WS Annotation of non-standard corpora

Modelling referential choice in natural spoken discourse

Multi-CAST, GRAID, and RefIND

Nils Norman Schiborr
University of Bamberg

16 September 2019
v1.1

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/

Geoffrey Haig Stefan Schnell Nils Schiborr

Multi-CAST: An overview

- **spoken**, **non-elicited**, **non-translated language**
- chiefly monologic, various narrative genres (folktales etc.)

- **11 corpora from typologically diverse languages**
- each corpus contains at least 1 000 clause units
- 20 000 clause units in total (c. 85 000 words)
- 10 additional corpora in preparation

- multiple layers of **standardized annotation** for morphosyntax and referent tracking
- designed as a tool for **quantitative**, **corpus-based typology**


Cypriot Greek
Northern Kurdish
Persian
Coastal Balochi
English
Tabasaran
Sanzhi Dargwa
Bora
Aguaruna
Jinghpaw
Mandarin
Sumbawa
Arta
Ainu
Japanese
Tondano
Tulil
Tolai
Teop
Vera'a
Nafsan
corpus —
available
in preparation

Corpus languages

language	affiliation	citation
Arta	Austronesian, Polynesian	(Kimoto 2019)
Cypriot Greek	I.E., Greek	(Hadjidas & Vollmer 2015)
English	I.E., Germanic	(Schiborr 2015)
Nafsan	Austronesian, Oceanic	(Thieberger & Brickell 2019)
Northern Kurdish	I.E., Iranian	(Haig et al. 2019)
Persian	I.E., Iranian	(Adibifar 2016)
Sanzhi Dargwa	Nakh-Daghest., Dargin	(Forker & Schiborr 2019)
Teop	Austronesian, Oceanic	(Mosel & Schnell 2015)
Tondano	Austronesian, Polynesian	(Brickell 2016)
Tulil	Papuan, Taulil-Butam	(Meng 2019)
Vera'a	Austronesian, Oceanic	(Schnell 2015)

- every corpus in the collection is an individually citable resource

Documentation

- **extensively documented**:
  - what's in it?
  - how's it structured?
  - what's changed?
    - → *collection overview*
  - how's it annotated?
    - → *guidelines* for basic schemes
    - → *annotation notes* for each corpus

Annotations

- time-aligned with audio recordings
- romanized transcriptions
  (alongside original orthographies where applicable)
- idiomatic English translations
- standard morphological glossing
  (as per *Leipzig Glossing Rules*)

Annotations

- **standardized annotations** for
  - morphosyntactic relations
    (with GRAID, Haig & Schnell 2014),
  - referent identification and tracking
    (with RefIND, Schiborr et al. 2018), and
  - the information status of newly introduced referents
    (with a reduced variant of RefLex, Riester & Baumann 2017)

GRAID

- *Grammatical Relations and Animacy in Discourse*
  (Haig & Schnell 2014)
- **form** and **syntactic function** of major clause constituents
- a uniform set of symbols captures generalized categories
- designed for **cross-linguistic comparability**
- complements, rather than replaces, morphological glossing

GRAID

(1) **Nafsan** (Austronesian, Oceanic)

	kineu	*a=*	*pam*	*natam̃ol*	*i=*	*tol*	*su*
	1SG	1SG.RS=	eat	person	3SG.RS=	three	PF
##ds	**pro.1:a**	=lv	v:pred	**np.h:p**	=rn	rn	rv

‘[The monster said,] “I have eaten three people.”’

[mc_nafsan_ntwam_0042]

GRAID

⟨ **np** . **h** : **p** ⟩

① ② ③

① full noun phrase (form)

② human, third person (animacy)

③ direct object (function)

GRAID

⟨ **pro** . **1** : **a** ⟩

① ② ③

① free definite pronoun (form)

② human, first person (animacy)

③ subject of a transitive clause (function)

GRAID

- glosses **align with the (lexical) head of NPs**,
  but target entire phrases

- definition of grammatical roles follows Andrews (2007),
- and is based on **language-specific benchmark constructions**

