The discourse basis of ergativity revisited: Online Appendices

Geoffrey Haig, Stefan Schnell

Language, Volume 92, Number 3, September 2016, pp. s1-s14 (Article)

Published by Linguistic Society of America
DOI: https://doi.org/10.1353/lan.2016.0044

THE DISCOURSE BASIS OF ERGATIVITY REVISITED: ONLINE SUPPLEMENTARY MATERIAL

GEOFFREY HAIG
University of Bamberg

STEFAN SCHNELL
University of Melbourne

This document contains illustration and justification of the methodology used in our article 'The discourse basis of ergativity revisited' (henceforth H&S). We begin by outlining the Multi-CAST database ('Multi-Language Corpora of Annotated Spoken Texts'), from which large portions of the data stem. Multi-CAST is a project designed for crosslinguistic, corpus-based studies into argument realization in spoken discourse, covering topics such as REFERENTIAL DENSITY (Bickel 2003, Noonan 2003), referentiality (Kibrik 2011), or PREFERRED ARGUMENT STRUCTURE (Du Bois 1987, 2003). The focus of H&S is preferred argument structure. It is worth pointing out that text-based, or corpus-based, typology is a field in its infancy (Schnell 2012, Cysouw & Wälchli 2007, Wälchli 2006, 2009), and researchers are still engaging in exploratory studies to gauge the validity of different data types. Our study and this document are also intended as a contribution to the ongoing methodological discussion.

Contents of this document:

1. CORPUS SIZE AND COMPOSITION (see H&S §3). The Multi-CAST corpus currently contains recordings of spontaneous spoken narrative texts from five languages, together with transcription, translation, and further linguistic annotation, available online at https://lac.uni-koeln.de/en/multicast/. Obviously, for quantitative approaches to cross-corpus comparison, corpora should aim at maximal size. However, manual annotation of natural spoken language data, in particular of lesser-described languages, is so time- and resource-costly that in published research, many of the available corpora (half of those listed in Table 2 below) do not exceed 1,000 clause units. The Multi-CAST corpora all contain a minimum of 1,000 clause units, making them broadly comparable to available data sets. Table 1 gives an overview of the Multi-CAST corpora, while Table 2 provides the relevant data from previously published sources that have been included in the analysis.

Our findings are based on the total number of clauses in Tables 1 and 2, thus a total of 25,618 clauses.

LANGUAGE	CLAUSES	GENRE	SOURCE
Vera'a	3,789	traditional narratives (11 texts)	Schnell 2016
Teop	1,328	traditional narratives (4 texts)	Mosel & Schnell 2016
N. Kurdish	1,205	traditional narratives (2 texts)	Haig & Thiele 2016
English	2,360	monologic oral history (Kent dialect of English, 1 text)	Schiborr 2016
Cypriot Greek	1,078	traditional narratives (3 texts)	Hadjidas & Vollmer 2016
TOTAL CLAUSES	9,670		

TABLE 1. The Multi-CAST corpus.

LANGUAGE	CLAUSES	GENRE	SOURCE
Sakapultek	456	Pear story retellings (18 texts)	Du Bois 1987, table 2
English	704	informal conversation	Kärkkäinen 1996
English	484	classroom interactions, teachers' contributions only	Kumpf 2003
English	1,313	televized interviews, interviewees' contributions only	Everett 2009
English	1,654	Pear story retellings (20 texts)	Kumagai 2006
Portugese	412	televized interviews, interviewees' contributions only	Everett 2009
Roviana	339	monologic texts, variety of topics and genres	Corston-Oliver 2003
Korean	4,363	children's speech (1 yr. 8 months to 2 yrs. 10 months)	Clancy 2003
To'aba'ita	1,278	six traditional narratives, third person only	Lichtenberk 1996
Mapudungun	700	transcriptions of spoken narratives, third person only	Arnold 2003
Yagua	1,156	traditional folkloric narrative	Payne 1993
Gorani	483	traditional folkloric narrative	Mahmoudveysi et al. 2012
French	1,056	structured interviews, mostly monologic	Ashby & Bentivoglio 1993
Spanish	1,550	structured interviews, mostly monologic	Ashby & Bentivoglio 1993
TOTAL CLAUSES	15,948		

TABLE 2. Previously published data included in H&S.

An issue that needs to be considered for cross-corpus comparability is that of variation in genre and discourse type. Some researchers advocate the use of standardized stimuli to elicit texts of broadly comparable content, for example the Frog story (e.g. Noonan 2003) or Pear film retellings (see Chafe 1980 for details on the stimulus and the elicitation experiment); the Sakapultek data from Du Bois 1987 are of the latter type. The Multi-CAST database, by contrast, contains what can be called 'original texts' (see discussion in Haig, Nau, Schnell, & Wegener 2011, Haig, Schnell, & Wegener 2011), that is, monologic

narratives involving spontaneous, nonelicited content, either oral history from the speech community or traditional folkloric tales. These have been recorded in their respective cultural contexts. Although these texts often represent 'staged events' (that is, they were produced for the sake of being recorded; cf. Himmelmann 1998) and are obviously not as natural as, for instance, unrecorded everyday conversations, they are undoubtedly closer to this ideal than the Pear film texts and could be expected to reflect more closely the kinds of habituated discourse structures that characterize and distinguish different language communities.

