ISSLaC3

Data-driven models of referential choice

Antecedent distance and beyond

Nils Norman Schiborr
University of Bamberg

7 December 2018
v1.1

Referential choice

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

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

. . .

(3) ○○ *used to have a team of four great horses.*

[english_kent03a_021;025]

Referential choice

influenced in some way by the preceding discourse, e.g.

- activation states (Chafe 1976, 1994)
- topic continuity (Givón 1983)
- accessibility (Ariel 1990[2014], 2004; Arnold 2010)
- givenness (Gundel et al. 1993)
- centering (Grosz et al. 1995)
- discourse prominence (Gordon & Hendrick 1997)
- and others (e.g. Kibrik 2000)

Accessibility marking scale

less accessible

full name
long definite lexical NP
short definite lexical NP
last name
first name
lexical NP with distal demonstrative
lexical NP with proximal demonstrative

distal demonstrative
proximal demonstrative
stressed pronoun
unstressed pronoun
cliticized pronoun
verbal person agreement

more accessible

zero

(adapted from Ariel 1990: 73)

Accessibility theory

accessibility theory

> *"provides one and the same account for expressions considered referential (e.g., proper names) [...] as well as for expressions considered anaphoric (e.g., pronouns) [...] It also does not view references to the speech situation (e.g., by deictics) as special."*
>
> (Ariel 2006: 15, emphasis added)

Hypothesis

in other words, speakers' choice between, e.g.

(A) between a full, lexical noun phrase (*the old man; Mr Barnes*) and a pronominal NP (e.g. *he*), and

(B) between a pronoun and zero anaphora,

should be predictable from **the same set** of explanatory variables

Corpus data

a subset of the **Multi-CAST collection**: (Haig & Schnell 2015)

- Cypriot Greek (IE, Greek)
- English (IE, Germanic)
- Northern Kurdish (IE, Iranian)
- Sanzhi Dargwa (Nakh-Daghestanian, Dargin)
- Teop (Austronesian, Oceanic)
- Vera'a (Austronesian, Oceanic)

spoken, non-elicited, monologic **narratives**

(Hadjidas & Vollmer 2015; Schiborr 2015; Haig & Thiele 2015; Forker & Schiborr in prep.; Mosel & Schnell 2015; Schnell 2015)

Annotations

GRAID (Haig & Schnell 2014)

'Grammatical relations and animacy in discourse'

- form of referring expressions
- marks zero anaphora
- delineates texts into clause units

RefIND (Schiborr & Schnell & Thiele 2017)

'Referent indexing in natural-language discourse'

- identification and tracking of discourse referents
- enables calculation of anaphoric distances and frequencies

Annotations

(4) Sanzhi Dargwa [sanzhi_devil_034]

	xun-ne-b		*suk*	*b-ič-ib*	*k:urt:a*
	road-SPR-N	∅	meet	N-occur.PFV-PRET	fox
##	np:l	0.h:s	other	v:pred	np.d:p
		0002			**0031**

‘On the road (he) met a fox.’

Annotations

(5) Sanzhi Dargwa [sanzhi_devil_038]

	kːurtːa-l	*b-ičː-ib*	*hel-i-j*	*cin-na*	*ʁez*
	fox-ERG	N-give.PFV-PRET	that-OBL-DAT	REFL.SG-GEN	hair
##	np.d:a	v:pred	pro.h:g	ln_refl.d:poss	np:p
	0031		**0002**	**0031**	**0032**

‘The fox gave him one of its hairs.’

Sampling criteria

1. only referents that can be identified throughout a discourse
   (i.e. that are "trackable" in the sense of Schiborr et al. 2017: 3)
2. only referents with $n \geq 2$ total mentions
3. only second and subsequent mentions
   (i.e. excluding new introductions)
4. only third person mentions
   (i.e. excluding first/second person)

