Multi-CAST

research context

Geoffrey Haig
University of Bamberg

Stefan Schnell
University of Melbourne

May 2018
v1.1

Language Archive Cologne

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Multi-CAST

research context

Geoffrey Haig
University of Bamberg

Stefan Schnell
University of Melbourne

May 2018
v1.1

Multi-CAST Multilingual Corpus of Annotated Spoken Texts

Citation for this document

Haig, Geoffrey & Schnell, Stefan. 2018[2016]. Multi-CAST research context. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts.* (archive handle) (date accessed)

Citation for the Multi-CAST collection

Haig, Geoffrey & Schnell, Stefan (eds.). 2018[2015]. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts.* (`https://lac.uni-koeln.de/en/multicast/`) (date accessed)

Archiving and versioning

This document has been archived at the *Language Archive Cologne* (LAC), accessible online at `lac.uni-koeln.de/en/multicast/`. The LAC is part of the Data Center for the Humanities (DCH) at the University of Cologne, Germany.

This is version 1.1 of the *Multi-CAST research context*, last updated 25 May 2018. The latest version is always available from the LAC.

This document was typeset with XeLaTeX and v2.0-10 of the *multicast2* class.

Contents

1 Introduction

The Multi-CAST collection (Haig & Schnell 2015) evolved out of collaborative research projects between 2009 and 2015, initially within the context of the documentation of endangered languages.[1] In the early phases, the main focus was on developing a system of syntactic annotations that would be sufficiently flexible to be applicable to spoken language corpora from typologically diverse languages, while also being sufficiently consistent to enable meaningful cross-corpus comparisons.

The resulting system, GRAID (Grammatical Relations and Animacy in Discourse, Haig & Schnell 2014), now in version 7.0, provides the basic foundation for the annotation of each of the corpora in Multi-CAST, and the main framework for comparative quantitative analysis. With the increasing diversification of the research questions that are being addressed, we have since developed additional annotation tiers, leading to richer annotations and an overall deeper archive structure.

Given the ongoing nature of the Multi-CAST research agenda, the present paper only deals with some of the initial – although no less relevant – research questions that have motivated the architecture and design philosophy of the annotations and the collection as a whole. For more detailed information on research based on Multi-CAST, both published and in progress, we refer to the *Research and publications* section of the Multi-CAST archive webpage.[2]

Much of the inspiration for Multi-CAST can be traced to the research tradition pioneered by Wallace Chafe and associates, which targetted natural spoken language and formulated functional explanations for the observed regularities. From the outset, this line of research was deeply informed by its cross-linguistic focus, and early studies were often based on natural language corpora from poorly-described languages (see, for instance, Du Bois 1987). In a sense, the research based on Multi-CAST originated as an attempt to harness the technical advances in corpus and documentary linguistics to the research agenda of the Chafe'ian paradigm.

The main focus of this research has been on understanding the choices speakers exercise when verbalizing a referent in a particular discourse context. Typically, there exists a choice between a lexical noun phrase (e.g. *the girl*), a pronoun (*she, her*), or zero. These choices are dependent on a variety of factors, which include local syntactic constraints (e.g. binding principles), the information status of the referents concerned (identifiability, accessibility, topicality, etc.), language-specific typological constraints (e.g. different degrees of tolerance of null-anaphora), and many more. A considerable body of literature addresses the interaction of these issues: see, among many others,

1 See `http://dobes.mpi.nl/`.

2 Online at `https://lac.uni-koeln.de/en/multi-cast-research-and-publications/`.

Chafe (1976; 1994), Prince (1981), Givón (1983), Ariel (1990) and 2000, Bickel (2003), Noonan (2003), Huang (2000), Holmberg (2009), and Du Bois (1987; 2003a). Systematic cross-linguistic studies of discourse are still a rarity however, and it is with this in mind that the texts in Multi-CAST were compiled and annotated.

In natural spoken discourse, a fair amount of work in effecting reference is actually achieved through covert, or zero, expression types, with the proportion varying from language to language. Any serious investigation of speakers' choices must thus take into account zero expressions, which conventional morphological glossing and part-of-speech tagging fail to register. In GRAID, zero expressions are methodically noted in the annotation, thereby "levelling the ground" between different languages and allowing for systematic cross-linguistic investigation of argument realization in discourse.

In the following, we outline two avenues of research which have figured prominently in Multi-CAST-related inquiries, and on which we continue to build in ongoing research.

