terminology). But this is not the case, at least with regard to newness and animacy. For both scales, the most extreme asymmetries are associated with A, for which high-prominence values far outweigh low-prominence values: In actual usage, a new A is a rare event (< 10%), as A is almost categorically given. A non-human A is similarly unlikely (< 15%). Thus, if languages are designed to mark the infrequent or unexpected, then we would expect DAM based on newness and animacy to be a very widespread phenomenon, at least more so than DOM.

16 The indication of baseline figures is similar in spirit at least to Levshina's (2021) concept of cue reliability, though we have calculated the baseline probability for the realization of a particular value (e.g. definiteness) across all NPs in the corpus, regardless of role (i.e. including those outside of transitive clauses), while her domain for calculating cue reliability is the sum of all tokens of A and P, thus disregarding the rest of the referential expressions in the corpora from which the transitive clauses have been extracted.

For P, on the other hand, a strong asymmetry is found only in animacy, where a human P is comparatively rare, though with much lower magnitude than found for the corresponding association with A. In terms of newness, role–reference expectations are essentially the same as for A: A new P is not a common event (more than 60% of P arguments are given), albeit not as rare as a new A. These observations would thus falsely predict a high frequency of DAM systems based on newness and animacy across the language of the world, or at least higher than the frequency of comparable DOM systems. But as mentioned above, DOM conditioned on newness and animacy is overall more frequent than DAM.

A more realistic perspective in terms of expectedness and efficiency, however, is to consider the dimensions of newness and humanness in relation to the baseline, the mean value for given and human referential NPs in all roles (i.e. not only A and P) in the corpora, indicated by the horizontal lines in Figure 3. In other words, the baseline gives us an indication of what percentage of referential expressions exhibit the respective high-prominence value across the entirety of discourse. With regard to newness, we see that the baseline rate (i.e. given rather than new arguments) in our corpora is above 80%. Most of what we talk about in actual usage is given, while new information by comparison is sparsely scattered in discourse. From this perspective, the rates for a given P are indeed unexpected, as they fall below the overall baseline. Once the baseline is taken into consideration, we can derive the correct prediction for DOM. Yet this perspective still fails to account for the paucity of newness and definiteness-based DAM (as opposed to humanness-based DOM) in the languages of the world. Nevertheless, this example highlights the fact that raw frequency data, as cited for instance in Jäger (2007) or Haspelmath (2021b), are potentially misleading and need to be qualified with some indication of baseline frequencies (cf. Levshina 2021 for a similar point).

The feature of person differs from newness and animacy in a number of ways, in itself a noteworthy finding. First, we note a reversal in the magnitude of dispersion for A and P respectively. With newness and humanness, the P values are more variable, while A clusters very tightly, but with person, this picture is reversed. The large range of values for A can be explained along the following lines: The actual rates of first and second person arguments in a particular corpus are largely dictated by its content. For example,

the low outlier value for A is the Persian corpus, which happens to contain exclusively *Pear story* retellings and hence scarcely any first or second person arguments. Conversational data, on the other hand, is characterized by high rates of first and second person (e.g. 64% in the conversational English data cited in Haig 2018: 810), and over 50% for the conversational data reported in Levshina (2021). The Multi-CAST corpora are composed predominantly of narrative texts, which yield overall low levels of first and second person arguments. Had we included more conversational data in our sample, the median rate would have risen accordingly. The rates for first and second person forms in the P role, however, are relatively impervious to content, as shown in Haig (2018: 810–811): Regardless of content, first and second person P do not rise above a rate of 20%. The uniform low frequency of first and second person P, and the general lack of a clear pattern for person with A, correctly predicts person-based DOM to be regularly attested, but falsely predicts person-based DAM to be rare or even absent entirely.

With regard to the reliability of our figures, we note that they exhibit overall very similar tendencies to those of "cue availability" for animacy and newness A and P in Levshina (2021: Figs. 1 and 2), despite notable differences in corpus size, language sample, and text type (conversational in Levshina vs. narrative in Multi-CAST) and minor differences in annotation and coding procedures. This provides independent confirmation that these proportions represent stable and consistent values that characterize discourse cross-linguistically. One point of difference concerns the higher rates of first and second person A arguments in Levshina's corpora (mean of over 50%). But as pointed out above as well as in Haig (2018), rates of first and second person referents are highly sensitive to text type; the higher values reported in Levshina (2021) are entirely predictable from the conversational nature of the data.