A related issue arises from the predominance of reference to speech-act participants versus third persons in different types of text. In the corpora in Tables 1 and 2, the majority of texts involve mostly third-person reference. The exceptions are the conversational English data from Kärkkäinen 1996 and the Korean data from Clancy 2003, as well as the English and Portuguese corpora used by Everett (2009), which consist of TV-broadcasted talk show interactions; these have predominantly reference to speech-act participants. Texts with different reference profiles of this sort are handled quite variably in different studies: some sources explicitly note the exclusion of first/second-person forms (e.g. Arnold 2003:229, Kumagai 2006:677), while others apparently include first/second-person forms in their counts (Everett 2009). The Multi-CAST texts involve predominantly third-person referents. Where first- or second-person reference occurs within direct speech, these have been coded in the corpus, but they have not been included in the data analysis (see the appendices for the raw data). This appears to be the most logical procedure, because only in the third person do speakers actually have a choice between a lexical and a nonlexical expression.

2. CORPUS MARK-UP: HOW THE MULTI-CAST DATA ARE ANNOTATED. The texts in Multi-CAST have been annotated according to a standardized glossing system, called GRAID (Grammatical Relations and Animacy In Discourse), developed by Geoffrey Haig and Stefan Schnell. GRAID was specifically designed for the purposes of crosslinguistic quantitative analysis of discourse, in the tradition of Du Bois 1987 or Bickel 2003. The annotations target linguistic categories of sufficient generality to be identified in the vast majority of—if not all—languages, for example, 'subject of an intransitive verb (S)' and so forth. GRAID does not, however, include annotation of information status (e.g. new vs. given), though the format of the annotation easily allows for additional layers of annotation to be added. Because the annotation is systematically linked to the sound files of the recordings, future investigation targeting the interface of prosody and syntax, for example, can be undertaken with these data.

Annotating a corpus with GRAID involves identifying the major syntactic constituents (predicates and their arguments) and tagging them according to form and function. Central among the functions identified are S (intransitive subject), A (transitive subject), and P (direct object). Two issues are important here: (i) the identification of S, A, and P in a given clause; (ii) distinguishing the different realization types of S, A, and P, in particular between lexical realizations (NPs with a lexical head) and nonlexical realizations (which generally subsume pronominal elements and referential null elements). GRAID works with a standardized set of tags (approximately forty) and a simple syntax for combining them. The complete set of glossing conventions together with detailed discussion of particular issues of linguistic analysis can be found in the GRAID manual 7.0 (Haig & Schnell 2014). Although GRAID makes no claim to solving the multifarious problems that arise in syntactic annotation, it does provide a set of decision-making procedures and principles that can be applied to resolve problems of analysis. In what follows, we outline and illustrate some of the recurrent issues in coding and analysis of discourse data.

All of the sources dealing with preferred argument structure work with the categories S, A, and P (or O in some publications). However, there are a number of different ways that these categories may be

interpreted (Donohue 2008, Haspelmath 2011), and there are undoubtedly interstudy differences in the procedures adopted. In the GRAID annotation system, Andrews's 1985, 2007 definition of S, A, and P is adopted, grounded in what Haspelmath (2011) calls the Comrie'an approach (see Comrie 1981). Hence, A is identified with the syntactic function of an NP that, in a two-argument clause construction, bears the same argument-encoding features (e.g. case and agreement properties) as the agent argument of a primary transitive predicate like 'kill' or 'smash', and P is the syntactic function of an NP that, in a two-argument clause construction, bears the same argument encoding as the patient argument of such a primary transitive predicate. S is the syntactic function of an argument that, in an intransitive clause construction, either has only this one single argument or has a second argument, the coding of which matches neither that of an agent nor that of a patient of a primary transitive predicate. Clauses with two arguments, one bearing A function and one P function, are transitive. All other types of clauses are intransitive.

This definition nevertheless leaves certain issues unsolved. To give one example: in English, numerous expressions of the type *take a walk*, *have a bath*, or *kick the bucket* exist, involving an apparently transitive verb and an object NP. Although these expressions are clearly transitive according to our definition, they are highly lexicalized, with the NP complement lacking specific reference (see Singer 2011 for discussion). Analysts coding discourse data must reach a decision on whether to consider them transitive (in which case the subjects are coded as A) or intransitive (leaving the subjects as S). Kumpf (2003), for example, opts to treat them as transitive, but most other researchers do not make their decisions explicit. In the Multi-CAST corpora, the decision is made on the basis of language-specific criteria and is explicitly outlined in the accompanying documentation. While decisions may thus not always lead to maximal cross-corpus comparability, they are at least explicit, and other researchers are free to peruse the relevant examples in context—and apply different decisions, if they consider them to be more appropriate.