GRAID

- GRAID primarily aims to **identify basic syntactic functions**
- other elements are left underspecified, or optionally glossed
  (e.g. NP subconstituents, verbal expressions, etc.)
- basic categories can be refined through optional tags
  (e.g. ⟨pro⟩ → ⟨dem_pro⟩; ⟨:s⟩ → ⟨:s_ds⟩)
- anomalous segments are noted, but left unanalyzed

- includes symbols for **zero anaphora** and **clause boundaries**

RefIND

- *Referent Indexing in Natural-language Discourse*
  (Schiborr et al. 2018)
- assigns unique **indices** to individual **discourse referents**, which are noted every time a referent is mentioned
- allows referents to be identified and tracked through a text
- also: metadata on ontological class of referents
  + hyponymic/meronymic relations between referents

GRAID + RefIND

(2) **Nafsan** (Austronesian, Oceanic)

	kineu	*a=*	*pam*	*nataṁol*	*i=*	*tol*	*su*
	1SG	1SG.RS=	eat	person	3SG.RS=	three	PF
##ds	pro.1:a	=lv	v:pred	np.h:p	=rn	rn	rv
	0026			**0048**			

‘[The monster said,] “I have eaten three people.”’

[mc_nafsan_ntwam_0042]

Structure and formats

- **WAV**, **MP3** recordings
- **TSV**, **XML**
  transcriptions, translations, annotations, and metadata;
  simple, flexible, and easily adaptable to analysts' needs
  and other existing formats (via XSLT etc.)
- **EAF**
  for the free, open annotation software ELAN,
  developed at the MPI Nijmegen;
  used by most of our annotators to annotate data
- ***multicastR***
  package for statistical programming language R

corpus	text	uid	gword	gloss	graid	refind
english	kent01	0164	#	#	##neg	
english	kent01	0164	*and*	and	other	
english	kent01	0164	*the*	the	ln_det	
english	kent01	0164	*house*	house	np:dt	0015
english	kent01	0164	#	#	#rc	
english	kent01	0164	0	0_house	rel_0:g	0015
english	kent01	0164	*we*	1PL	pro.1:s	0014
english	kent01	0164	*come*	come.PST	v:pred	
english	kent01	0164	*in*	in	adp	
english	kent01	0164	*first*	first	other	
english	kent01	0164	%	%	%	
english	kent01	0164	*we*	1PL	pro.1:s	0014
english	kent01	0164	*did-n't*	do.PST-NEG	lv_aux	
english	kent01	0164	*stop*	stop.INF	v:pred	
english	kent01	0164	*long*	long	other	

Companion R package

- ***multicastR*** (Schiborr 2018)
  - for the free statistical programming language R
  - accesses corpus data (and metadata) directly in R, downloaded from our servers
  - allows selection of specific versions
  - plus a few convenience functions
  - can be installed from CRAN or manually from source files on our website

Open science

- **restriction free**
  licensed under a *Creative Commons* (CC-BY 4.0) licence or in the public domain
- **freely accessible**
  from the servers of the University of Bamberg
- **open software**
  based on open software and formats
- **extensively documented**
  design, structure, and annotations

Replicability

- continuously updated with new and revised material
- keep older versions of **the entire collection**
  on record as complete ‘snapshots’
- allows exact replication of published research results
  (if methods are published as well, e.g. as online appendices)
- for our own work using Multi-CAST,
  plan to include associated scripts in our R package,
  keeping data and code side-by-side

Archival

- currently:
  all files stored on a webserver
  hosted by the University of Bamberg
- long-term storage:
  ???

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

`multicast.aspra.uni-bamberg.de/`

first stop: *collection overview* (PDF)

contact information at the bottom of the webpage

Case study

- examine some of the dimensions of **referential choice**:
  - referent semantics: humanness
  - discourse context: recency

  (from a broad top-down perspective, glancing over most detail!)

- using latest Multi-CAST data (from August 2019 + extras) and associated tools (regex, R and *multicastR*)

Referential choice

(3) *I went along with this old man, Mr Barnes.*

(4) ***He*** *was a nice old man.*

. . .