The model

explain the selection between

(A) **lexical** vs. **non-lexical** expressions, and

(B) among non-lexical expressions,
pronouns vs. **zero**

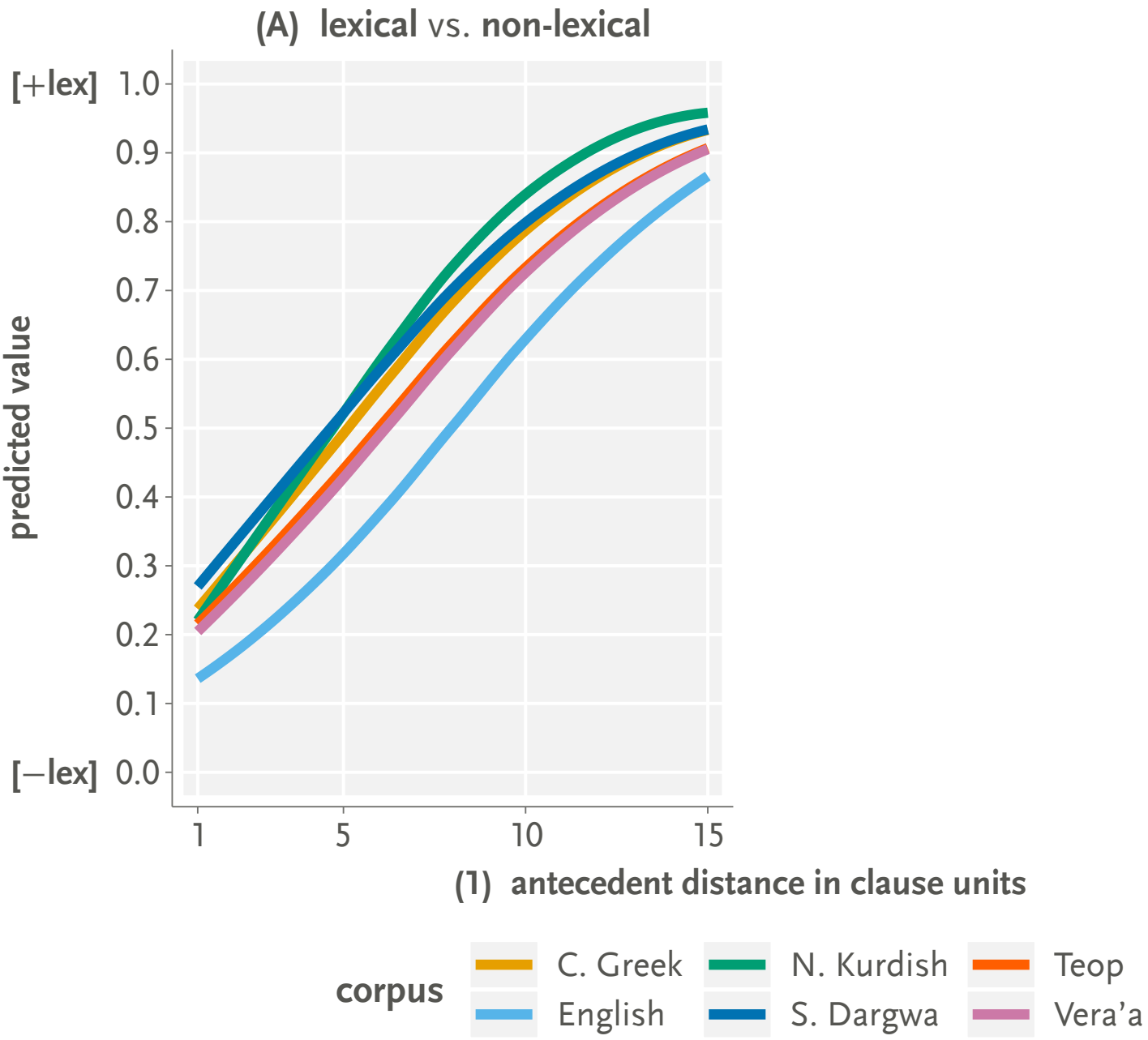
via the **explanatory variables**

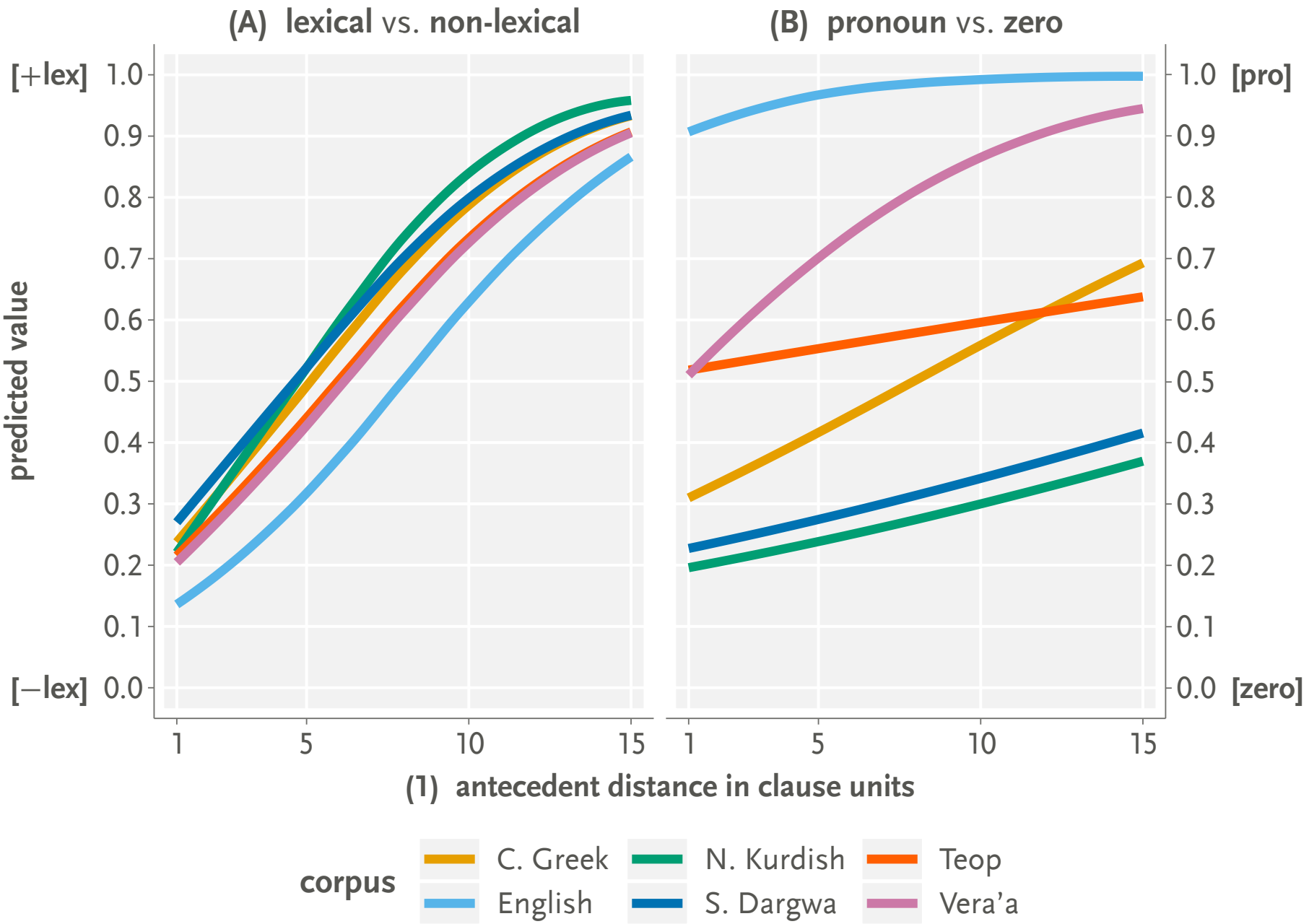
1. **antecedent distance** in clause units
2. **frequency** in recent discourse
3. mention **in previous clause** ($d = 1$)