Another example involves the conception of S, A, and P as syntactic functions, as reflected in Andrews's 2007 definition adapted in GRAID, or semantic macroroles, as done in RRG (Van Valin & LaPolla 1997). For instance, a study based on a semantic macrorole approach (see Haspelmath 2011 for discussion) will consider incorporated objects to be P (or O) arguments, while actor-like arguments would be coded as A (see for instance Naess 2007). According to the GRAID conventions, one would not assume a P argument at all, and would treat the other argument as S rather than A. A further issue is the classification of certain constructions as transitive or intransitive: in Kumagai 2006, the second NP of an English existential construction *there is NP* is treated as P, on the grounds that it directly follows the predicate. This decision appears to be a minority one, and it is not followed for the Multi-CAST English corpus (Schiborr 2016), but again, few researchers are as explicit as Kumagai (2006) in laying out their coding procedures.

The second key issue concerns the distinction between lexical (i.e. referential expressions involving an NP) and nonlexical (referential expressions involving either an overt pronominal element or a phonologically unexpressed element). While for most languages a broad binary division into lexical and nonlexical can be implemented, and researchers evidently feel little necessity to elaborate on the coding practices, a number of problematic cases arise, for instance: indefinite or interrogative expressions may be either pronouns or nouns (heading NPs) in different languages; quantifiers, numerals, demonstratives, and so on defy ready classification as either NPs or pronouns. According to GRAID conventions, a gloss for 'pronoun' is generally used in cases of what Lyons (1968) calls 'definite pronouns', which will have context-retrievable identifiable referents. This includes first- and second-person pronouns referring to speech-act participants. In Multi-CAST corpora, such first- and second-person pronouns will usually refer to characters in a given story when they are participating in depicted speech acts. Third-person pronouns are in most cases anaphoric pronouns referring back to aforementioned entities. A particularly problem-

atic issue is the identification of referential zero elements, which is only rarely explicitly discussed. The GRAID annotation system provides the following guidelines for identifying zero arguments. A zero argument is counted where the following three conditions are met: (i) the argument is licensed by the lexical argument structure of the verbal lexeme (or the serial verb construction) involved; (ii) the relevant argument position must be clearly associated with a specific referent, retrievable from the discourse context; (iii) the syntactic construction must allow for the overt realization of the argument in question. The second criterion will thus exclude 'semantic arguments' with generic, nonspecific reference: for instance, where a verb of consumption like 'eat' occurs without an expression of something eaten, we do not assume a zero argument, unless it can be construed as referring to a specific referent mentioned elsewhere in the context. The third criterion excludes arguments for which overt realization is prevented by a particular syntactic configuration, for instance, in different types of nonfinite clause construction in which overt expression of the highest-ranking argument role is systematically blocked; hence no contrast in form is evoked (see Bickel 2003 for the same approach and justification).

By way of illustration, we provide a short section of GRAID annotated text in 2 and 3 below, extracted from the Multi-CAST corpus. The basic principles of GRAID are as follows: each word of the object language is provided with a GRAID gloss, which combines a symbol indicating its form (e.g. zero, pronoun, verb, etc.) with a symbol indicating its syntactic function. Form glosses may further be modified with a value indicating person or animacy features. Full details are available in the GRAID manual 7.0 (Haig & Schnell 2014); the basic principles are illustrated in the following examples.

(1) He swallowed a snake
pro.h:a v:pred ln np:p

The gloss 'pro.h:a' means: 'independent pronoun, human referent, in A function (subject of transitive clause)'. The gloss 'np:p' is read as 'head of a lexical NP, with nonhuman reference, and in P function'. The gloss 'ln' covers elements within an NP, standing to the left of the lexical head of the NP (cf. §2.8 of the GRAID manual 7.0).

Of course, natural spoken discourse does not consist solely of strings of well-formed clauses such as 1. Many utterances pose considerable difficulties of analysis, which cannot be readily accounted for by the categories available in GRAID. Such unaccountable structures can be glossed 'nc' ('not classifiable'), and thereby be excluded from the quantitative analysis. Obviously, the amount of discourse not considered for a specific text should not exceed a critical proportion, and we note that for none of the texts in Multi-CAST does the proportion of 'nc'-chunks exceed more than 10% of the total.