(5) **Ø** *used to have a team of four great horses.*

[mc_english_kent03_0021;0025]

Referential choice

- referring expressions differ in **informativity** and **specificity**
  (e.g. zero vs. full NPs)
- speakers needs to select appropriate forms
  to facilitate identification of the intended referents
  by listeners (“recipient design”)

Referential choice

- referential choice is influenced in some way by the properties of **the preceding discourse**
  - activation states (Chafe 1976, 1994)
  - accessibility (Ariel 1990, 2004; Arnold 2010)
  - centering (Grosz et al. 1995)
  - *and others* (e.g. Kibrik 2000)
  - topic continuity (Givón 1983)
  - givenness (Prince 1981; Gundel et al. 1993)
  - discourse prominence (Gordon & Hendrick 1997)

Working with Multi-CAST

1. access the data, e.g. via *multicastR*
2. establish sampling criteria
3. identify forms of referring expressions
4. identify properties of referents and individual mentions

Sampling criteria

- **only subjects** (“`:(a|s|ncs)($|_)`”)
- only mentions of fully referential expressions
  (i.e. those tracked by RefIND)
- **only given mentions**
  (i.e. second and subsequent mentions)
- only positions where a pragmatic choice is possible
  (e.g. no reflexives, gaps in relative clauses)
- only third person mentions
  (i.e. not first or second person)

The sample

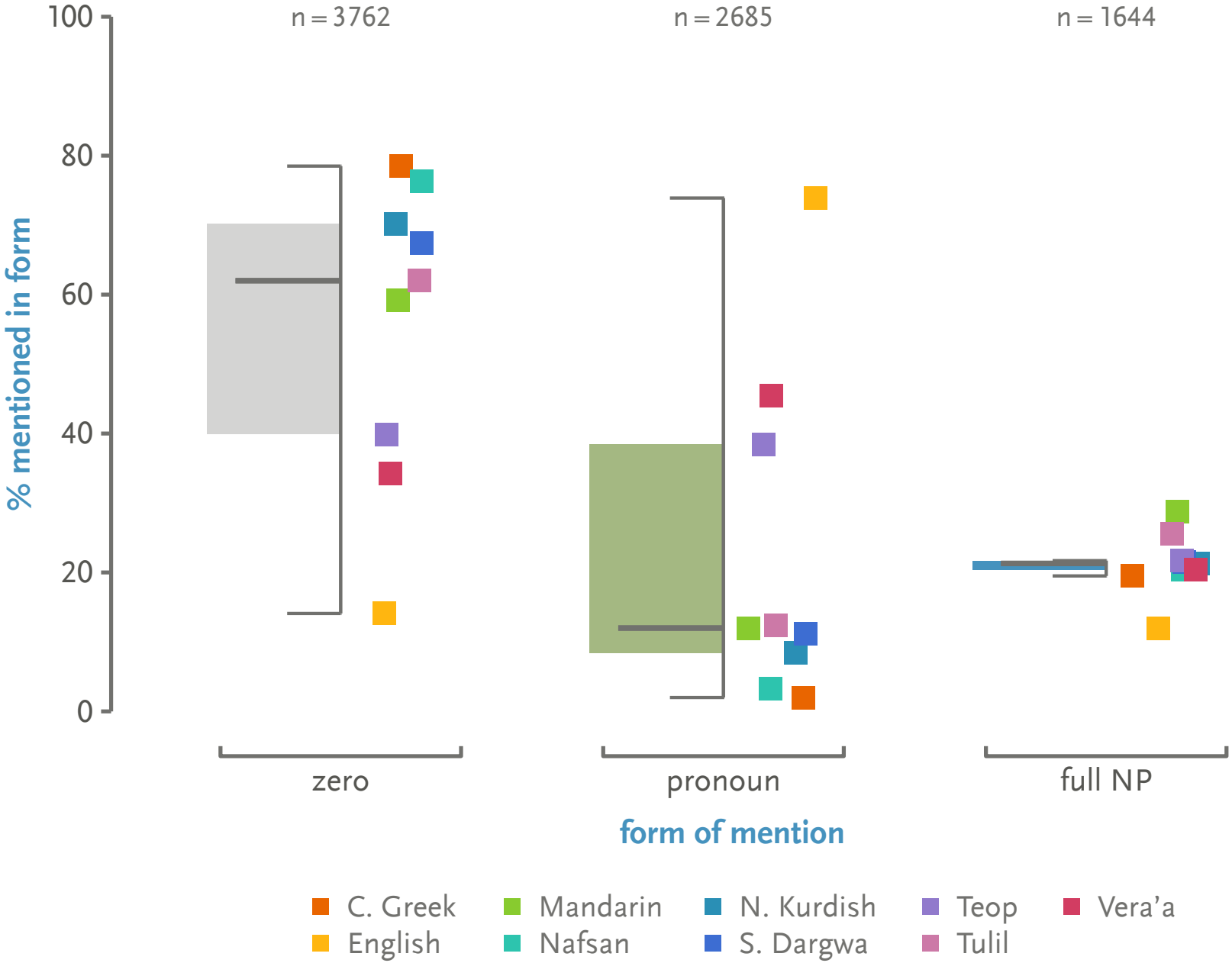
corpus	clause units	unique referents	mentions as subject
Cypriot Greek	1 071	99	441
English	4 184	509	1 343
Mandarin*	1197	109	715
Nafsan	1 012	118	692
Northern Kurdish	1 359	120	642
Sanzhi Dargwa	1 066	103	475
Teop	1 302	101	771
Tulil	1 264	148	590
Vera'a	3 608	293	2 422
totals	**14 866**	**1 600**	**8 091**