2 Research outlines

2.1 The 'discourse basis of ergativity' and related issues

The best-known association between syntactic relation and information status is that of subjects with given information, and hence with reduced (pronominal or zero) expression. It has furthermore been claimed that direct objects (P) and the subjects of intransitive clauses (S) are typically associated with new information and hence with expression as lexical noun phrases (Du Bois 1987; 2003a; b). This grouping of the S and P roles has been referred to as the "discourse basis of ergativity". However, the claimed unity of S and P has proved rather elusive, and up until now, more representative cross-linguistic data have not been available.

It is a straightforward matter to extract the levels of lexical expression for A (subjects of transitive clauses), S (subjects of intransitive clauses), and P (direct objects) from the Multi-CAST data. The results are shown in Table 1 and Figure 1. As can be seen, there exists little evidence for the claimed unity of S and P. The implication of these findings are discussed in detail in Haig & Schnell (2016).

2.2 The zero-expression of different argument types

2.2.1 Referential null subjects and objects

Since the 1970's it has been assumed that there are significant cross-linguistic differences in the extent to which languages tolerate clauses without overt subjects. While earlier literature referred to a binary 'pro-drop parameter',

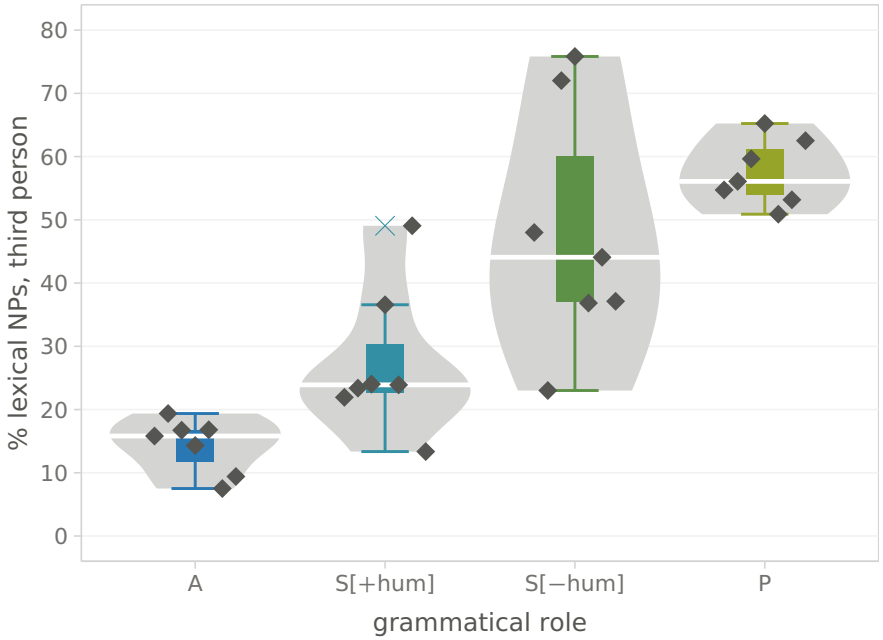

Figure 1 Lexicality of arguments by grammatical role and humanness in the Multi-CAST collection, third person only. Each dot represents a corpus, in the same order left-to-right as in Table 1 below, top-to-bottom.

	A			S[+hum]			S[−hum]			P		
corpus	lex	all	%	lex	all	%	lex	all	%	lex	all	%
C. Greek	37	234	16	50	228	22	24	50	48	255	466	55
English	80	413	19	80	342	23	78	339	23	563	1 004	56
N. Kurdish	46	275	17	86	358	24	121	168	72	232	389	60
Persian	82	573	14	181	495	37	160	211	76	332	509	65
Teop	62	369	17	120	502	24	63	171	37	228	448	51
Tondano	37	493	8	26	53	49	93	211	44	285	536	53
Vera'a	74	786	9	208	1 558	13	134	361	37	492	787	63

Table 1 Proportions of lexical expression of third person core arguments in the Multi-CAST collection.
A = subject of a transitive clause,
S = subject of an intransitive clause,
P = direct object.

	subjects (A+S)			objects (P)		
corpus	null	all	%	null	all	%
C. Greek	752	934	81	77	495	16
English	256	2 023	13	39	1 013	4
N. Kurdish	542	1 069	51	137	417	33
Persian	778	1 311	59	92	509	18
Teop	375	1 276	29	97	447	22
Tondano	455	885	51	217	543	40
Vera'a	916	4 092	22	192	919	21

Table 2 Proportions of referential null subjects and objects in the Multi-CAST collection.

dividing languages into pro-drop and non-pro-drop classes, attempts at more refined typologies have since been developed in the realm of referential null subjects (e.g. Roberts & Holmberg 2009).