Below we illustrate a short stretch of discourse from the Northern Kurdish (West Iranian, Iranian, Indo-European; Eastern Turkey) subcorpus in Multi-CAST (Haig & Thiele 2016). All texts are initially segmented into utterance units, on the basis of pauses and major syntactic boundaries, and transcribed. Example 2 shows the beginning of utterance unit (nkurd_muserz02_015), together with a free translation.

(2) Tîne, dibê: 'Tu šerjêbike.' Waya jî dibê: 'Ez çer çêlekê šerjêkim?' […]
'(She) brings (it), says: "you will kill (it)." And he says: "Why should I kill the cow?" '
(Northern Kurdish, Multi-CAST nkurd_muserz02_015)

Each annotation unit is then segmented into minimal clauses, consisting of a single predicate and its associated arguments and adjuncts. At this point, any zero arguments (see above for conditions on assuming zero arguments) are inserted into the clause units, signaled by the digit '0'. A second annotation tier

('gloss') contains a word-for-word gloss of each grammatical word of the clause-unit tier.[1] A third tier, aligned with the word-for-word gloss, contains the GRAID annotation. The result is as follows, with an explanation of the GRAID symbols provided in Table 3.

(3) Sample of GRAID annotations from Northern Kurdish (Haig & Thiele 2016)

gram. words	##	0	0	Tîne	#	0	dibê
gloss	##	0_she	0_it	brings	#	0_she	says
GRAID	##	0.h:a	0:p	v:pred	#	0.h:s_ds	ds_v:pred
gram.words	#ds	Tu	0	šerjêbike			
gloss	#ds	you	0_it	will.slaughter			
GRAID	#ds	pro.2:a	0:p	v:pred			
gram.words	##	Waya	jî	dibê:			
gloss	##	he	too	says			
GRAID	##	pro.h:s_ds	other	ds_v:pred			
gram.words	#ds	Ez	çer	çêlekê	šerjêkim		
gloss	#ds	I	why	cow	will.kill		
GRAID	#ds	pro.1:a	other	np:p	v:pred		

SYMBOL	EXPLANATION
##	left-hand clause boundary, main clause
0.h:a	referential, but nonovert element (zero), with human reference and in A function
0:p	zero, nonhuman referent (here: the cow) with P function (direct object)
v:pred	finite verb
#	left-hand clause boundary, dependent clause
0.h:s_ds	zero with human referent, intransitive subject function (S), of a verb of speech (the transitivity status of verbs of speech is often contentious; we have therefore introduced the additional gloss s_ds for subjects of verbs of speech)
ds_v:pred	finite verb of speech
#ds	left-hand clause boundary, direct speech
pro.2:a	independent pronoun, second person, in A function
other	any element that is irrelevant to argument/predicate relations
np:p	lexical noun phrase, nonhuman referent, in P function

TABLE 3. Explanation for the GRAID symbols in 3.

Once texts have been annotated in this manner, it is a straightforward task to extract and quantify the GRAID glosses regarding, for example, lexical versus nonlexical expression types: they can be exported as a string of symbols and then analyzed using regular expressions in a variety of software applications. The raw figures for the categories relevant in the present context are presented below in Appendix B.

3. INTERPRETING QUANTITATIVE DATA ON PREFERRED ARGUMENT STRUCTURE: CONTRASTING TWO APPROACHES. As regards quantitative analysis of the data, two distinct methods have been applied in research on preferred argument structure, each yielding a different outcome. Occasionally, these differences

[1] The contributions to Multi-CAST vary in the degree of detail provided in the word-for-word glossing. The Northern Kurdish gloss represents the simplest end of the scale, with just a word-for-word translation, while others provide morphemic segmentation and identification of individual morphemes.

have been neglected in comparisons of figures from studies applying different methods. This is problematic, as will become clear.

The first approach takes the number of referential expressions within each of the categories S, A, and P and calculates the proportion of lexical expressions within each total, yielding a percentage value of 'lexicality' for each role. On this approach, what is compared is the propensity of each role to host lexical expressions. The second approach, by contrast, takes as its frame of reference the total number of lexical expressions in the entire corpus and calculates the percentage of that total which is accounted for by each of the three roles. Thus the question here is: Where do lexical arguments go? Note that on this approach, the total number of arguments (i.e. lexical and nonlexical) is in fact not relevant; only the number of lexical arguments and their distributions across S, A, and P are considered. In our work we have adopted exclusively the first approach.

It is worth considering the differences between these two approaches in a little more detail. Table 4, from Du Bois 2003:37, provides data from seven languages analyzed according to the second approach. The values for S and P have been highlighted.

