BICLCE 2019

Quantitative models of referential choice

Lexical anaphora in English

Nils Norman Schiborr
University of Bamberg

26 September 2019
v1.0

Referential choice

- How do speakers choose the most appropriate **form of expression** to refer back to an established **discourse referent**?

Referential choice

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

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

…

(3) *[**The man** | **He** | **Ø**] used to have a team of four great horses.*

[mc_english_kent03_0021;0025]

Referential choice

- **referring expressions** differ in
  informativity, specificity of reference

- **recipient design**:
  choice of form should facilitate identification of the intended referent

Discourse factors

- ideal choice of form is influenced
  by **the properties of the preceding discourse**
  and of **the referent itself**

- activation states, accessibility
  (Chafe 1976, 1994; Ariel 1990, 2004)

- topic continuity, focus
  (Givón 1983; Lambrecht 2010, etc.)

(cf. further Prince 1981; Gundel et al. 1993;
Grosz et al. 1995; Gordon & Hendrick 1997, etc.)

Discourse universality

- from a typological perspective,
  same factors are relevant across languages

- **but:**
  relative weighting between factors differs

Case study

- **lexical subjects**:
  choice of a full NP over other forms for subjects
  (i.e. pronouns, zero)
- from a comparative, discourse-structural perspective

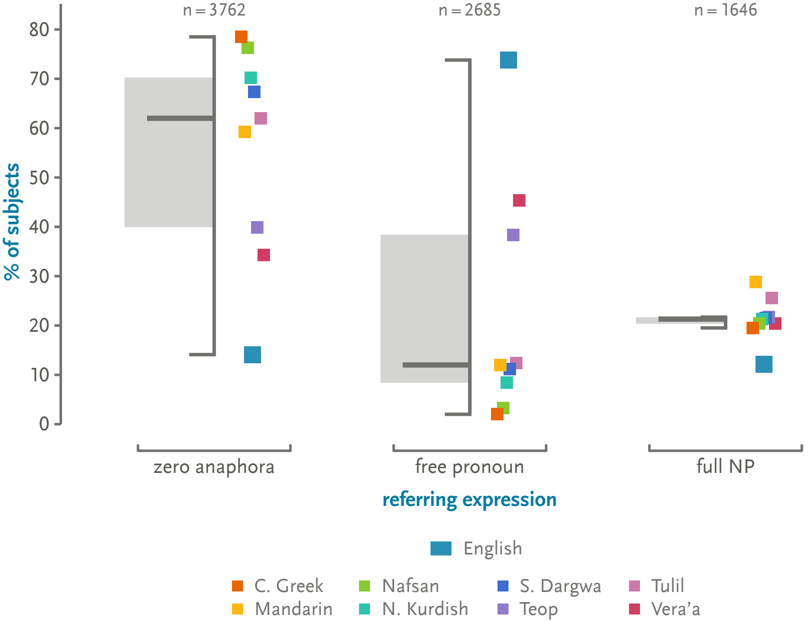
Corpus data

- corpus data from **eight languages** vs. **English**
  - spontaneous, natural spoken language
  - chiefly monologic narratives
    (folktales, personal narratives)
  - uniformly annotated for cross-corpus comparability
    (incl. zero anaphora and syntactic boundaries)

- only referential NP subjects
- only third person subjects
  (i.e. excluding first/second person)

The sample

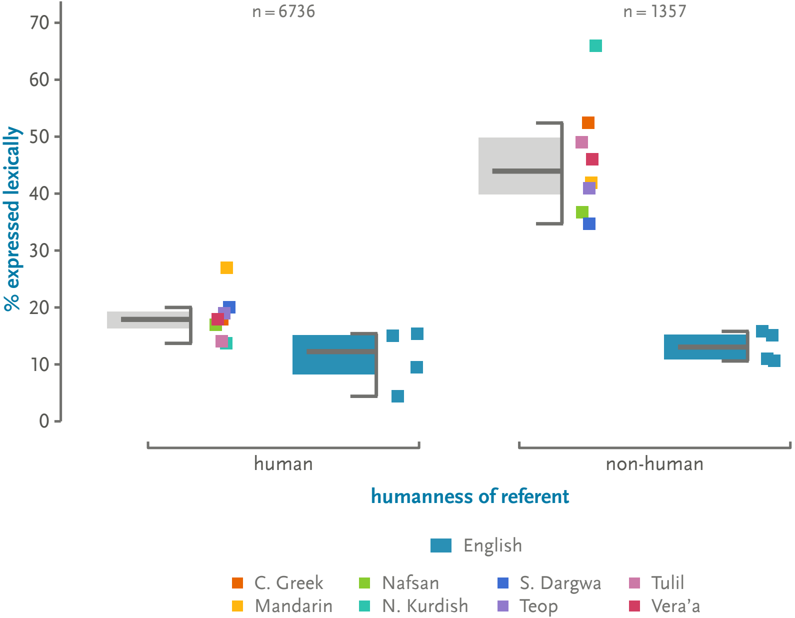
corpus	clause units	unique referents	sampled subjects
English	4 184	509	1 345
Cypriot Greek	1 071	99	441
Mandarin	1 197	109	715
Nafsan	1 012	118	692
Northern Kurdish	1 359	120	642
Sanzhi Dargwa	1 066	103	475
Teop	1 302	101	771
Tulil	1 264	148	590
Vera'a	3 608	293	2 422
totals	**14 866**	**1 600**	**8 093**