* (Vollmer, in prep.)

Form of mentions

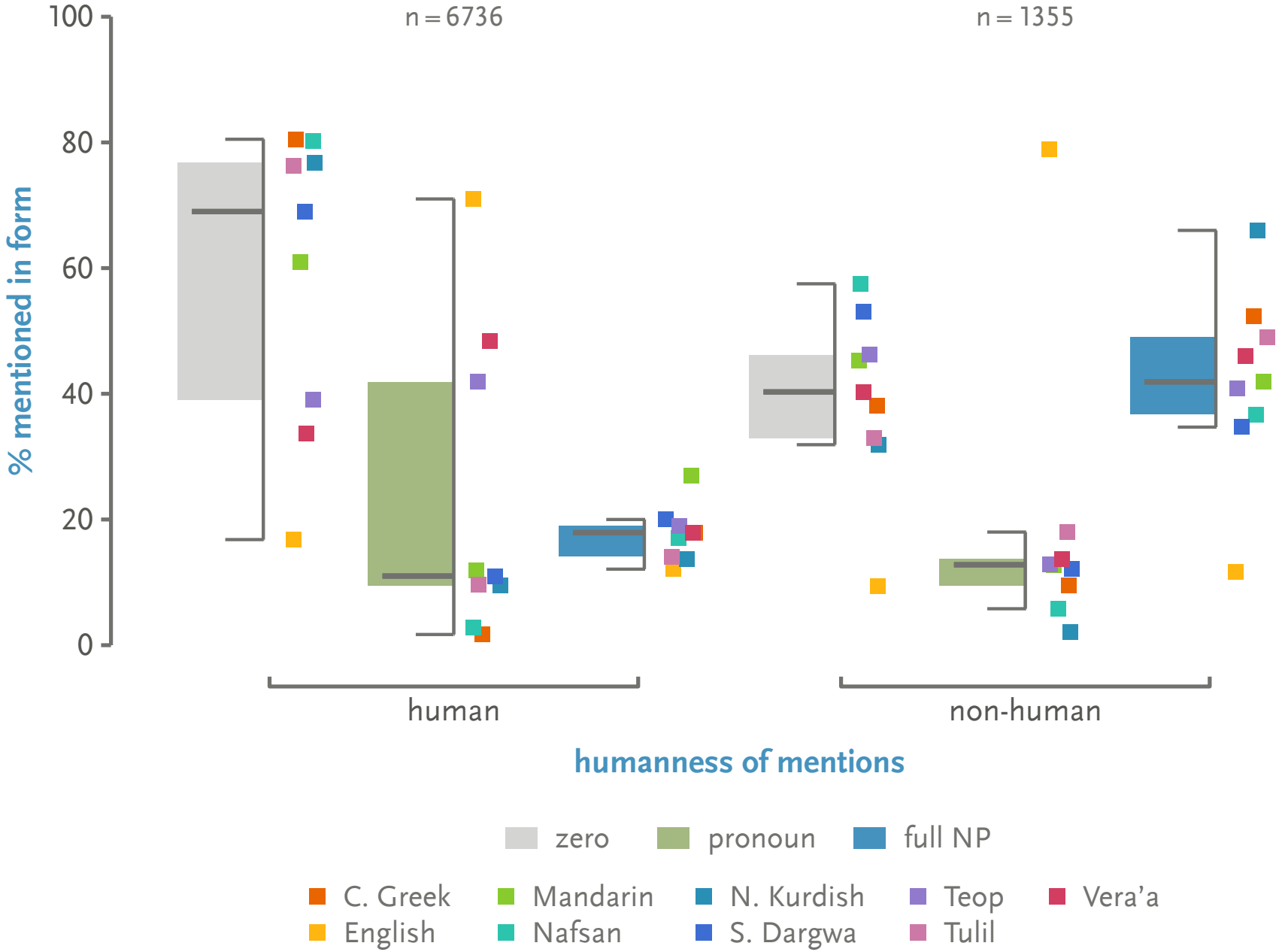
- **three basic form types**:
  - full noun phrases ('lexical' NPs, e.g. *the woman*),
  - free pronouns (e.g *she*, *her*), and
  - zero anaphora