ROLE	A		S		P		TOTAL	
LANGUAGE	*N*	%	*N*	%	*N*	%	*N*	%
Hebrew	18	8	103	44	111	48	232	100
Sakapultek	11	5	126	58	81	37	218	100
Papago	37	10	169	47	152	42	358	99
English	21	8	90	35	146	57	257	100
Spanish	35	6	215	36	341	58	591	100
French	32	5	290	45	324	50	646	100
Japanese	48	7	320	48	293	44	661	100

TABLE 4. Percentage of lexical arguments in different syntactic roles (as totals of all lexical arguments in the respective texts (in other words, answering the question: 'Where do lexical arguments go?').

On the face of it, the data in Table 4 seem to support the claim for the unity of S and P in terms of lexicality. However, as Haspelmath (2006) and Everett (2009) point out, this take on the data obscures the fact that in natural language discourse (at least in the majority of corpora known to us), intransitive clauses outnumber transitive clauses, with 60/40 being a typical ratio. Therefore, for any given lexical expression, the odds are stacked in favor of it being hosted by the S function, as opposed to A or P. Thus, a higher percentage for lexical S trivially reflects the higher overall proportion of S functions within all argument expressions. The apparent unity of S and P in Table 4 is thus in part an artifact of a particular take on the data.

To what extent different perspectives yield quite different results can be demonstrated by applying them to the exact same data sets and monitoring the differences. Consider the French and Spanish data in Ashby & Bentivoglio 1993, included in Table 4 above. When analyzed according to the second perspective ('Where do lexical arguments go?'), the figures for the lexicality of S are 36% for Spanish and 45% for French. But if we interpret precisely the same data, but from the first perspective (i.e. 'How lexical is each role?'), the respective percentages for the lexicality of the S role drop dramatically. The percentages are given in 4.

(4) Lexicality of S: 'What proportion of the S arguments are lexical?'
Spanish: 22% (215 of total of 979 S arguments are lexical)
French: 28% (290 of total of 1,025 S arguments are lexical)

It is evident that the figures for S in Spanish and French given in Table 4, based on identical data, are far higher than those given in 4, and are in fact very close to the figures for P. The same effect can readily be replicated for any of the data sets from Table 1 above by analyzing the raw data from the appendices. The second perspective thus suggests an overall higher lexicality of S, due to the higher numbers of S in the data, and hence a greater similarity of S to P; the percentages for A and P, by contrast, change only minimally. The net effect of interpreting the data in this manner is to increase the proximity of S and P, heightening the impression of an ergative bias in discourse.

In our analysis, we consistently adopt the first of the two approaches, for the following reasons. First, speakers' prime concern is to verbalize interlinked, individual states of affairs with different numbers and types of participatory roles. In other words, discourse unfolds in a linear fashion, rather than being preplanned in a monolithic chunk. Speakers do not plan how to deploy a given inventory of lexical expressions, as suggested by the second approach ('Where do lexical expressions go?'). Second, relevant grammaticalization processes of ergative patterns of argument encoding would presumably be triggered by the high proportion of identical forms of expression in particular argument functions, comparable to the frequency effects on the reanalysis of word boundaries (Bybee 2001, Thompson & Hopper 2001). Finally, the alternative perspective only considers lexical core arguments, neglecting nonlexical arguments (for instance, pronouns). In fact, lexical arguments make up only a small percentage of arguments in a text, which may vary considerably across corpora (see, for example, Stoll & Bickel 2009 for a crosslinguistic investigation of 'Lexical referential density' and below for data from Papago).

This is not to deny that the alternative perspective may yield significant insights concerning the distribution of new or contrasted (hence primarily lexically expressed) referents in discourse. But for the specific research questions targeted in H&S, the first approach (as indeed originally adopted in Du Bois 1987) seems to be the more relevant one.

A final example of the issues involved comes from the data on Papago (Uto-Aztecan, Arizona; Payne 1987). The Papago data are included in Table 4 above and regularly cited as evidence for ergative patterning in discourse (Du Bois 2003:37, Everett 2009:3). They are repeated here in Table 5 for convenience.

ROLE	A		S		P		TOTAL	
	N	%	*N*	%	*N*	%	*N*	%
	37	10	169	47	152	42	358	99

TABLE 5. Papago: where do lexical arguments go?

In the literature, the shaded percentages in Table 5 are cited as evidence for the unity of S and P, and in fact, the value for S actually exceeds that for P, apparently strong evidence for an ergative bias in discourse. However, the data in Table 5 represent only the lexical NPs in core argument function, just a fraction of the total data. We know that the corpus includes 759 'usable clauses' (Payne 1987:759), with a corresponding total of approximately 759 subjects (S or A). However, the figures in Table 5 are based on just 206 (37 + 169) lexical S or A tokens, implying that approximately 553 of the available S/A positions in the corpus are nonlexical. The figures provided in the table give no information on those 553 nonlexical subjects (how many are S, and how many are A?); thus we cannot reliably reconstruct the respective degree of lexicality of the individual roles. It is quite possible that the share of lexical forms within all S arguments is considerably lower than the 47% indicated in Table 5. However, it is simply not possible to cross-check these data on this perspective, and we therefore do not include Papago in our own corpus. Finally, the unity of S and P suggested in Table 5 does not line up with Payne's own assessment of the