3.2.3 Summary

In sum, our data confirm the basic hypothesis outlined above in (7), repeated here for convenience:

(9) An A argument is more frequently associated with the high-prominence values, while a P argument is more frequently associated with the low-prominence values in (6).

However, our data also reveal quite striking differences in the frequency asymmetries across the three referential scales. To the extent that these differences represent consistent patterns in spoken language, an efficiency-based account for the emergence of grammar would predict that these differences would be reflected in some way in the typological distribution of DAM and DOM across the languages of the world. We find little evidence for this; reliance on absolute measures of frequency would actually make the wrong predictions, namely a predominance of DAM as opposed to DOM, the existence of newness-based DOM, and a lack of person-based DAM. However, once baseline values are taken into account, the relative differences between A and P appear less extreme, and the overall picture appears more compatible with the cross-linguistic findings, though there is still no straightforward mapping of usage frequency to the generally estimated cross-linguistic distribution of DOM and DAM.

Finally, it is worth considering whether frequencies of role–reference associations considered for individual roles (A and P) are the relevant parameter. Instead, one might invoke a more general principle that overt case marking of any kind makes most sense for arguments that are overtly realized – and those are, all other things being equal, much more likely to be P than A (see Schiborr 2021: 153–160). Thus the comparative rarity of DAM is connected to the fact that A is not reliably available as the locus for indicating the unexpected because it is so often left unexpressed.

A further possible avenue of inquiry is to focus not on the individual roles (A, P, etc.) but on a more general notion of frequent constellations of A and P (which Haspelmath 2021b refers to as "scenarios"), characterized by the relative prominence of A and P. Wherever expected prominence constellations are absent, additional marking is predicted to occur – somewhere in the clause. But crucially, the actual location of that marking may be dictated by fairly

random language specific factors, and is hence impervious to the specifics of frequency distributions. This view has the advantage of subsuming inverse systems and other variant systems, which emerge simply as a subtype of a more general phenomenon of marking the unexpected, much in the spirit of Haspelmath's (2021a; b) conclusions. Finally, efficiency considerations may come to a natural limit where marking of extremely unpredictable and rare instances of an asymmetrical category are concerned: Given their overall rarity, efficiency would predict that languages do not afford the development of a specialized marker for these cases – which would make the system overall less efficient – but rather take the risk of leaving matters vague on rare occasions during communication. This is much in the same sense that ambiguity can be seen as an overall advantageous efficiency property of human languages (Piantadosi et al. 2012).

4 Conclusions

In the preceding sections we have exemplified how data from Multi-CAST can be exploited in significant ways for linguistic typology, most specifically for research on features that exhibit variability in usage. While this is already a major focus of corpus-based typological research, the current reliance on written language corpora (as with Universal Dependencies and others) is not fully consonant with research questions that probe the role of language processing and efficiency in shaping the grammars of the world's languages. For these purposes, we maintain that spoken-language data from typologically diverse languages provide the gold standard, though we readily concede the considerable methodological challenges in data compilation and annotation. But even working with the current modest dimensions of Multi-CAST, we are able to identify previously unnoticed regularities in the organization of discourse cross-linguistically. Our findings on the distribution of lexical forms in discourse (Section 3.1) point to a hitherto unnoticed degree of cross-linguistic unity in this regard, which counterbalances claims in the literature regarding language-specific differences. In Section 3.2, we demonstrated that the coding asymmetries characteristic of DOM and DAM reflect broad frequency differences in usage. But we also demonstrated that the features of givenness, animacy, and person do not pattern alike in discourse, and on closer

inspection, do not translate straightforwardly into an explanation of the typological distribution of DAM and DOM across the world's languages. We also noted that a better fit is obtained if baseline figures are included, again underscoring the necessity for more comprehensively annotated data sets. In view of the limited corpus representativity of Multi-CAST, we hasten to stress the preliminary nature of these findings, but the degree of uniformity is sufficient to allow us to challenge some cherished assumptions of language typology, and to formulate novel and testable hypotheses for future research.