- all captured by the GRAID annotations:
  - "`(^|\W|_)?np`" → full NPs
  - "`(^|\W|_)?pro`" → pronouns
  - "`(^|\W|_)?0`" → zero


100
80
60
40
20
0
% mentioned in form
n = 3762
n = 2685
n = 1644
zero
pronoun
full NP
form of mention
C. Greek
English
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a

Humanness of mentions

- **two values**:
  - human or
  - non-human
- regular expressions matching the GRAID annotations:
  - "\.h" → human (+ third person)
  - (non-human third person is unmarked)
  - then filter for first/second person mentions, "\.[12]"


100
80
60
40
20
0
% mentioned in form
n = 6736
n = 1355
human
non-human
humanness of mentions
zero
pronoun
full NP
C. Greek
English
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a

Recency effects

- one factor deemed highly influential:
  textual distance to a co-referential antecedent
  (Ariel 1990, Kibrik 2000; NLP pronoun resolution, etc.)
- in other words,
  how long ago was a specific referent last mentioned?
- unit of measurement:
  elapsed time, words, **clauses**, intervening referents, ...

Antecedent distance

gword	graid	refind
	##	
I	pro.1:s	0000
went	v:pred	
with	adp	
this	ln	
man	np.h:obl	0036
	##	
bla	other	
bla	other	
	##	
he	pro.h:a	0036
had	v:pred	
horses	np:p	0042