Papago data. In her summary discussion, she explicitly notes the lack of evidence for a unified function of S and P (her 'absolutive' category) in information management, and her findings are in fact much more similar to those of Chafe (1994:85–86) and (Prince 1981) for English, which posit the unity of S and A:

> Of all subjects, 79% encode given information; 86% encode definite information. Of all given subjects, 83% are S's, and 83% of all definite subjects are S's. These percentages show that S plays a strong role in encoding continuity. With regard to overt objects, 64% are given and 60% are definite. … Given that S is strongly associated with continuous definite information, and given that O [= P] is not clearly more strongly associated with new indefinite information than with continuous definite information, questions arise as to the strength and universality of a correlation of new information with the absolutive. [footnote omitted] (Payne 1987:800–801)

In sum, the same raw data can be interpreted to yield quite distinct claims. We have argued above that the first perspective, which considers the totality of arguments in each function, is more relevant given the nature of the research questions at hand, and in H&S all data are analyzed in this manner. Where cited sources have used the other perspective, we have either recalculated the figures, where possible, or excluded the data from this study.

APPENDIX A: LANGUAGES AND SOURCES FOR TABLE 2 IN H&S

LANGUAGE	SOURCE
Sakapultek	Du Bois 1987, table 2
English	Kärkkäinen 1996; total for each role taken from table 5; number of lexical vs. nonlexical mentions from Table 2.
English	Kumpf 2003, table 4
English	Everett 2009, table 4
English	Kumagai 2006, table 8
Portuguese	Everett 2009, table 4
Roviana	Corston-Oliver 2003, table 4
Korean	Clancy 2003; based on data from two children (Korean L1), recorded over a period of one year, commencing at age one year, eight months and one year, ten months, respectively; the percentages provided are taken from Clancy's figure 1 and represent the mean figures of the two children.
To'aba'ita	Lichtenberk 1996, table 8
Mapudungun	Arnold 2003, figures 2a–c; the figures given in H&S differ slightly from those given in Arnold's figure 3, because the latter includes pronouns under 'lexical' (Arnold 2003: 234). In the interests of consistency, pronouns have been removed from 'lexical' in our figures.
Yagua	Payne 1993, table 26; data stem from four folkloric narrative texts, three of which are reproduced in Payne 1993, altogether comprising 1,156 clauses (Payne 1993:57). For our purposes, 'lexical' includes any kind of mention involving an NP either by itself or in combination with some bound pronominal device (verbal prefix (VC = verb coding), enclitic (E), or head coding (HC)); zero coding of arguments is apparently only rarely attested in Yagua narratives, mostly only for Sp and P arguments (Payne 1993:29); they are not mentioned in table 26. We assume that the discrepancy between the absolute numbers of A and P arguments is due to uncounted zero arguments.
Gorani	Mahmoudveysi et al. 2012; two stories: 'Mard and Nāmard' (pp. 96–103) and 'Titila and Bibila' (pp. 89–95)
French	Ashby & Bentivoglio 1993, table 3; the authors distinguish between an 'X' role for the subject of the copula, and an 'S' role for other intransitive verbs. In our figures we have collapsed these two roles into the S-role.
Spanish	Ashby & Bentivoglio 1993; see details for French.
Vera'a	Schnell 2016; see Multi-CAST corpus description and annotation notes.
Teop	Mosel & Schnell 2016; see Multi-CAST corpus description and annotation notes.
Northern Kurdish	Haig & Thiele 2016; see Multi-CAST corpus description and annotation notes.
English	Schiborr 2016; see Multi-CAST corpus description and annotation notes.
Cypriot Greek	Hadjidas & Vollmer 2016; see Multi-CAST corpus description and annotation notes.

APPENDIX B: RAW DATA FROM THE MULTI-CAST DATA SET

VERA'A (Schnell 2016)

A (*n* = 795)				S (*n* = 2,026)				P (*n* = 905)				
[+hum]		[−hum]		[+hum]		[−hum]		[+hum]		[−hum]		
[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	TOTAL
89	655	26	25	367	1,330	171	158	107	162	473	163	3,726

traditional narratives; third person only

TEOP (Mosel & Schnell 2016)

A (*n* = 319)				S (*n* = 640)				P (*n* = 470)				
[+hum]		[−hum]		[+hum]		[−hum]		[+hum]		[−hum]		
[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	TOTAL
55	258	7	9	122	340	94	84	53	97	181	139	1,429

traditional narratives; third person only

NORTHERN KURDISH (Haig & Thiele 2016)