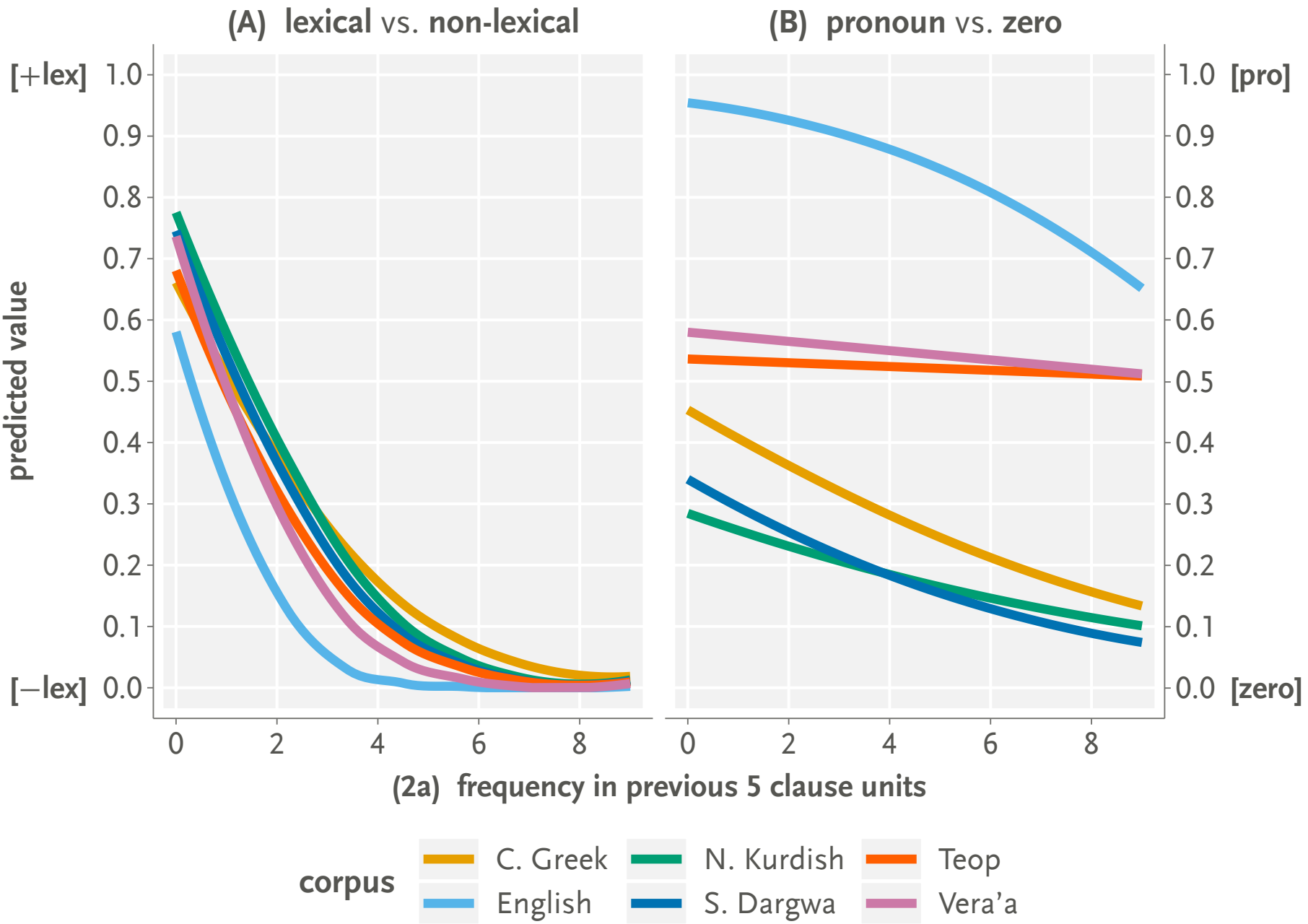
Sample statistics

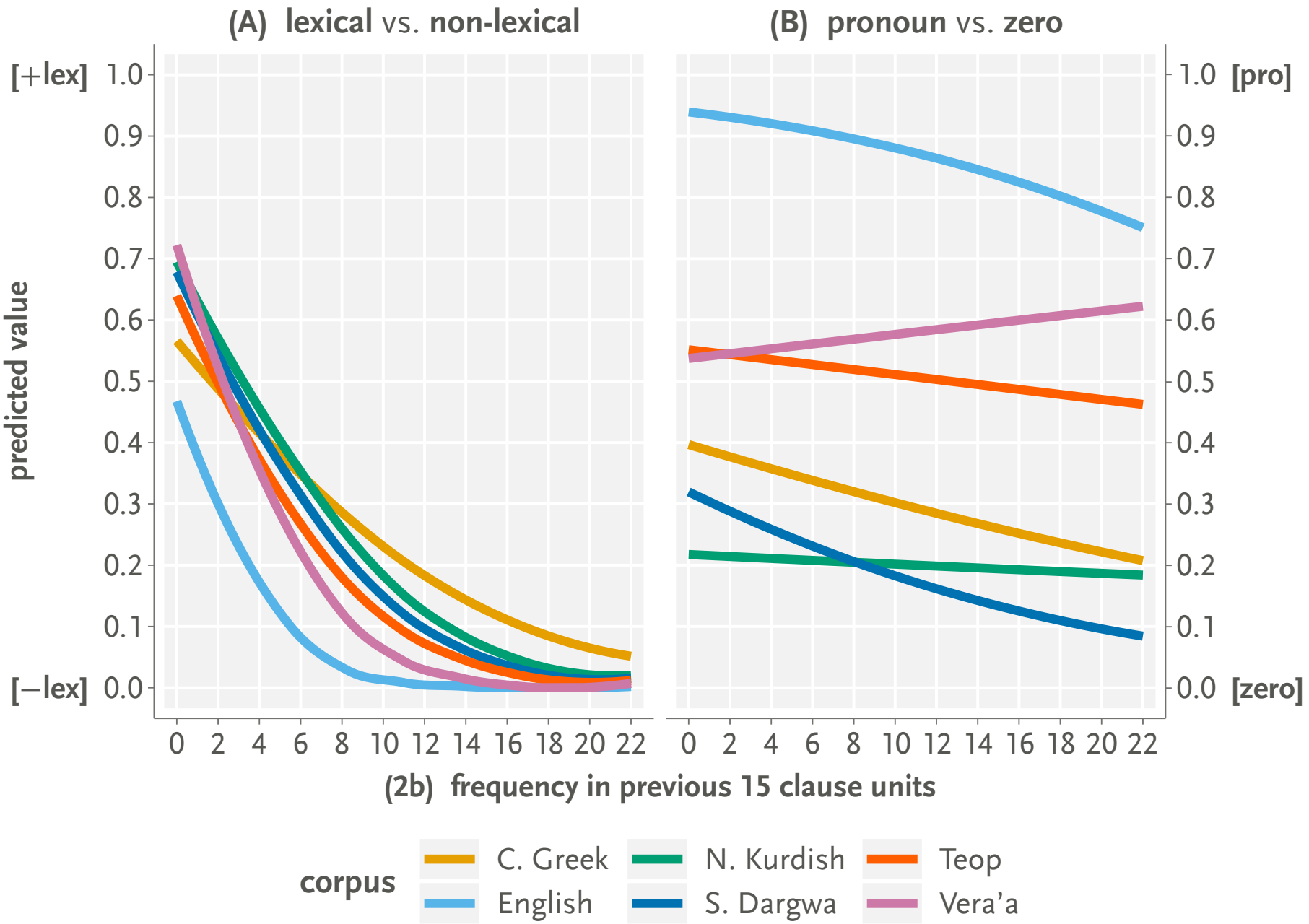
corpus	clauses	referents*	mentions
Cypriot Greek	1 071	165	1 653
English	1 244	261	1 452
Northern Kurdish	1 389	153	2 167
Sanzhi Dargwa	1 618	275	1 977
Teop	1 272	180	1 688
Vera'a	2 377	241	3 158
totals	**8 871**	**1 175**	**12 095**