The Multi-CAST project demonstrates that typological research in discourse and grammar based on typologically diverse, spoken language corpora is possible. At the same time, it requires a number of quite laborious steps of pre-processing and annotation, as well as engagement with descriptive and analytical issues relevant to each individual language sampled (see Schnell et al., forthcoming, for a more detailed account of annotation and analysis procedures). Corpus-based typology accomodates a broad spectrum of corpus designs (see Schnell et al., this volume); we here hope to have demonstrated that spoken-language corpora with minimal content control, a typical product of language documentation efforts, can contribute significantly towards a usage-based approach to language typology.

References

Adibifar, Shirin. 2016. Multi-CAST Persian. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (`https://multicast.aspra.uni-bamberg.de/#persian`).

Amberber, Mengistu. 2008. Semantic primes in Amharic. In Goddard, Cliff (ed.), *Cross-linguistic semantics*, 83–119. Amsterdam: John Benjamins.

Andrews, Avery. 2007. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description, volume 1*: *Clause structure*, 132–223. Cambridge: Cambridge University Press.

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).

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

Bickel, Balthasar & Witzlack-Makarevich, Alena & Zakharko, Taras. 2015. Typological evidence against universal effects of referential scales on case alignment. In Bornkessel-Schlesewsky, Ina & Malchukov, Andrej L. & Richards, Marc D. (eds.), *Scales and hierarchies*: *A cross-disciplinary perspective*, 7–44. Berlin: Mouton de Gruyter.

Bisang, Walter. 2015. Hidden complexity – The neglected side of complexity and its implications. *Linguistic Vanguard* 1(1). 177–187. (https://doi.org/10.1515/lingvan-2014-1014).

BNC Consortium. 2008. *The British National Corpus*. Oxford: Oxford University Computing Services. (https://www.natcorp.ox.ac.uk/).

Bogomolova, Natalia & Ganenkov, Dmitry & Schiborr, Nils N. 2021. Multi-CAST Tabasaran. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#tabasaran).

Bresnan, Joan & Dingare, Shirpa & Manning, Christopher. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Butt, Miriam & King, Tracy H. (eds.), *Proceedings of the LFG 01 Conference*, 13–32. Stanford, CA: CSLI Publications.

Brickell, Timothy C. 2016. Multi-CAST Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#tondano).

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, Charles N. (ed.), *Subject and topic*, 25–55. New York: Academic Press.

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

Comrie, Bernard. 1979. Definite and animate direct objects: A natural class. *Linguistica silesiana* 3. 13–21.

Coon, Jessica & Preminger, Omer. 2017. Split ergativity is not about ergativity. In Coon, Jessica & Massam, Diane & Travis, Lisa D. (eds.), *The Oxford handbook of ergativity*, 226–252. Oxford: Oxford University Press.

Cysouw, Michael. 2009. The asymmetry of affixation. *Snippets* 20(3). 10–14.

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).

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA).* (https://english-corpora.org/coca/).

Donaldson, Tamsin. 1980. *Ngiyambaa: The language of the Wangaaybuyan.* Cambridge: Cambridge University Press.

Du Bois, John. 1985. Competing motivations. In Haiman, John (ed.), *Iconicity in syntax*, 343–366. Amsterdam: John Benjamins.

Du Bois, John. 2003. Argument structure: Grammar in use. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*: *Grammar as architecture for function*, 11–60. Amsterdam: John Benjamins.

Engelhardt, Paul E. & Bailey, Karl G. D. & Ferreira, Fernanda. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language* 54(4). 554–573. (https://doi.org/10.1016/j.jml.2005.12.009).

Evans, Nicholas. 2020. Why the comparability problem is still central in typology. *Linguistic Typology* 24(3). 417–425. (https://doi.org/10.1515/lingty-2020-2055).

