Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

Research context

June 2016

Geoffrey Haig
University of Bamberg

Stefan Schnell
University of Melbourne

cite this document as:

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

file v1.0

Introduction

The Multi-CAST collection 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 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 (Du Bois 1987a). 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, Chafe (1976, 1994), Prince (1981), Givón (1983), Ariel (1990, 2000), Bickel (2003), Noonan (2003), Huang (2000), Holmberg (2009), and Du Bois (1987a, 2003a).

[1] `http://dobes.mpi.nl/`

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

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.

Testing the proposed correlation of syntactic relations with information status: 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 1987a, 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 to appear).

The frequency of zero-expression of different argument types: non-referential subjects, referential density, and related issues

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

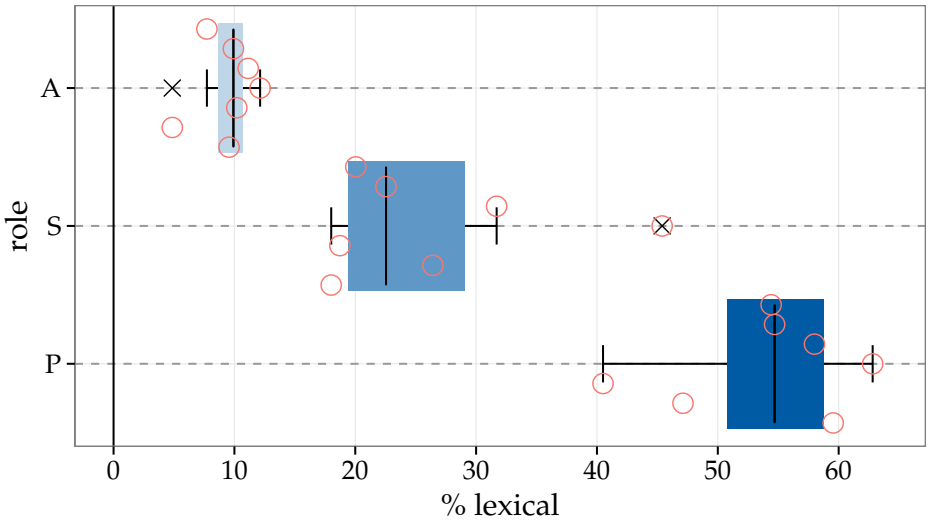

Figure 1. Distribution of lexical arguments by role across the Multi-CAST corpora. The value of each corpus is indicated by a red circle, following the same order from top to bottom as in Table 1.

	A			S			P		
corpus	**all**	**[+lex]**	**%**	**all**	**[+lex]**	**%**	**all**	**[+lex]**	**%**
cypgreek	492	38	*8*	439	88	*20*	476	259	*54*
english	2986	289	*10*	4085	921	*23*	2914	1594	*55*
nkurd	413	46	*11*	653	207	*32*	400	232	*58*
persian	602	73	*12*	760	345	*45*	519	326	*63*
teop	461	47	*10*	769	144	*19*	437	177	*41*
tondano	782	38	*5*	439	116	*26*	590	278	*47*
veraa	1203	115	*10*	2987	538	*18*	974	580	*60*

Table 1. Lexical core arguments in the Multi-CAST corpora.

Relevant data on this topic can be extracted readily 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.

Related to this, although stemming from a different research tradition, is the notion 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):

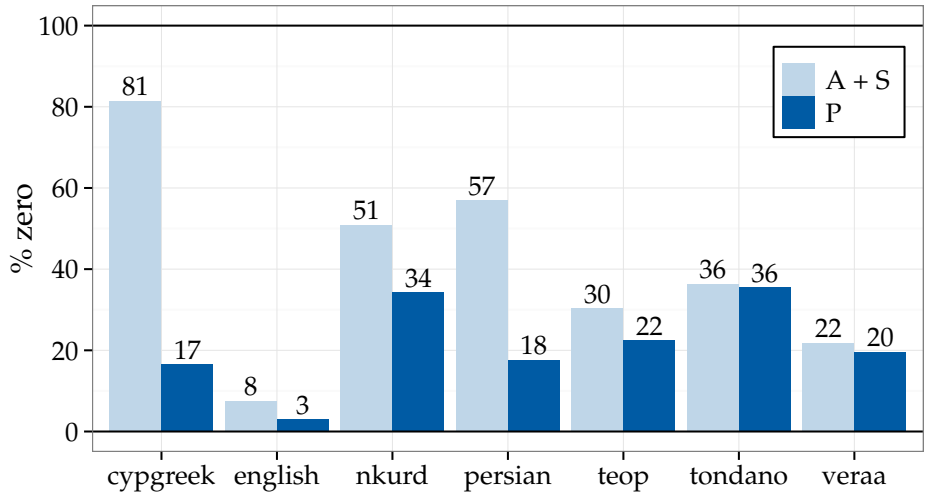

Figure 2. Referential null subjects and objects in the Multi-CAST corpora.