A (*n* = 277)				S (*n* = 527)				P (*n* = 396)				
[+hum]		[−hum]		[+hum]		[−hum]		[+hum]		[−hum]		
[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	TOTAL
40	223	6	8	87	275	120	45	22	53	210	111	1,200

traditional narratives; third person only

ENGLISH (Schiborr 2016)

A (*n* = 422)				S (*n* = 688)				P (*n* = 1,111)				
[+hum]		[−hum]		[+hum]		[−hum]		[+hum]		[−hum]		
[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	TOTAL
63	292	20	47	82	266	77	263	37	36	525	513	2,221

oral history interviews (interviewee only), traditional Kentish dialect; third person only

CYPRIOT GREEK (Hadjidas & Vollmer 2016)

A (*n* = 243)				S (*n* = 300)				P (*n* = 483)				
[+hum]		[−hum]		[+hum]		[−hum]		[+hum]		[−hum]		
[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	[+lex]	[−lex]	TOTAL
29	191	9	14	64	186	24	26	40	60	219	164	1,026

traditional narratives; third person only

REFERENCES

ANDREWS, AVERY. 1985. The major functions of the noun phrase. *Language typology and syntactic description, vol. 1: Clause structure*, ed. by Timothy Shopen, 62–154. Cambridge: Cambridge University Press.

ANDREWS, AVERY. 2007. The major functions of the noun phrase. *Language typology and syntactic description, vol. 1: Clause structure*, 2nd edn., ed. by Timothy Shopen, 132–223. Cambridge: Cambridge University Press.

ARNOLD, JENNIFER E. 2003. Multiple constraints on reference form: Null, pronominal, and full reference in Mapudungun. In Du Bois et al., 225–45.

ASHBY, WILLIAM, and PAOLA BENTIVOGLIO. 1993. Preferred argument structure in spoken French and Spanish. *Language Variation and Change* 5.61–76. DOI: 10.1017/S095439450000140X.

BICKEL, BALTHASAR. 2003. Referential density in discourse and syntactic typology. *Language* 79.708–36. DOI: 10.1353/lan.2003.0205.

BYBEE, JOAN L. 2001. Frequency effects on French liaison. *Frequency and the emergence of linguistic structure* (Typological studies in language 45), ed. by Joan L. Bybee and Paul Hopper, 337–59. Amsterdam: John Benjamins.

CHAFE, WALLACE L. (ed.) 1980. *The Pear stories: Cognitive, cultural and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

CHAFE, WALLACE L. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

CLANCY, PATRICIA M. 2003. The lexicon in interaction: Developmental origins of preferred argument structure in Korean. In Du Bois et al., 81–108.

COMRIE, BERNARD. 1981. *Language universals and linguistic typology: Syntax and morphology*. Oxford: Blackwell.

CORSTON-OLIVER, SIMON. 2003. Core arguments and the inversion of the nominal hierarchy in Roviana. In Du Bois et al., 273–300.

CYSOUW, MICHAEL, and BERNHARD WÄLCHLI. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF—Sprachtypologie und Universalienforschung* 60.2.95–99. DOI: 10.1524/stuf.2007.60.2.95.

DONOHUE, MARK. 2008. Semantic alignment systems. *The typology of semantic alignment*, ed. by Mark Donohue and Søren Wichmann, 24–75. Oxford: Oxford University Press.

DU BOIS, JOHN W. 1987. The discourse basis of ergativity. *Language* 63.4.805–55. DOI: 10.2307/415719.

DU BOIS, JOHN W. 2003. Argument structure: Grammar in use. In Du Bois et al., 11–60.

DU BOIS, JOHN W.; LORRAINE E. KUMPF; and WILLIAM J. ASHBY (eds.) 2003. *Preferred argument structure: Grammar as architecture for function*. Amsterdam: John Benjamins.

EVERETT, CALEB. 2009. A reconsideration of the motivations for preferred argument structure. *Studies in Language* 33.1.1–24. DOI: 10.1075/sl.33.1.02eve.

HADJIDAS, HARRIS, and MARIA VOLLMER. 2016. Cypriot Greek. In Haig & Schnell 2016. Online: https://lac.uni-koeln.de/en/multicast-cypriot-greek/.

HAIG, GEOFFREY; NICOLE NAU; STEFAN SCHNELL; and CLAUDIA WEGENER (eds.) 2011. *Documenting endangered languages: Achievements and perspectives*. Berlin: Mouton de Gruyter.

HAIG, GEOFFREY, and STEFAN SCHNELL. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators. Version 7.0. Online: https://lac.uni-koeln.de/multicast/.

HAIG, GEOFFREY, and STEFAN SCHNELL (eds.) 2016. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). Online: https://lac.uni-koeln.de/multicast/.

HAIG, GEOFFREY; STEFAN SCHNELL; and CLAUDIA WEGENER. 2011. Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Haig, Nau, et al., 55–86.