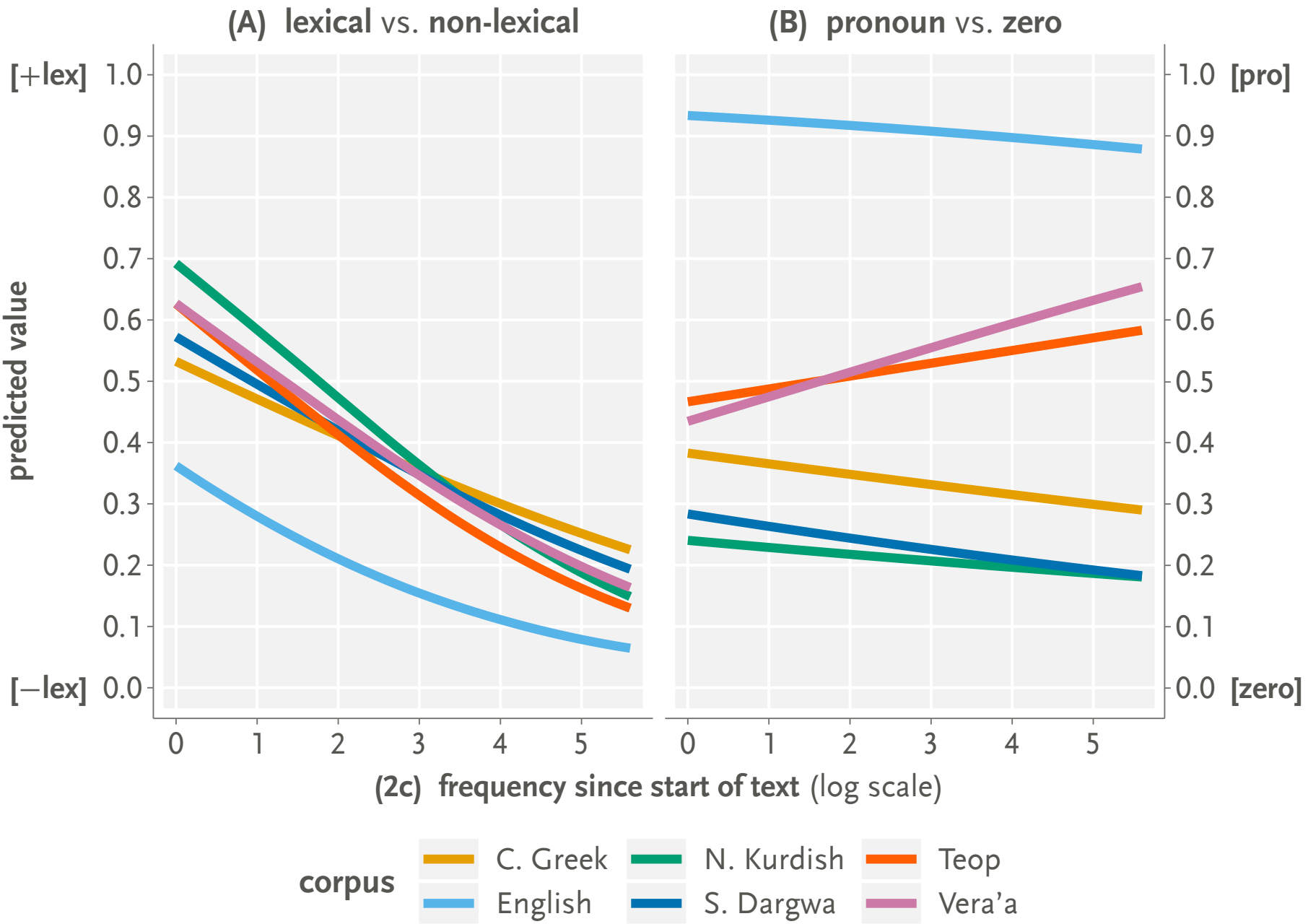
* with $n \geq 2$ mentions

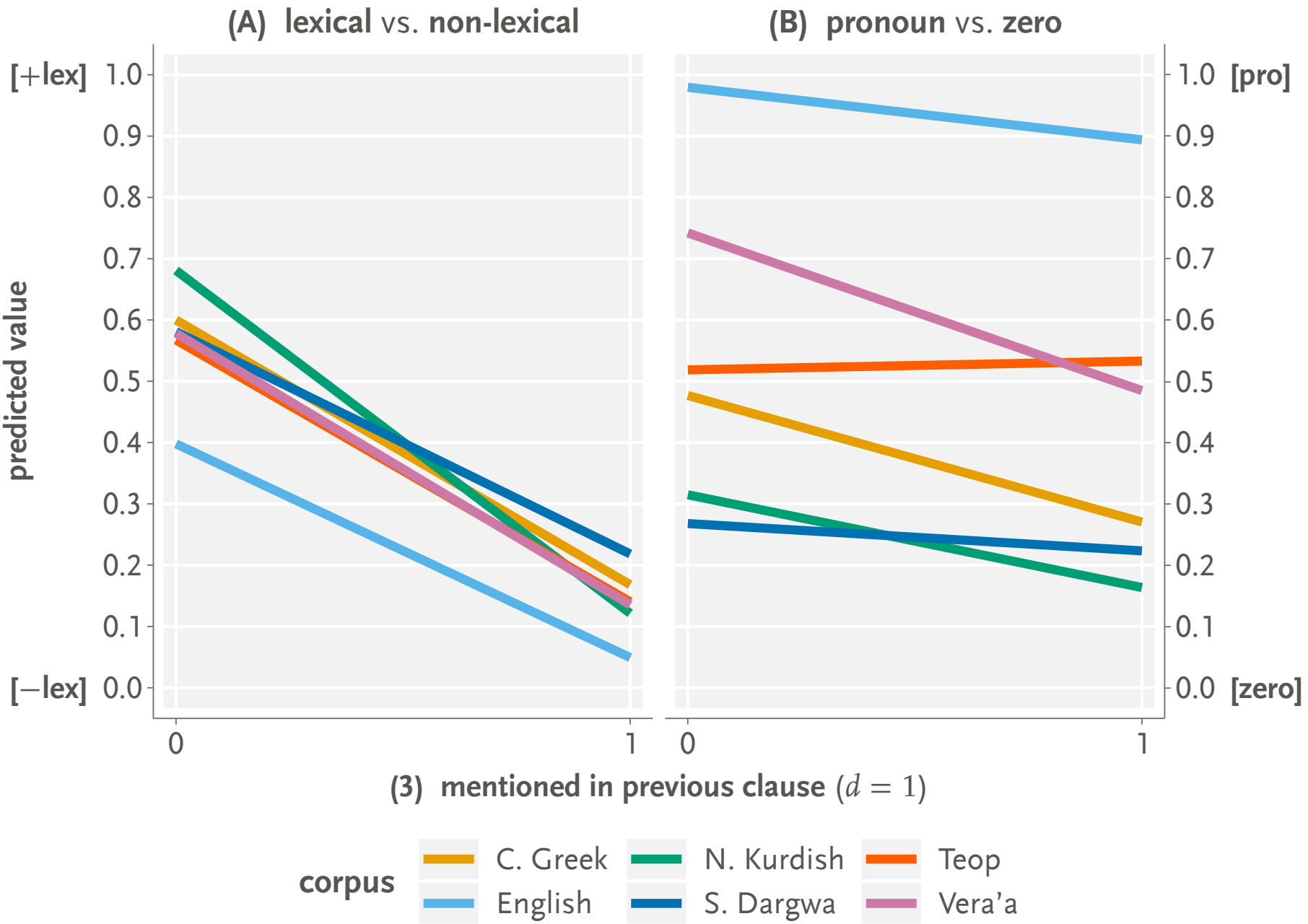

(A) lexical vs. non-lexical
[+lex]
[−lex]
predicted value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
5
10
15
(1) antecedent distance in clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


(A) lexical vs. non-lexical
(B) pronoun vs. zero
[+lex]
[−lex]
predicted value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
[pro]
[zero]
1
5
10
15
(1) antecedent distance in clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


(A) lexical vs. non-lexical
(B) pronoun vs. zero
[+lex]
[−lex]
[pro]
[zero]
predicted value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
(2a) frequency in previous 5 clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


(A) lexical vs. non-lexical
(B) pronoun vs. zero
[+lex]
[−lex]
[pro]
[zero]
predicted value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0 2 4 6 8 10 12 14 16 18 20 22
(2b) frequency in previous 15 clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


(A) lexical vs. non-lexical
(B) pronoun vs. zero
[+lex]
[−lex]
[pro]
[zero]
predicted value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
(2c) frequency since start of text (log scale)
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


(A) lexical vs. non-lexical
(B) pronoun vs. zero
[+lex]
[−lex]
[pro]
[zero]
predicted value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
(3) mentioned in previous clause ($d = 1$)
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a

Summary

