Multi-CAST
Multilingual Corpus of Annotated Spoken Texts

Geoffrey Haig & Nils Schiborr
Department of General Linguistics
University of Bamberg

8 June 2016

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 hataraite iru.*

Example: Conventional

- conventional typology **categorizes** languages:

  English = not pro-drop,

  Spanish = pro-drop,

  Japanese = radical/discourse pro-drop

  etc. (cf. Holmberg 2009)

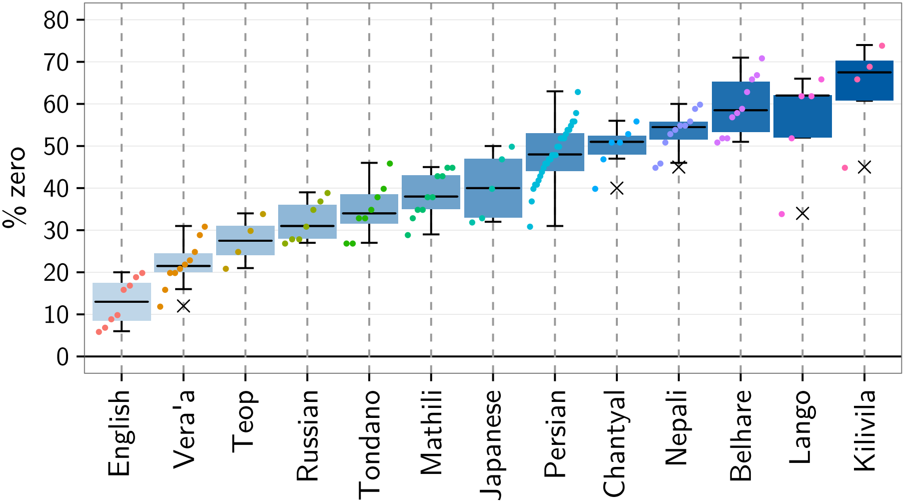
Example: Corpus-based

- **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/

LAC Language Archive Cologne

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)

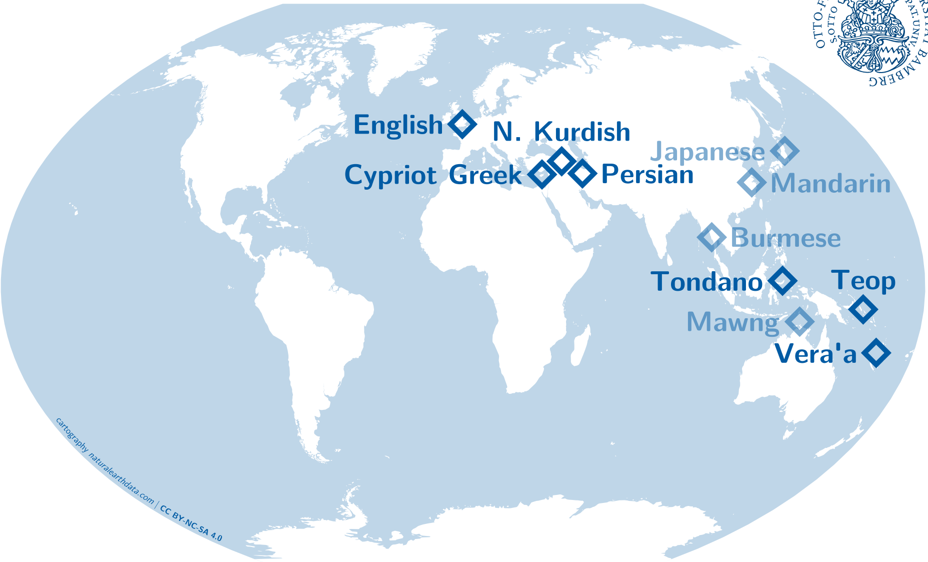
  IMDI for metadata definitions
  (ISLE Metadata Initiative)

The collection

		clause	
language	**affiliation**	**units**	**annotators**
Cyp. Greek	Greek	1,071	Hadjidas & Vollmer 2016
English	Germanic	7,278	Schiborr 2016b
N. Kurdish	Iranian	1,101	Haig & Thiele 2016
Persian	Iranian	1,421	Adibifar 2016
Teop	Oceanic	1,302	Mosel & Schnell 2016
Tondano	Philippine	1,086	Brickell 2016
Vera'a	Oceanic	3,606	Schnell 2016
collection totals		16,965	*Haig & Schnell 2016*

Future additions

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


English
N. Kurdish
Cypriot Greek
Persian
Japanese
Mandarin
Burmese
Tondano
Teop
Mawng
Vera'a
cartography naturalearthdata.com | CC BY-NC-SA 4.0

Syntactic annotation

-. (audio recording)

1. utterance unit
2. translation
3. 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
  (Du Bois 1987, 2003a, 2003b, 2006)

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?