Antecedent distance

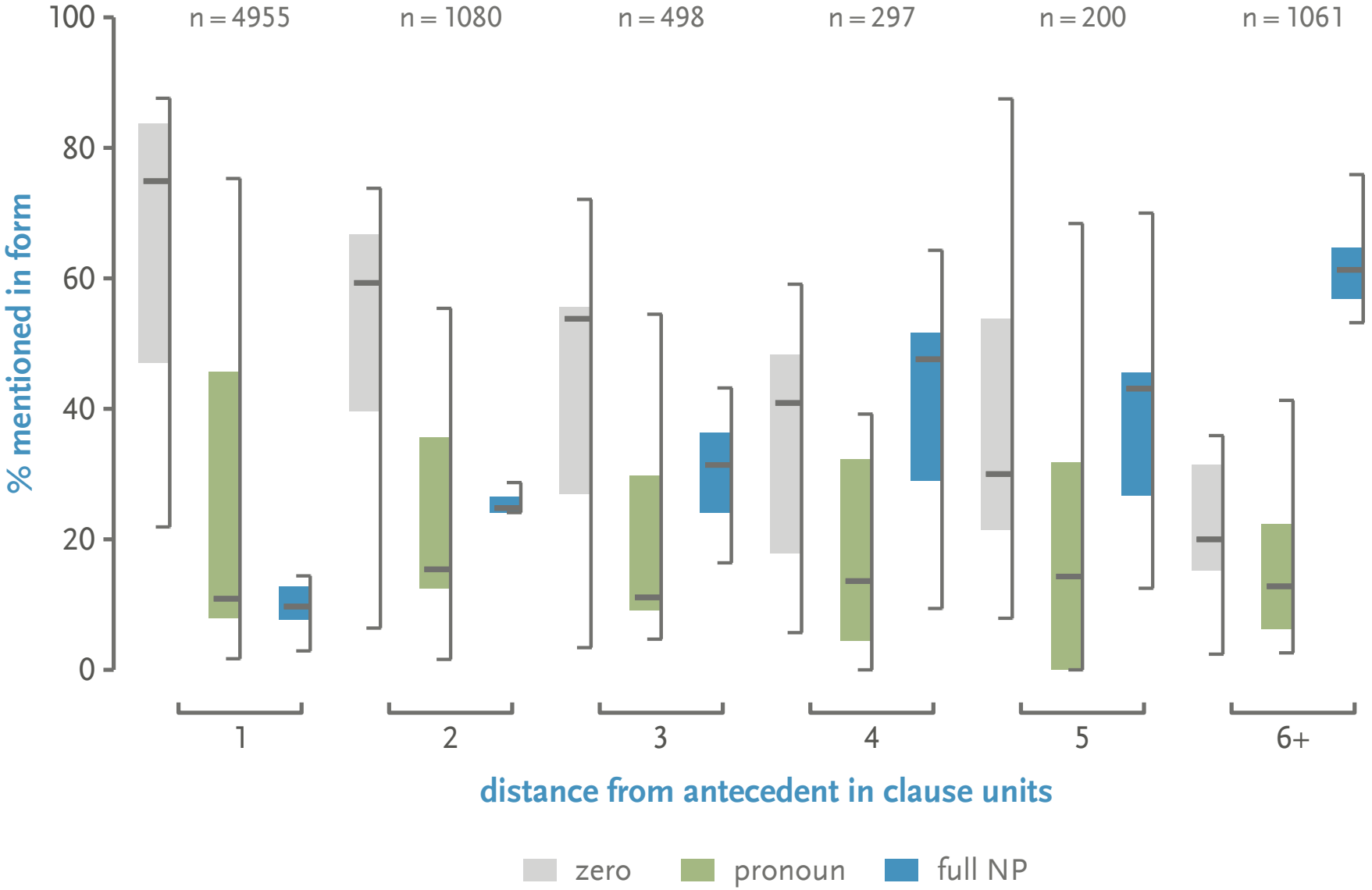
gword	graid	refind	*clause index*
	##		**1** ← *clause boundary*
I	pro.1:s	0000	
went	v:pred		
with	adp		
this	ln		
man	np.h:obl	0036	
	##		**2** ← *clause boundary*
bla	other		
bla	other		
	##		**3** ← *clause boundary*
he	pro.h:a	0036	
had	v:pred		
horses	np:p	0042	