in a sample of narrative data from six languages,

the properties of the preceding discourse

- explain the broad distinction between **lexical** and **non-lexical expressions** reasonably well,
- but largely fail to do so for the distinction between **pronouns** and **zero**

Conclusions

in essence,

the data do not corroborrate initial assumption of all types of referring expression being selected based on **different thresholds of the same criteria**

Conclusions

for the selection of zero over pronominal NPs,
factors outside of discourse are at play, e.g.

- morphosyntax (e.g. number),
- agreement paradigms, entrenchment,
- humanness, 'avoid non-human zero' (Genetti & Crain 2003),
- priming, esp. with subjects,
- prosody (cf. Fretheim 1996; Mithun 1996),
- etc.

see also variationist studies on pronoun use
(e.g. Travis & Torres Cacoullos 2012; Meyerhoff & Walker 2015)

Going forward

not just hypothesis-testing,

but creation of cross-linguistic, “bottom-up” models of referential choice

further development and refinement of corpora and quantitative methods

all data will in the near future be freely available online at

Multi-CAST Multilingual Corpus of Annotated Spoken Texts

https://lac2.uni-koeln.de/multicast/

— *normally at* —

https://lac.uni-koeln.de/multicast/


English
Abkhaz
Hinuq
Tabasaran
Sanzhi Dargwa
Cypriot Greek
Northern Kurdish
Persian
Coastal Balochi
Japanese
Mandarin
Jinghpaw
Burmese
Tondano
Totoli
Sumbawa
Tulil
Teop
Vera'a
Nafsan
Multi-CAST

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.

