Multi-CAST
Multilingual Corpus of Annotated Spoken Texts

Geoffrey Haig & Nils Schiborr
8 June 2016
Department of General Linguistics
University of Bamberg
Overview

1. **Research context:**
   the probabilistic grammar of discourse

2. **Multi-CAST:**
   content and design

3. Syntactic annotations

4. Analysis procedures

5. Case studies
Research context

- corpus-based language typology: how are the resources of grammar deployed in connected spoken language?
Language typology

- investigates range and limits of variation in human language
- draws on samples of genetically and areally diverse languages
- traditionally compares grammars, yielding categorical feature values and correlations
Corpus-based approaches

- compares **corpora** rather than grammars
- requires **consistent annotation schemes**
- yields **probabilistic assessments** rather than categorical statements
Example: Conventional

- conventional typology: pronoun deletion

**ENGLISH**

*I work here.*

**SPANISH**

___ *trabajo aqui.*

**JAPANESE**

___ *koko de hataraitte iru.*
Example: Conventional

- conventional typology categorizes languages:

  English = not pro-drop,
  Spanish = pro-drop,
  Japanese = radical/discourse pro-drop
  etc. (cf. Holmberg 2009)
Example: Corpus-based

- corpus-based approach to pronoun deletion:
- characterizes texts, rather than languages

(Bickel 2003; Noonan 2003; Haig & Adibifar, under review)
Example: Corpus-based
(Bickel 2003; Noonan 2003; Kumagai 2006; Haig & Adibifar, submitted)
Research background

- typologically informed, functionalist syntax
- **Wallace Chafe, Talmy Givón**, and associates in the 1970’s and 1980’s
Research background

- applies more **sophisticated methodologies** from corpus linguistics, variationist linguistics, and statistical analysis
Research focus

- information management in discourse
  how is *new information* introduced?
  how are *referents* tracked in grammar?
  how ‘persistent’ are pronouns?
  how does *syntactic function* relate to *information status*?
Multi-CAST
Multilingual Corpus of Annotated Spoken Texts
Collaboration

- collaborative research
  Bamberg / Melbourne / Cologne

- principal investigators
  Geoffrey Haig (U of Bamberg)
  Stefan Schnell (U of Melbourne)
Collaboration

- supported by
  ARC-DECRA (Schnell),
  CoEDL (Thieberger & Schnell),
  U of Bamberg department funding (Haig);
  further funding applied for
Collaboration

- supported by

CLARIN F-AG-3 (Felix Rau)

hosted at the Language Archive Cologne
lac.uni-koeln.de/multicast/
Purpose and scope

- research tool for corpus-based typology
- includes sub-corpora from seven languages
- ongoing expansion; aim: 20–25 languages
- corpus description in Schiborr 2016a
  
  http://bit.ly/1Wz1pg6
Corpus design

- typologically diverse languages
- natural spoken language
- annotated on multiple levels
Corpus design

- typical data set: folkloric texts, rec’d in situ
- minimum of 1000 clause units per language
- consistent corpus structure
- consistent annotation scheme
Open Science

- explicit documentation
- unrestricted access
  all data released under
  a Creative Commons license
  (BY-NC-SA 4.0)
Compilation

- uses **CLARIN**-developed tools:
  - **ELAN** as an annotation platform
    (EUDICO Linguistic Annotator, MPI Nijmegen)
  - **IMDI** for metadata definitions
    (ISLE Metadata Initiative)
# The collection

<table>
<thead>
<tr>
<th>language</th>
<th>affiliation</th>
<th>clause units</th>
<th>annotators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyp. Greek</td>
<td>Greek</td>
<td>1,071</td>
<td>Hadjidas &amp; Vollmer 2016</td>
</tr>
<tr>
<td>English</td>
<td>Germanic</td>
<td>7,278</td>
<td>Schiborr 2016b</td>
</tr>
<tr>
<td>N. Kurdish</td>
<td>Iranian</td>
<td>1,101</td>
<td>Haig &amp; Thiele 2016</td>
</tr>
<tr>
<td>Persian</td>
<td>Iranian</td>
<td>1,421</td>
<td>Adibifar 2016</td>
</tr>
<tr>
<td>Teop</td>
<td>Oceanic</td>
<td>1,302</td>
<td>Mosel &amp; Schnell 2016</td>
</tr>
<tr>
<td>Tondano</td>
<td>Philippine</td>
<td>1,086</td>
<td>Brickell 2016</td>
</tr>
<tr>
<td>Vera’a</td>
<td>Oceanic</td>
<td>3,606</td>
<td>Schnell 2016</td>
</tr>
</tbody>
</table>

*collection totals* 16,965 Haig & Schnell 2016
Future additions

- in the pipeline:
  - Burmese, Japanese, Mawng, Mandarin
Syntactic annotation

1. (audio recording)
2. utterance unit
3. translation
4. grammatical words
   + morphological glossing
4. GRAID
   Grammatical Relations and Animacy in Discourse,
   (Haig & Schnell 2014)
Syntactic annotation

- in the pipeline:
  - referent indexing (RefIND, Schiborr et al. 2016)
  - semantic predicate types
Analysis

- **ELAN**: multi-tier export of annotations
- **complex structural and statistical analysis**
  using statistics software (R, SPSS)
  and text mining tools
Case study

- The Discourse Basis of Ergativity Revisited
  (Haig & Schnell, to appear)

- re-examines widely accepted claims in typology
Discourse Ergativity

- **basic idea:**
  syntactic function ↔ information status

- transitive subjects (A)
  = favoured position for **given information**

- intransitive subjects (S), direct objects (P)
  = favoured position for **new information**
Discourse Ergativity

thus:

- transitive subjects (A) = mostly *pronouns/zero*, few lexical NPs
- intransitive subjects (S), direct objects (P) = few pronouns/zero, many *lexical NPs*
Discourse Ergativity

- S+P: shared information-structure profile, in opposition to A

- Du Bois’ claim: unity of S+P mirrors ergative alignment

- based on a corpus of spoken Sakapultek (Mayan)
Discourse Ergativity?

- doubts have been raised
  (Haspelmath 2006, Everett 2009)
- but claims not **representatively** tested,
  i.e. using a **larger data base and standardized testing procedures**
Discourse Ergativity?

- we tested the claims against Multi-CAST (+ 14 other corpora)

- result:
  no significant grouping of $S+P$
No Discourse Ergativity
(Haig & Schnell 2016; Haig & Schnell, to appear)
Publications & Resources

- GRAID Manual 7.0  
  (Haig & Schnell 2014)
- Corpus overview  
  (Schiborr 2016a)
- Haig & Schnell (to appear)
- Brickell & Schnell (accepted)
- Haig & Adibifar (under review)
- RefIND Manual  
  (Schiborr et al. 2016)
References (1/6)


References (2/6)


References (3/6)


Haig, Geoffrey & Schnell, Stefan. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators, ver. 7.0. (https://lac.uni-koeln.de/multicast/) (accessed 2016-06-08.)


References (4/6)


References (5/6)


Thank you!