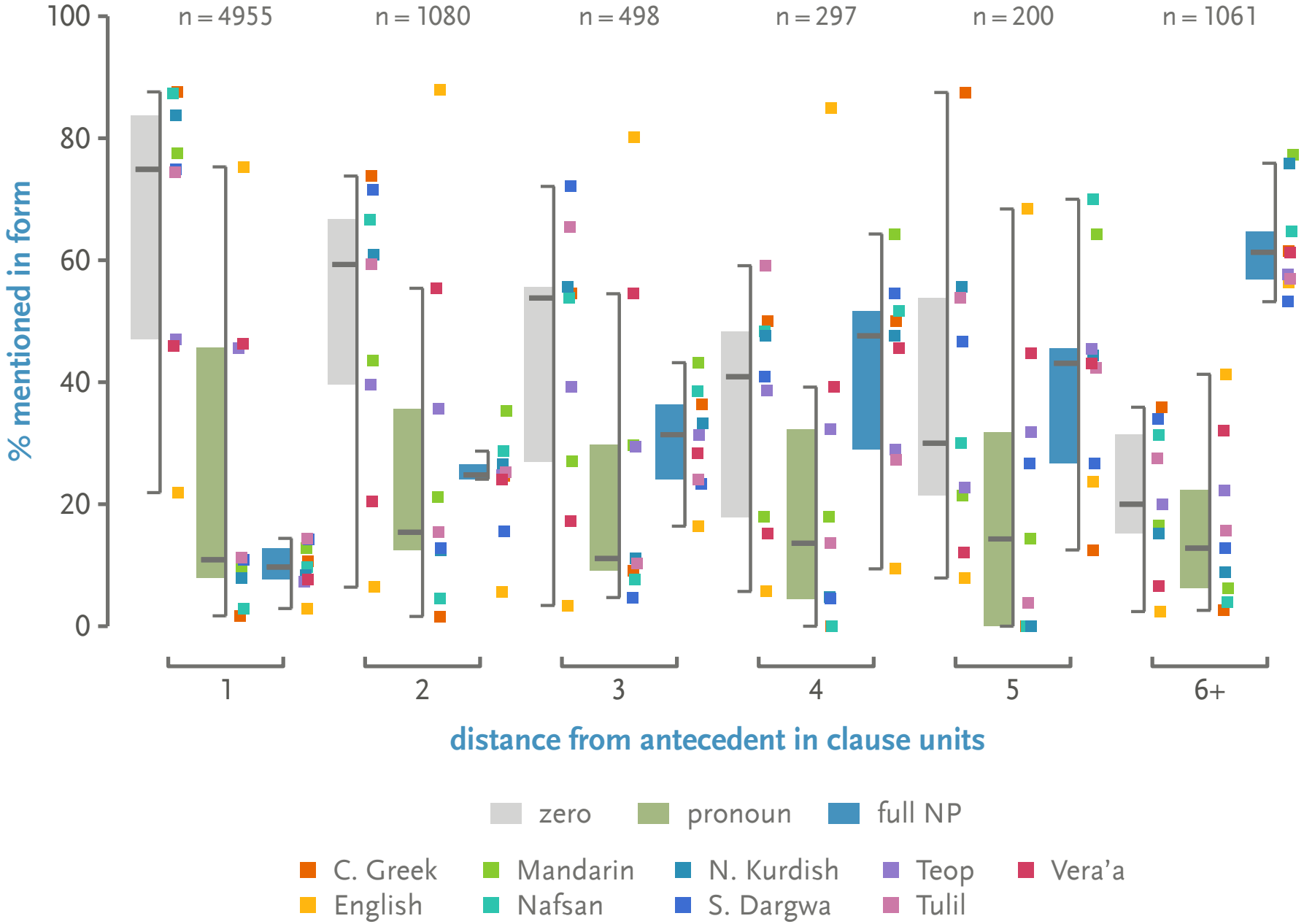
Antecedent distance

gword	graid	refind	*clause index*	
	##		**1**	← *clause boundary*
I	pro.1:s	0000	1	
went	v:pred		1	
with	adp		1	
this	ln		1	
man	np.h:obl	0036	1	
	##		**2**	← *clause boundary*
bla	other		2	
bla	other		2	
	##		**3**	← *clause boundary*
he	pro.h:a	0036	3	
had	v:pred		3	
horses	np:p	0042	3	

Antecedent distance

gword	graid	refind	*clause index*	
	##		**1**	← *clause boundary*
I	pro.1:s	0000	1	
went	v:pred		1	
with	adp		1	
this	ln		1	
man	np.h:obl	0036	1	← *antecedent*
	##		**2**	← *clause boundary*
bla	other		2	
bla	other		2	
	##		**3**	← *clause boundary*
he	pro.h:a	0036	3	← *anaphor @ 3 − 1 = 2 clauses distance*
had	v:pred		3	
horses	np:p	0042	3	


n = 4955
n = 1080
n = 498
n = 297
n = 200
n = 1061
100
80
60
40
20
0
% mentioned in form
1
2
3
4
5
6+
distance from antecedent in clause units
zero
pronoun
full NP


n = 4955
n = 1080
n = 498
n = 297
n = 200
n = 1061
100
80
60
40
20
0
% mentioned in form
1
2
3
4
5
6+
distance from antecedent in clause units
zero
pronoun
full NP
C. Greek
English
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a

In summary

- most languages have a preferred default form of reference (zero or pronouns)
- strongest inter-corpus variation in zero/pronoun choice
- selection criteria for full NPs are similar across corpora
- human referents less likely to be full NPs than non-human
- rate of zero drops as antecedent distance increases; inverse for full NPs

- choice of full NPs over other forms
  → candidate for a discourse universal?