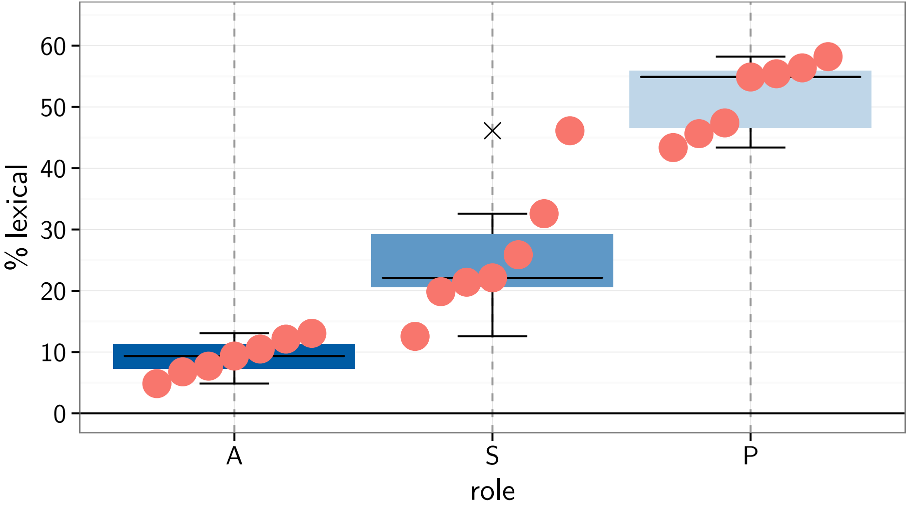
- we tested the claims against **Multi-CAST** (+ 14 other corpora)

- **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)

Adibifar, Širin. 2016. Persian. In Haig, Geoffrey & Schnell, Stefan(eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-persian`) (accessed 2016-06-08.)

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

Brickell, Timothy C. 2016. Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-tondano`) (accessed 2016-06-08.)

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.

References (2/6)

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

Du Bois, John. 2006. The Pear Story in Sakapultek Maya: A case study of information flow and preferred argument structure. In Sedano, Mercedes & Bolivar, Adriana & Shiro, Martha (eds.), *Haciendo Lingüística: Homenaje a Paola Bentivoglio* [Doing linguistics: A tribute to Paola Bentivoglio], 191–221. Caracas: Universidad Central de Venezuela.

Everett, Caleb. 2009. A reconsideration of the motivations for preferred argument structure. Studies in Language 33(1): 1–24.

Hadjidas, Harris & Vollmer, Maria C. 2016. Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-cypriot-greek`) (accessed 2016-06-08.)

References (3/6)

Haig, Geoffrey & Adibifar, Širin. Under review. Does addresse familiarity impact on referential density? Evidence from spoken Persian, and implications for language typology.

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.)

Haig, Geoffrey & Schnell, Stefan (eds.). 2016. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast/`) (accessed 2016-05-11.)

Haig, Geoffrey & Schnell, Stefan. To appear. The discourse basis of ergativity revisited.

Haig, Geoffrey & Thiele, Hanna. 2016. Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-northern-kurdish`) (accessed 2016-06-08.)

References (4/6)

Haspelmath, Martin. 2006. Review of *Preferred argument structure: Grammar as architecture for function*, by John Du Bois, Lorraine Kumpf, and William Ashby. Language 82(4): 908–912.

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.

Kumagai, Yoshiharu. 2006. Information management in intransitive subjects: Some implications for the preferred argument structure theory. Journal of Pragmatics 38(5): 670–694.

Mosel, Ulrike & Schnell, Stefan. 2016. Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-teop`) (accessed 2016-06-08.)

References (5/6)

Noonan, Michael. 2003. A crosslinguistic investigation of referential density. (Unpublished manuscript.) (`http://crossasia-repository.ub.uni-heidelberg.de/190/`) (accessed 2016-06-08.)

Schiborr, Nils N. 2016a. Multi-CAST corpus overview and description. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://www.uni-bamberg.de/fileadmin/aspra/Multi-CAST_corpus-overview.pdf`) (accessed 2016-06-08.)

Schiborr, Nils N. 2016b. English. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-english`) (accessed 2016-06-08.)

Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2016. RefIND (Referent Indexing in Natural-language Discourse): Annotation guidelines, ver. 1.0. Bamberg/Melbourne: University of Bamberg/University of Melbourne. (Unpublished Manuscript.)

References (6/6)

Schnell, Stefan. 2016. Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (`https://lac.uni-koeln.de/multicast-veraa`) (accessed 2016-06-08.)

Multi-
CAST

Multilingual Corpus of
Annotated Spoken Texts

Thank you!