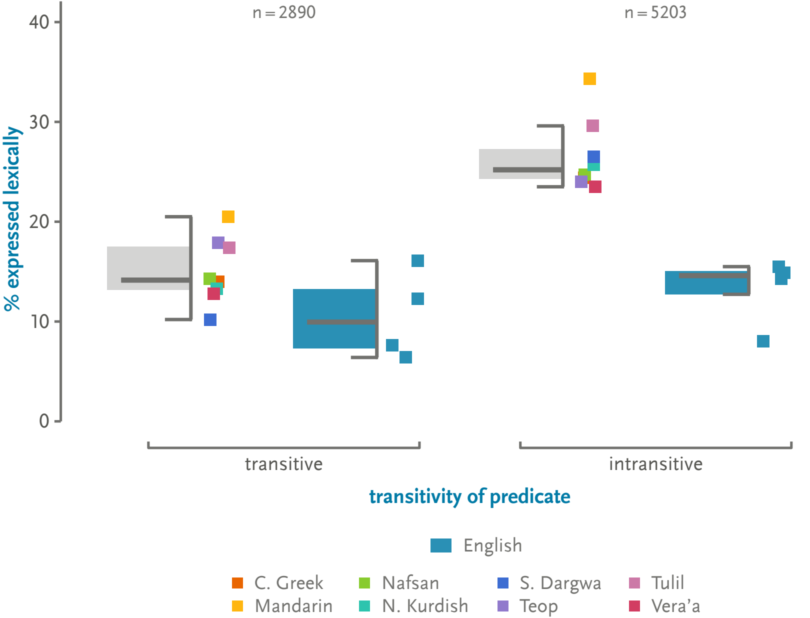
n = 3762
n = 2685
n = 1646
80
70
60
50
40
30
20
10
0
% of subjects
zero anaphora
free pronoun
full NP
referring expression
English
C. Greek
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a