And more

- also possible with Multi-CAST:
  - phrase weight,
  - role of demonstratives,
  - finer distinctions of referent types (beyond humanness),
  - positional cues (e.g. word order alternations),
  - role continuity,
  - local information pressure,
  - competition between candidate antecedents,
  - semantic predicate types [t.b.a.],

    etc.

Multi-CAST

- **spoken corpora from 11 typologically diverse languages**
- chiefly monologic, non-elicited, non-translated language
- 10 additional corpora in preparation

- time-aligned with audio recordings
- minimum 1 000 clauses per corpus
- 20 000 clause units in total (c. 85 000 words)

- multiple layers of **annotation** for morphosyntax, referent tracking
- for quantitative, corpus-based typology
- **restriction-free, designed for replicability**

- have a dataset that fits? Contact us!

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/

References

Adibifar, Shirin. **2016**. Multi-CAST Persian. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Arnold, Avery. **2007**. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description*, Vol. 1, 132–223. Cambridge: Cambridge University Press.

Ariel, Mira. **1990**[2014]. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. **2004**. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes* 37(2). 91–116.

Arnold, Jennifer E. **2003**. Multiple constraints on reference form: Null, pronominal, and full reference in Mapudungun. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure: Grammar as architecture for function,* 225–245. Amsterdam: John Benjamins.

Brickell, Timothy. **2016**. Multi-CAST Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

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.

References

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

Forker, Diana & **Schiborr**, Nils N. **2019**. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST.*

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

Gordon, Peter C. & **Hendrick**, Randall. **1997**. Intuitive knowledge of linguistic co-reference. *Cognition* 62(2). 325–370.

Grosz, Barbara J. & **Joshi**, Aravind K. & **Weinstein**, Scott. **1995**. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2). 203–225.

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

Hadjidas, Harris & **Vollmer**, Maria C. **2015**. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST.*

Haig, Geoffrey & **Schnell**, Stefan. **2014**. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators.* Version 7.0. (`https://multicast.aspra.uni-bamberg.de/#annotations`)

References

Haig, Geoffrey & **Schnell**, Stefan. **2019**[2015]. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (`https://multicast.aspra.uni-bamberg.de/`)

Haig, Geoffrey & **Vollmer**, Maria & **Thiele**, Hanna. **2019**. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Kibrik, Andrej A. **2000**. A cognitive calculative approach towards discourse anaphora. In Baker, Paul & Hardie, Andrew & McEnery, Tony & Siewierska, Anna (eds.), *Proceedings from the 3rd Discourse Anaphora and Reference Resolution Colloquium (DAARC 2000)*, 72–82. Lancaster: Lancaster University Centre for Computer Corpus Research on Language.

Kimoto, Yukinori. **2019**. Multi-CAST Arta. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Meng, Chenxi. **2019**. Multi-CAST Tulil. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Mosel, Ulrike & **Schnell**, Stefan. **2015**. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

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

References

Riester, Arndt & **Baumann**, Stefan. **2017**. The RefLex scheme — Annotation guidelines. *SpinSpec: Working papers of the SFB 732* 14. (http://elib.uni-stuttgart.de/handle/11682/9028)

Schiborr, Nils N. **2015**. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Schiborr, Nils N. **2018**. multicastR: A companion to the Multi-CAST collection. R package version 1.3.0. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://cran.r-project.org/package=multicastR)

Schiborr, Nils N. & **Schnell**, Stefan & **Thiele**, Hanna. **2018**. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines*. Version 1.1. University of Bamberg. (https://multicast.aspra.uni-bamberg.de/#annotations)

Schnell, Stefan. **2015**. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Thieberger, Nick & **Brickell**, Timothy. **2019**. Multi-CAST Nafsan. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.