Relevant data on this topic can be readily extracted from Multi-CAST. Table 2 provides the percentages of zeroes for A (subjects of transitive clauses), S (subjects of intransitive clauses), and P (direct objects). Combining A and S to yield a (reasonably coherent) notion of "subject", the degree of tolerance of referential null subjects in each language is shown in Figure 2 together with the degree of tolerance of zero objects, a hitherto largely neglected typological parameter.

2.2.2 Referential density

Related to the notion of referential null, although stemming from a different research tradition, is the theory of referential density (RD). Referential density is defined as the ratio of overtly expressed arguments to the possible (i.e. notional) arguments in a given texts (Bickel 2003; Noonan 2003):

(1) $$RD = \frac{n\,(\text{overt arguments})}{n\,(\text{available argument positions})}$$

In Bickel (2003), referential density is explicitly defined to include non-core arguments, but excluding adjuncts. However, we have found that in practice, implementing the core–non-core distinction is fraught with considerable difficulty. In calculating the referential density values for the Multi-CAST corpora as given in Table 3, we have decided to consider only A (subjects of transitive clauses), S (subjects of intransitive clauses), and P (direct objects), essentially yielding an aggregation of the same measures as in Table 2. The rationale

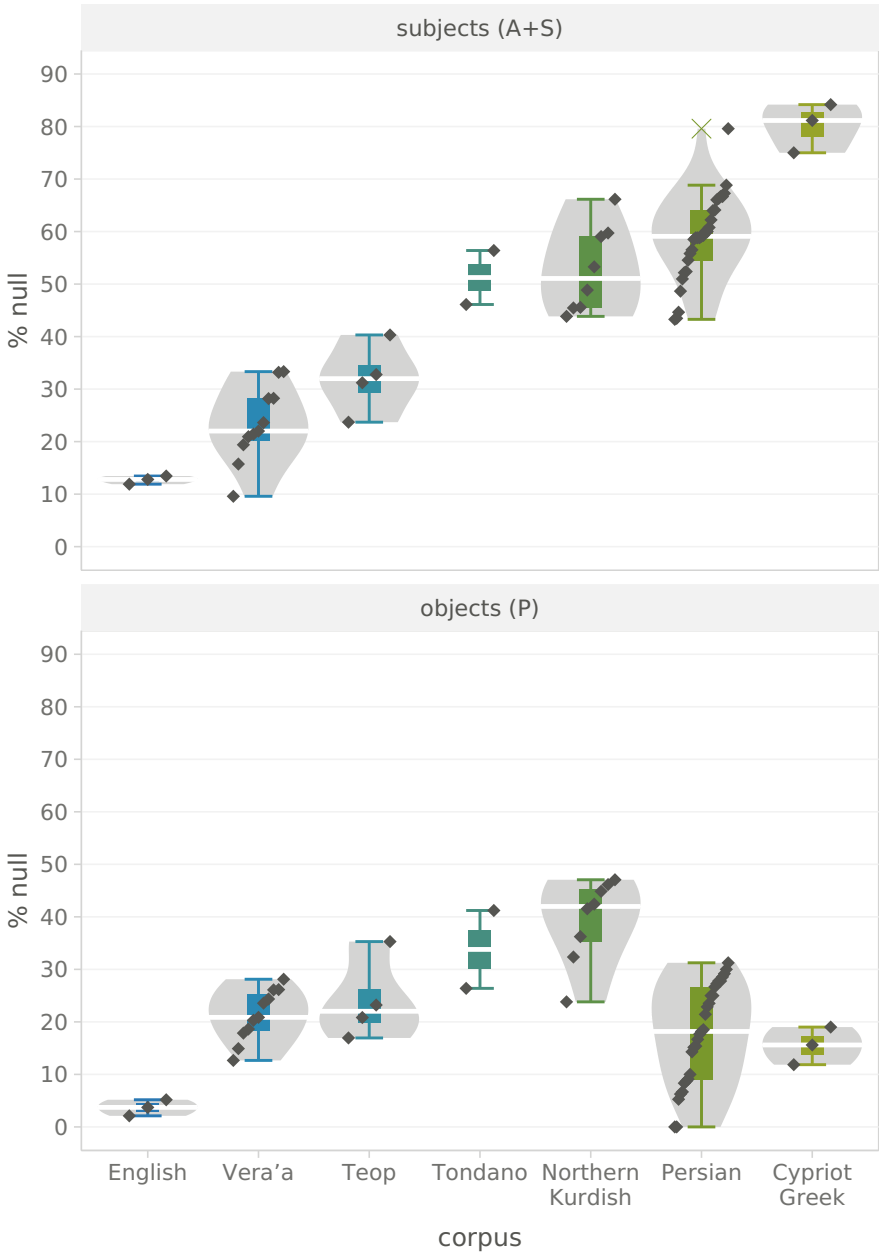

Figure 2 Distribution of referential null subjects and objects in the Multi-CAST collection. Each dot represents a corpus text.