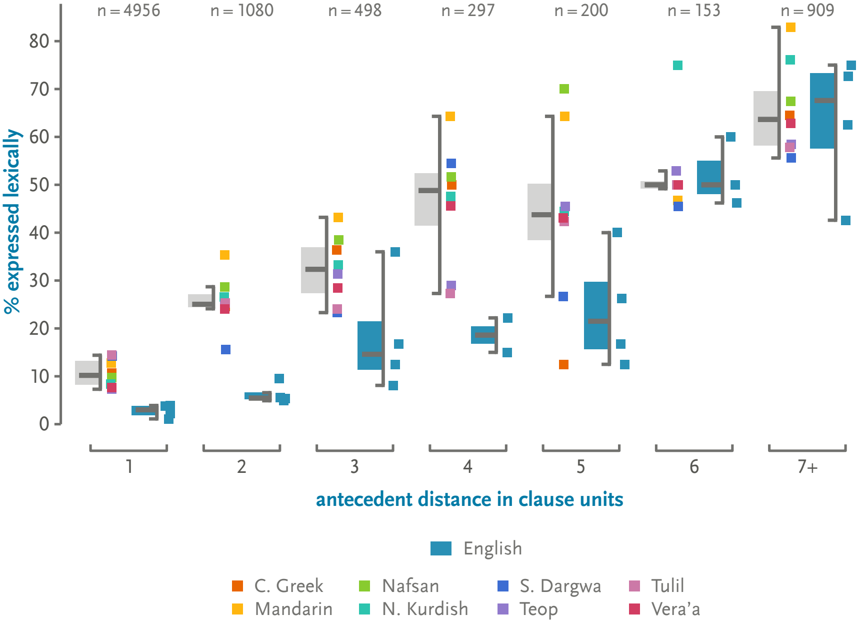
Factors

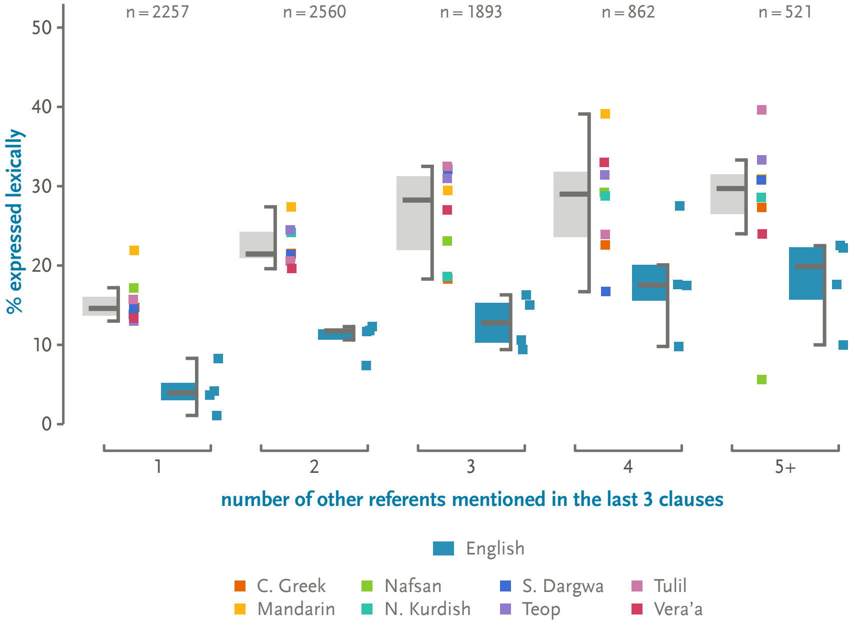
- inherent semantic properties of the referent:
  → **humanness**
- properties of the predicate:
  → **transitivity**
- properties of the preceding discourse:
  → **recency**
  → **local information pressure**
- properties of the antecedent:
  → **form of the antecedent**
  → **function of the antecedent**



n = 6736
n = 1357
% expressed lexically
70
60
50
40
30
20
10
0
human
non-human
humanness of referent
English
C. Greek
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a


n = 2890
n = 5203
40
30
20
10
0
% expressed lexically
transitive
intransitive
transitivity of predicate
English
C. Greek
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a


n = 4956
n = 1080
n = 498
n = 297
n = 200
n = 153
n = 909
% expressed lexically
0
10
20
30
40
50
60
70
80
1
2
3
4
5
6
7+
antecedent distance in clause units
English
C. Greek
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a



n = 2257
n = 2560
n = 1893
n = 862
n = 521
50
40
30
20
10
0
% expressed lexically
1
2
3
4
5+
number of other referents mentioned in the last 3 clauses
English
C. Greek
Mandarin
Nafsan
N. Kurdish
S. Dargwa
Teop
Tulil
Vera'a

Correlation coefficients

- Pearson’s *r* for various factors by lexicality, subjects only

factor	8 corpora	English
referent is human	**0.887**	0.275
clause is transitive	**0.870**	0.348
antecedent distance	**0.843**	**0.885**
information pressure	0.557	**0.749**
antecedent is lexical	**0.894**	**0.747**
antecedent is subject	−**0.840**	−0.453

In summary

- in **English**, compared to broad cross-corpus tendencies,
- the selection of full NPs in subject position is less sensitive to animacy and morphosyntax
  (other corpora likewise, but to a lesser degree)

- conversely, discourse properties have a greater influence

How come?

- rather than applying universally,
- the factors influencing referential choice may be **parametrized across languages**

- e.g. for humanness in English:
  - prounouns are the preferred default form of reference
  - pronouns are marked for humanness (in the singular)
  - hence less need to disambiguate via full NPs

How come?

- other possible explanations:
  - differences in text type
    (**but:** in other corpora text types cluster together)
  - differences in content
    (**but:** texts are long and varied in subject matter)
  - speaker idiosyncracies

- English is anomalous in a number of other respects

Summary

- from a cross-linguistic perspective,
- **choice of form for subjects** is influenced by
  - properties of the preceding discourse,
  - semantic properties of the referent,
  - properties of the predicate,
  - function and form of the antecedent
- **but:** in some corpora (e.g. English), certain factors less relevant (for English, **animacy** and **role**)
- **factors influencing referential choice not universal, but weighted differently across languages?**

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

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

— spoken language corpora from 11 languages —

— annotated for cross-corpus typological research —

— fully documented, freely accessible —

Corpus data

language	affiliation	citation
English	I.E., Germanic	(Schiborr 2015)
Cypriot Greek	I.E., Greek	(Hadjidas & Vollmer 2015)
Mandarin	Sino-Tibetan, Sinitic	(Vollmer, in prep.)
Nafsan	Austronesian, Oceanic	(Thieberger & Brickell 2019)
Northern Kurdish	I.E., Iranian	(Haig et al. 2019)
Sanzhi Dargwa	Nakh-Daghest., Dargin	(Forker & Schiborr 2019)
Teop	Austronesian, Oceanic	(Mosel & Schnell 2015)
Tulil	Papuan, Taulil-Butam	(Meng 2019)
Vera’a	Austronesian, Oceanic	(Schnell 2015)

References

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.

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.

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.

References

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`)

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.*

Lambrecht, Knud. **2010**. Constraints on subject-focus mapping in French and English: A contrastive analysis. In Breul, Carsten & Göbbel, Edward (eds.), *Comparative and contrastive studies in information structure,* 77–100. Amsterdam: John Benjamins.

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

References

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.

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

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.*