HAIG, GEOFFREY, and HANNA THIELE. 2016. Northern Kurdish. In Haig & Schnell 2016. Online: https://lac.uni-koeln.de/en/multicast-northern-kurdish/.

HASPELMATH, MARTIN. 2006. Review of Du Bois et al. *Language* 82.4.908–12. DOI: 10.1353/lan. 2006.0203.

HASPELMATH, MARTIN. 2011. On S, A, P, T, and R as comparative concepts for alignment typology. *Linguistic Typology* 15.3.535–67. DOI: 10.1515/LITY.2011.035.

HIMMELMANN, NIKOLAUS P. 1998. Documentary and descriptive linguistics. *Linguistics* 36.161–95. DOI: 10.1515/ling.1998.36.1.161.

KÄRKKÄINEN, ELISE. 1996. Preferred argument structure and subject role in American English conversational discourse. *Journal of Pragmatics* 25.675–701. DOI: 10.1016/0378-2166(95)00010-0.

KIBRIK, ANDREJ A. 2011. *Reference in discourse*. Oxford: Oxford University Press.

KUMAGAI, YOSHIHARU. 2006. Information management in intransitive subjects: Some implications for the preferred argument structure theory. *Journal of Pragmatics* 38.670–94. DOI: 10.1016/j.pragma. 2006.02.003.

KUMPF, LORRAINE E. 2003. Genre and preferred argument structure: Sources of argument structure in classroom discourse. In Du Bois et al., 109–30.

LICHTENBERK, FRANTIŠEK. 1996. Patterns of anaphora in To'aba'ita narrative discourse. *Studies in anaphora*, ed. by Barbara Fox, 379–411. Amsterdam: John Benjamins.

LYONS, JOHN. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

MAHMOUDVEYSI, PARWIN; DENISE BAILEY; LUDWIG PAUL; and GEOFFREY HAIG. 2012. *The Gorani language of Gawraju (Gawrajuyi), a village of West Iran: Texts, grammar and lexicon*. Wiesbaden: Reichert.

MOSEL, ULRIKE, and STEFAN SCHNELL. 2016. Teop. In Haig & Schnell 2016. Online: https://lac.uni-koeln.de/en/multicast-teop/.

NAESS, ASHILD. 2007. *Prototypical transitivity*. Amsterdam: John Benjamins.

NOONAN, MICHAEL. 2003. A crosslinguistic investigation of referential density. Milwaukee: University of Wisconsin–Milwaukee, MS.

PAYNE, DORIS. 1987. Information structuring in Papago narrative discourse. *Language* 63.783–804. DOI: 10.2307/415718.

PAYNE, THOMAS. 1993. *The twins stories: Participant coding in Yagua narrative*. Los Angeles: University of California Press.

PRINCE, ELLEN F. 1981. Toward a taxonomy of given/new information. *Radical pragmatics*, ed. by Peter Cole, 223–54. New York: Academic Press.

SCHIBORR, NILS NORMAN. 2016. English. In Haig & Schnell 2016. Online: https://lac.uni-koeln.de/en/multicast-english/.

SCHNELL, STEFAN. 2012. Data from language documentations in research on referential hierarchies. *Potentials of language documentation: Methods, analyses, and utilization*, ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, 64–72. Honolulu: University of Hawai'i Press. Online: http://nflrc.hawaii.edu/ldc/sp03/, and http://hdl.handle.net/10125/4518.

SCHNELL, STEFAN. 2016. Vera'a. In Haig & Schnell 2016. Online: https://lac.uni-koeln.de/en/multicast-veraa/.

SINGER, RUTH. 2011. Typologising idiomaticity: Noun-verb idioms and their relations. *Linguistic Typology* 15.625–59. DOI: 10.1515/LITY.2011.037.

STOLL, SABINE, and BALTHASAR BICKEL. 2009. How deep are differences in referential density? *Cross-linguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*, ed. by Elena Lieven Guo, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Seyda Özçaliskan, 543–55. London: Psychology Press.

THOMPSON, SANDRA A., and PAUL HOPPER. 2001. Transitivity, clause structure, and argument structure: Evidence from conversation. *Frequency and the emergence of linguistic structure* (Typological studies in language 45), ed. by Joan L. Bybee and Paul Hopper, 27–60. Amsterdam: John Benjamins.

VAN VALIN, ROBERT D., JR, and RANDY J. LAPOLLA. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.

WÄLCHLI, BERNHARD. 2006. Descriptive typology, or, the typologist's expanded toolkit. Konstanz: University of Konstanz, MS.

WÄLCHLI, BERNHARD. 2009. *Motion events in parallel texts: A study in primary data typology*. Bern: University of Bern Habilitation thesis.

[geoffrey.haig@uni-bamberg.de]
[stefan.schnell@unimelb.edu.au]