Multi-CAST
research context

Geoffrey Haig
University of Bamberg

Stefan Schnell
University of Melbourne

May 2018
v1.1
## Contents

1. **Introduction**  
2. **Research outlines**  
   2.1 The ‘discourse basis of ergativity’ and related issues  
   2.2 The zero-expression of different argument types  
      2.2.1 Referential null subjects and objects  
      2.2.2 Referential density  
   2.3 Further lines of inquiry  
3. **References**

1 Introduction 2
2 Research outlines 2
   2.1 The ‘discourse basis of ergativity’ and related issues 2
   2.2 The zero-expression of different argument types 2
      2.2.1 Referential null subjects and objects 2
      2.2.2 Referential density 4
   2.3 Further lines of inquiry 6
3 References 7
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. 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.

Much of the inspiration for Multi-CAST can be traced to the research tradition pioneered by Wallace Chafe and associates, which targeted 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/.
Research context

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.

Table 1  Proportions of lexical expression of third person core arguments in the Multi-CAST collection.

<table>
<thead>
<tr>
<th>corpus</th>
<th>A lex</th>
<th>A all</th>
<th>A %</th>
<th>S[+hum] lex</th>
<th>S[+hum] all</th>
<th>S[+hum] %</th>
<th>S[−hum] lex</th>
<th>S[−hum] all</th>
<th>S[−hum] %</th>
<th>P lex</th>
<th>P all</th>
<th>P %</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Greek</td>
<td>37</td>
<td>234</td>
<td>16</td>
<td>50</td>
<td>228</td>
<td>22</td>
<td>48</td>
<td>566</td>
<td>55</td>
<td>255</td>
<td>466</td>
<td>55</td>
</tr>
<tr>
<td>English</td>
<td>80</td>
<td>413</td>
<td>19</td>
<td>80</td>
<td>342</td>
<td>23</td>
<td>78</td>
<td>339</td>
<td>23</td>
<td>563</td>
<td>1004</td>
<td>56</td>
</tr>
<tr>
<td>N. Kurdish</td>
<td>46</td>
<td>275</td>
<td>17</td>
<td>86</td>
<td>358</td>
<td>24</td>
<td>121</td>
<td>168</td>
<td>72</td>
<td>232</td>
<td>389</td>
<td>60</td>
</tr>
<tr>
<td>Persian</td>
<td>82</td>
<td>573</td>
<td>14</td>
<td>181</td>
<td>495</td>
<td>37</td>
<td>160</td>
<td>211</td>
<td>76</td>
<td>332</td>
<td>509</td>
<td>65</td>
</tr>
<tr>
<td>Teop</td>
<td>62</td>
<td>369</td>
<td>17</td>
<td>120</td>
<td>502</td>
<td>24</td>
<td>63</td>
<td>171</td>
<td>37</td>
<td>228</td>
<td>448</td>
<td>51</td>
</tr>
<tr>
<td>Tondano</td>
<td>37</td>
<td>493</td>
<td>8</td>
<td>26</td>
<td>53</td>
<td>49</td>
<td>93</td>
<td>211</td>
<td>44</td>
<td>285</td>
<td>536</td>
<td>53</td>
</tr>
<tr>
<td>Vera’a</td>
<td>74</td>
<td>786</td>
<td>9</td>
<td>208</td>
<td>1558</td>
<td>13</td>
<td>134</td>
<td>361</td>
<td>37</td>
<td>492</td>
<td>787</td>
<td>63</td>
</tr>
</tbody>
</table>
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):

\[ RD = \frac{n \text{ (overt arguments)}}{n \text{ (available argument positions)}} \]

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.
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