Fauconnier, Stefanie & Verstraete, Jean-Christophe. 2014. A and O as each other's mirror image? *Linguistic Typology* 18(1). 3–49.

Foley, William A. & Van Valin, Robert D., Jr. (eds.). 1984. *Functional syntax and universal grammar.* Cambridge: Cambridge University Press.

Forker, Diana & Schiborr, Nils N. 2019. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts.* (https://multicast.aspra.uni-bamberg.de/#sanzhi).

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.

Garside, Roger. 1993. The marking of cohesive relationships: Tools for the construction of a large bank of anaphoric data. *ICAME* 17. 5–27.

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).

Givón, Talmy (ed.). 1979. *Discourse and syntax*. New York: Academic Press.

Givón, Talmy (ed.). 1983. *Topic continuity in discourse* (Typological Studies in Language 3). Amsterdam: John Benjamins.

Hadjidas, Harris & Vollmer, Maria C. 2015. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#cypgreek).

Haig, Geoffrey. 2017. Deconstructing Iranian ergativity. In Coon, Jessica & Massam, Diane & Travis, Lisa (eds.), *The Oxford handbook of ergativity*, 465–500. Oxford: Oxford University Press.

Haig, Geoffrey. 2018. The grammaticalization of object pronouns: Why differential object indexing is an attractor state. *Linguistics* 56(4). 781–818. (https://doi.org/10.1515/ling-2018-0011).

Haig, Geoffrey. 2020. *Stability and adaptivity of word order in the Western Asian transition zone: Evidence from West Iranian*. Paper presented at the workshop Tracing Contact in Closely Related Languages, Zürich, Switzerland, 19–20 November 2020.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (Version 7.0)*. (https://multicast.aspra.uni-bamberg.de/#annotations).

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

Haig, Geoffrey & Vollmer, Maria C. & Thiele, Hanna. 2019. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#nkurd).

Hardie, Andrew. 2003. Developing a tagset for automated parts-of-speech tagging in Urdu. In Archer, Dawn & Rayson, Paul & Wilson, Andrew & McEnery, Tony (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, 298–307. Lancaster: Lancaster University.

Haspelmath, Martin. 2021a. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633. (https://doi.org/10.1017/S0022226720000535).

Haspelmath, Martin. 2021b. Role-reference associations and the explanation of argument coding splits. *Linguistics* 59(1). 123–174. (https://doi.org/10.1515/ling-2020-0252).

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

Huang, Yan. 2000. *Anaphora: A cross-linguistic study*. Oxford: Oxford University Press.

Hundt, Marianne & Lehmann, Hans M. & Schneider, Gerold (eds.). 2016. *The International Corpus of English (ICE)*. (https://www.ice-corpora.uzh.ch/en.html).

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.

Jäger, Gerhard. 2007. Evolutionary game theory and typology: A case study. *Language* 83(1). 74–109. (https://doi.org/10.1353/lan.2007.0020).

Karttunen, Lauri. 1976. Discourse referents. In McCawley, James D. (ed.), *Syntax and Semantics 7*: *Notes from the Linguistic Underground*, 363–385. New York: Academic Press.

Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.

Kehler, Andrew. 2004. Discourse coherence. In Horn, Laurence & Ward, Gregory (eds.), *Handbook of pragmatics*, 241–265. Malden, MA: Blackwell.

Kimoto, Yukinori. 2019. Multi-CAST Arta. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#arta).

Kurabe, Keita. 2021. Multi-CAST Jinghpaw. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#jinghpaw).

Labov, William & Waletzky, Joshua. 1967. Narrative analysis. In Helm, June (ed.), *Essays on the verbal and visual arts*, 12–44. Seattle, WA: University of Washington Press.

Leech, Geoffrey & Barnett, Ross & Kahrel, Peter. 1996. EAGLES final report and guidelines for the syntactic annotation of corpora. *EAGLES report* EAG-TCWG-SASG/1.5. (https://www.ilc.pi.cnr.it/EAGLES96/home.html).

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Languages in Contrast* 23(3). 533–572. (https://doi.org/10.1515/lingty-2019-0025).