corpus	overt args	all args	RD
C. Greek	600	1 429	0.42
English	2 782	3 078	0.90
N. Kurdish	807	1 486	0.54
Persian	950	1 820	0.52
Teop	1 267	1 740	0.73
Tondano	756	1 428	0.53
Vera'a	3 903	5 011	0.78

Table 3 Core referential density (RD) of the texts in the Multi-CAST collection.

behind this decision and a discussion of RD in the context of extra-linguistic factors is given in a pilot study, Haig & Adibifar (2016).

2.3 Further lines of inquiry

Other research approaches which could be investigated with the help of Multi-CAST include, among many others:

- What is the impact of syntactic function on the form of a referential expression (cf. Du Bois's 1987 "avoid lexical A", or Chafe's 1994 "light subject constraint")?
- What is the preferred configuration (syntactic function, type of clausal construction, etc.) for the introduction of new referents into discourse?
- What impact does the factor of humanness have on the formal expression of discourse participants in different syntactic functions and/or constructions?
- What is its profile (or "depth" in terms of Stoll & Bickel 2009) of a text or language, that is, the proportion of lexical to pronominal forms of reference?
- What (if any) is the impact of object realization on subject realization (and vice versa)?
- Under which conditions do referential null objects occur?
- What are the conditions and pathways of the grammaticalization of argument-predicate agreement?
- What types of syntactic constructions prevail in different text varieties across different languages (cf. Biber 1995)?

References

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. 2000. The development of person agreement markers: From pronouns to higher accessibility markers. In Barlow, Michael & Kemmer, Suzanne (eds.), *Usage-based models of language*, 197–220. Stanford: Center for the Study of Language and Information.

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

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

Bybee, Joan & Hopper, Paul (eds.). 2001. *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45). Amsterdam: John Benjamins.

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

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

Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63(4). 805–855.

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

Du Bois, John. 2003b. Discourse and grammar. In Tomasello, Michael (ed.), *The new psychology of language: Cognitive and functional approaches to language structure*, vol. 2, 47–88. Mahwah, NJ: Erlbaum.

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

Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307.

Haig, Geoffrey & Adibifar, Shirin. 2016. Does addresse familiarity impact on referential density? Evidence from spoken Persian, and implications for language typology. Unpublished manuscript. University of Bamberg.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (`https://lac.uni-koeln.de/en/multicast/`) (Accessed 2015-12-30).

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (`https://lac.uni-koeln.de/multicast/`) (Accessed 2016-02-08).

Haig, Geoffrey & Schnell, Stefan. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618.

Holmberg, Anders. 2009. Null subject parameters. In Biberauer, Theresa & Holmberg, Anders & Roberts, Ian & Sheehan, Michelle (eds.), *Parametric variation: Null subjects in minimalist theory*, 88–124. Cambridge: Cambridge University Press.

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

Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.

Li, Charles N. & Thompson, Sandra A. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. In Givón, Talmy (ed.), *Discourse and syntax*, vol. 12, 311–335. New York: Academic Press.

Lichtenberk, František. 1996. Patterns of anaphora in To'aba'ita narrative discourse. In Fox, Barbara (ed.), *Studies in anaphora*, 379–411. Amsterdam: John Benjamins.

Meyerhoff, Miriam. 2000. The emergence of creole subject-verb agreement and the licensing of null subjects. *Language Variation and Change* 12(2). 203–230.

Noonan, Michael. 2003. *A crosslinguistic investigation of referential density*. Milwaukee: University of Wisconsin-Milwaukee. (`http://crossasia-repository.ub.uni-heidelberg.de/190/`) (Accessed 2016-02-08).

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

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

Stoll, Sabine & Bickel, Balthasar. 2009. How deep are differences in referential density? In Guo, Jiansheng & Lieven, Elena & Budwig, Nancy & Ervin-Tripp, Susan & Nakamura, Keiko & Özçaliskan, Seyda (eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of dan isaac slobin*, 543–555. London: Psychology Press.

Wälchli, Bernhard. 2006. *Descriptive typology, or: The typologist's expanded toolkit*. Stockholm: Stockholm University.

Wälchli, Bernhard. 2009. *Motion events in parallel texts: A study in primary data typology*. Bern: University of Bern Unpublished Habilitationsschrift.

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

lac.uni-koeln.de/multicast/