	A			S			P		
corpus	**all**	**zero**	**%**	**all**	**zero**	**%**	**all**	**zero**	**%**
cypgreek	492	434	*88*	439	324	*74*	476	79	*17*
english	2986	331	*11*	4085	197	*5*	2914	87	*3*
nkurd	413	245	*59*	653	298	*46*	400	137	*34*
persian	602	465	*77*	760	311	*41*	519	92	*18*
teop	461	109	*24*	769	264	*34*	437	98	*22*
tondano	782	323	*41*	439	122	*28*	590	210	*36*
veraa	1203	300	*25*	2987	616	*21*	974	192	*20*

Table 2. Referential null arguments in the Multi-CAST corpora.

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

In (Bickel 2003), referential density is explicitly defined to include non-core arguments, though excluding adjuncts. However, we have found that in practice, implementing the argument–adjunct distinction is fraught with considerable difficulty; in calculating the referential density values for the Multi-CAST corpora as given in Table 3, we have instead 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 behind this decision, and the discussion of RD with extra-linguistic factors, is provided in a pilot study, Haig & Adibifar (Submitted).

corpus	all args	overt args	RD
cypgreek	1407	570	*0.41*
english	9913	9298	*0.94*
nkurd	1466	786	*0.54*
persian	1881	1013	*0.54*
teop	1667	1196	*0.72*
tondano	1811	1156	*0.64*
veraa	5164	1108	*0.79*

Table 3. Referential density across texts in the Multi-CAST corpora.

Further questions

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

- What is the impact of syntactic function on the form of a referential expression (cf. Du Bois's 1987a '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 (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 grammaticalisation 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. [p. 1]

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. [p. 1]

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. [p. 5]

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736. [pp. 1, 3, 4]

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. [p. 1]

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press. [pp. 1, 5]

Du Bois, John. 1987a. Absolutive zero: Paradigm adaptivity in Sacapultec Maya. *Lingua* 71(2). 203–222. [pp. 1, 2, 5]

Du Bois, John. 1987b. 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*, 11–60. Amsterdam: John Benjamins. [pp. 1, 2]

Du Bois, John. 2003b. Discourse and grammar. In Tomasello, Michael (ed.), *The psychology of language*, vol. 2, 47–88. Mahwah, NJ: Erlbaum. [p. 2]

Givón, Talmy (ed.). 1983. *Topic continuity in discourse* (Typological Studies in Language 3). Amsterdam: John Benjamins. [p. 1]

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, Širin. Submitted. Does addresse familiarity impact on referential density? Evidence from spoken Persian, and implications for language typology . [p. 4]

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotat-*

ors (version 7.0). (https://lac.uni-koeln.de/en/multicast/) (accessed 2015-12-30.) [p. 1]

Haig, Geoffrey & Schnell, Stefan. to appear. The discourse basis of ergativity revisited . [p. 2]

Holmberg, Anders. 2009. Null subject parameters. In Biberauer, Theresa & Holmberg, Anders & Roberts, Ian & Sheehan, Michelle (eds.), *Parametric variation*, 88–124. Cambridge: Cambridge University Press. [p. 1]

Huang, Yan. 2000. *Anaphora: A cross-linguistic study*. Oxford: Oxford University Press. [p. 1]

Kibrik, Andrej. 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.). (Unpublished manuscript.) [pp. 1, 3]

Prince, Ellen. 1981. Toward a taxonomy of given-new information. In Cole, Peter (ed.), *Radical pragmatics*, 223–255. New York: Academic Press. [p. 1]

Roberts, Ian & Holmberg, Anders. 2009. Introduction: Parameters in minimalist theory. In Biberauer, Theresa & Holmberg, Anders & Roberts, Ian & Sheehan, Michelle (eds.), *Parametric variation*, 1–57. Cambridge: Cambridge University Press. [p. 2]

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

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