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

Mardale, Alexandru & Montrul, Silvina. 2020. *The acquisition of differential object marking*. Amsterdam: John Benjamins. (https://doi.org/10.1075/tilar.26).

McLuhan, Marshall. 1964. *Understanding media: The extension of man*. New York: New American Library.

Meng, Chenxi. 2019. Multi-CAST Tulil. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#tulil).

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).

Mettouchi, Amina & Vanhove, Martine & Caubet, Dominique (eds.). 2015. *Corpus-based studies of lesser-described languages: The CorpAfroAs corpus of spoken AfroAsiatic languages*. Amsterdam: John Benjamins.

Mitkov, Ruslan. 2000. Corpora for anaphora resolution: Mouton de Gruyter. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics*, 579–598. Berlin: John Benjamins.

Mosel, Ulrike & Schnell, Stefan. 2015. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#teop).

Ozerov, Pavel. 2018. Tracing the sources of information structure: Towards the study of interactional management of information. *Journal of Pragmatics* 138(1). 77–97. (https://doi.org/10.1016/j.pragma.2018.08.017).

Perlmutter, David. 1971. *Deep and surface constraints in syntax*. New York, NY: Holt, Rinehart and Winston.

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.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Cole, Peter (ed.), *Radical pragmatics*, 223–255. New York: Academic Press.

Riester, Arndt & Baumann, Stefan. 2017. *The RefLex scheme — Annotation guidelines* (SinSpeC: Working papers of the SFB 732 14). Stuttgart: University of Stuttgart. (https://elib.uni-stuttgart.de/handle/11682/9028) (Accessed 2018-03-01).

Schiborr, Nils N. 2015. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#english).

Schiborr, Nils N. 2018. Multi-CAST collection overview. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/).

Schiborr, Nils N. 2019. *multicastR: A companion to the Multi-CAST collection*. R package version 2.0.0. (https://cran.r-project.org/package=multicastR).

Schiborr, Nils N. 2021. *Lexical anaphora: A corpus-based typological study of referential choice*. Unpublished Ph.D. dissertation, University of Bamberg.

Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines (Version 1.1)*. Bamberg: University of Bamberg. (https://multicast.aspra.uni-bamberg.de/#annotations).

Schmidtke-Bode, Karsten & Levshina, Natalia. 2018. Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal. In Seržant, Ilja A. & Witzlack-Makarevich, Alena (eds.), *Diachronic typology of differential argument marking*, 509–537. Berlin: Language Science Press.

Schnell, Stefan. 2015. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#veraa).

Schnell, Stefan. 2018. Attention focus and information packaging in Vera'a demonstratives. In Riesberg, Sonja & Shiohara, Asako & Utsumi, Atsuko (eds.), *Perspectives on information structure in Austronesian languages*, 81–113. Berlin: Language Science Press. (https://doi.org/10.5281/zenodo.1402539).

Schnell, Stefan & Haig, Geoffrey & Schiborr, Nils N. & Vollmer, Maria C. Forthcoming. Are referent introductions sensitive to forward planning in discourse? Evidence from Multi-CAST. In Mattiola, Simone & Barotto, Alessandra (eds.), *Discourse phenomena in typological perspective*. Amsterdam: John Benjamins.

Schnell, Stefan & Haig, Geoffrey & Seifart, Frank. This volume. The role of language documentation in corpus-based typology. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 1–28. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74656).

Schnell, Stefan & Schiborr, Nils N. 2018. Corpus-based typological research in discourse and grammar: GRAID and Multi-CAST. *Asian and African Languages and Linguistics* 12. 1–16. (https://hdl.handle.net/10108/91145).

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

Sinnemäki, Kaius. 2014. A typological perspective on differential object marking. *Linguistics* 52(2). 281–313. (https://doi.org/10.1515/ling-2013-0063).

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.

Thieberger, Nick & Brickell, Timothy. 2019. Multi-CAST Nafsan. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#nafsan).

Visser, Eline. 2021. Multi-CAST Kalamang. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#kalamang).

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.

Vollmer, Maria C. 2020. Multi-CAST Mandarin. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*: *Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/#mandarin) (Accessed 2020-01-03).

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