Ariel, Mira. **2006**. Accessibility theory. In Brown, Keith (ed.), *Encyclopedia of language & linguistics,* 15–18. Amsterdam: Elsevier Science.

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.

Meyerhoff, Miriam & **Walker**, James A. **2015**. Subject and object pronoun use in Bequia (St Vincent and the Grenadines). In Prescod, Paula (ed.), *Language issues in Saint Vincent and the Grenadines,* 67–86. Amsterdam: John Benjamins.

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.

References

Fretheim, Thorstein. **1996**. Accessing contexts with intonation. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility,* 89–112. Amsterdam: John Benjamins.

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

Genetti, Carol & **Crain**, Laura. **2003**. Beyond preferred argument structure: Sentences, pronouns, and given referents in Nepali. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure: Grammar as architecture for function,* 197–223. Amsterdam: John Benjamins.

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.

References

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://lac2.uni-koeln.de/en/multicast/`)

Haig, Geoffrey & **Schnell**, Stefan. **2018**[2015]. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts.* (`https://lac2.uni-koeln.de/en/multicast/`)

Haig, Geoffrey & **Thiele**, Hanna. **2015**. 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.

Mithun, Marianne. **1996**. Prosodic cues to accessibility. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 223–234. Amsterdam: John Benjamins.

References

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

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

Schiborr, Nils N. & **Schnell**, Stefan & **Thiele**, Hanna. **2017**. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines*. Version 1.0. MS. University of Bamberg / University of Melbourne.

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

Travis, Catherine & **Torres Cacoullos**, Rena. **2012**. What do subject pronouns do in discourse? Cognitve, mechanical and constructional factors in variation. *Cognitive Linguistics* 23(4). 711–748.

Addendum A

Summaries of regression models

logistic regression model:

(1) expression ~ antecedent distance

choice	corpus	obs(0)	obs(1)	model p	C	coeff. χ^2	coeff. p
±**lex**	cypgreek	587	368	<0.00001	0.771	127.97	<0.00001
±**lex**	english	492	184	<0.00001	0.812	114.30	<0.00001
±**lex**	nkurd	700	480	<0.00001	0.824	181.99	<0.00001
±**lex**	sanzhi	725	538	<0.00001	0.751	161.30	<0.00001
±**lex**	teop	844	484	<0.00001	0.778	199.89	<0.00001
±**lex**	veraa	1549	903	<0.00001	0.790	422.89	<0.00001
±**pro**	cypgreek	390	197	0.00110	0.602	10.02	0.00155
±**pro**	english	38	454	0.01003	0.660	3.50	0.06136
±**pro**	nkurd	556	144	0.06474	0.587	3.68	0.05506
±**pro**	sanzhi	550	175	0.05215	0.543	3.97	0.04633
±**pro**	teop	398	446	0.16305	0.501	1.90	0.16769
±**pro**	veraa	681	868	<0.00001	0.617	47.78	<0.00001

logistic regression model:

(2a) expression ~ recent frequency (5 clauses)

choice	corpus	obs(0)	obs(1)	model p	C	coeff. χ^2	coeff. p
±lex	cypgreek	587	368	<0.00001	0.739	121.19	<0.00001
±lex	english	492	184	<0.00001	0.781	88.69	<0.00001
±lex	nkurd	700	480	<0.00001	0.811	236.23	<0.00001
±lex	sanzhi	725	538	<0.00001	0.760	198.78	<0.00001
±lex	teop	844	484	<0.00001	0.771	206.75	<0.00001
±lex	veraa	1549	903	<0.00001	0.809	456.73	<0.00001
±pro	cypgreek	390	197	0.00044	0.585	11.81	0.00059
±pro	english	38	454	0.02233	0.626	5.55	0.01853
±pro	nkurd	556	144	0.00621	0.570	7.27	0.00703
±pro	sanzhi	550	175	0.00063	0.581	11.20	0.00082
±pro	teop	398	446	0.77061	0.515	0.09	0.77060
±pro	veraa	681	868	0.32722	0.519	0.96	0.32722

logistic regression model:

(2b) expression ~ recent frequency (15 clauses)

choice	corpus	obs(0)	obs(1)	model p	C	coeff. χ^2	coeff. p
±lex	cypgreek	587	368	<0.00001	0.703	75.43	<0.00001
±lex	english	492	184	<0.00001	0.729	49.62	<0.00001
±lex	nkurd	700	480	<0.00001	0.774	175.66	<0.00001
±lex	sanzhi	725	538	<0.00001	0.722	144.24	<0.00001
±lex	teop	844	484	<0.00001	0.738	151.40	<0.00001
±lex	veraa	1549	903	<0.00001	0.787	395.59	<0.00001
±pro	cypgreek	390	197	0.02848	0.551	4.71	0.02999
±pro	english	38	454	0.21058	0.561	1.66	0.19737
±pro	nkurd	556	144	0.61956	0.512	0.25	0.62041
±pro	sanzhi	550	175	0.00183	0.573	9.26	0.00234
±pro	teop	398	446	0.40993	0.523	0.68	0.41005
±pro	veraa	681	868	0.24941	0.517	1.32	0.24997

logistic regression model:

(2c) expression ~ frequency since start of text (log scale)

choice	corpus	obs(0)	obs(1)	model p	C	coeff. χ^2	coeff. p
±**lex**	cypgreek	587	368	<0.00001	0.617	35.58	<0.00001
±**lex**	english	492	184	0.00001	0.617	18.57	0.00002
±**lex**	nkurd	700	480	<0.00001	0.696	125.58	<0.00001
±**lex**	sanzhi	725	538	<0.00001	0.617	48.39	<0.00001
±**lex**	teop	844	484	<0.00001	0.680	113.02	<0.00001
±**lex**	veraa	1549	903	<0.00001	0.652	167.10	<0.00001
±**pro**	cypgreek	390	197	0.15899	0.542	1.98	0.15905
±**pro**	english	38	454	0.42880	0.552	0.64	0.42489
±**pro**	nkurd	556	144	0.29621	0.524	1.10	0.29463
±**pro**	sanzhi	550	175	0.12121	0.539	2.40	0.12127
±**pro**	teop	398	446	0.08530	0.549	2.95	0.08581
±**pro**	veraa	681	868	0.00003	0.552	17.02	0.00004

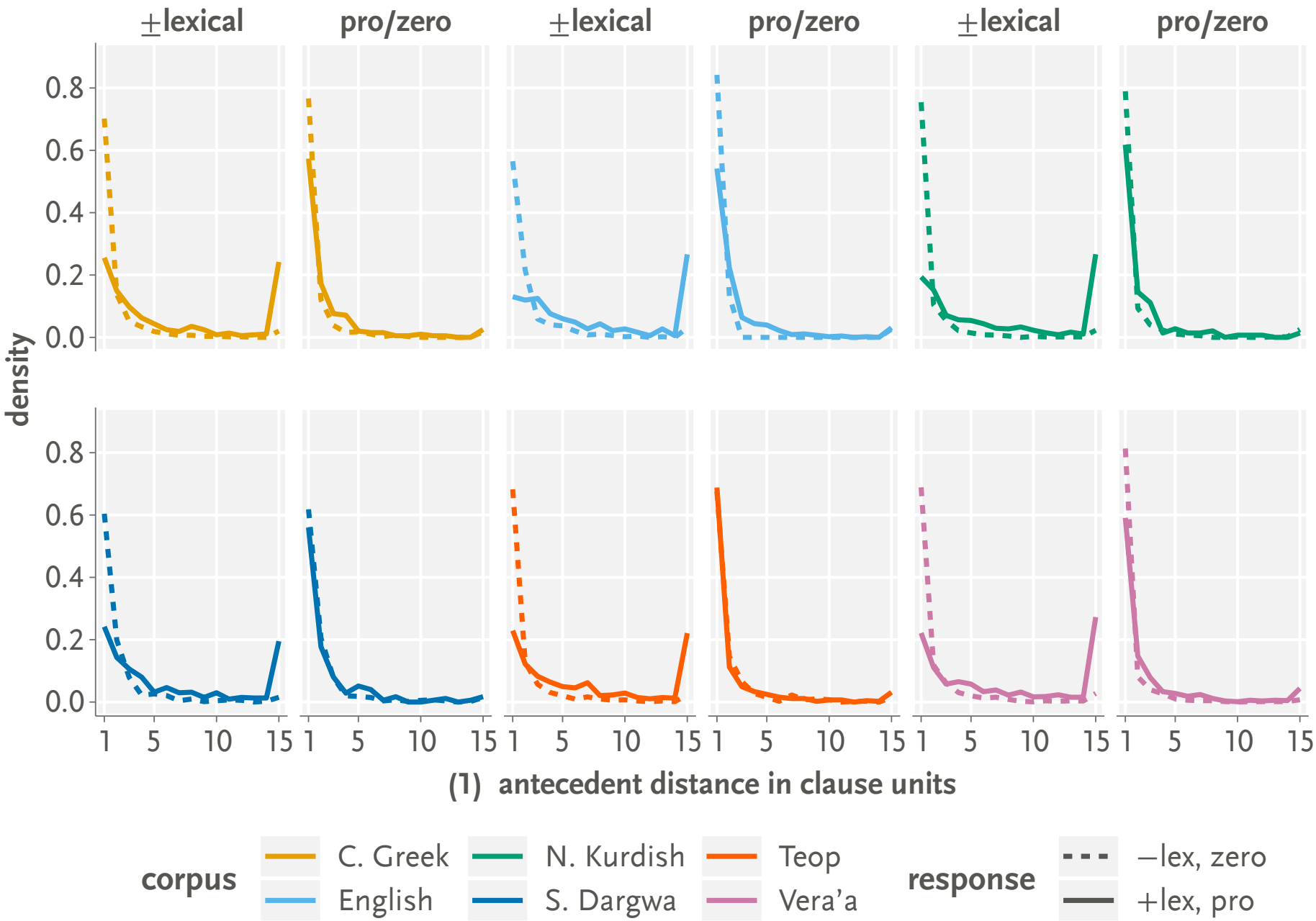
logistic regression model:

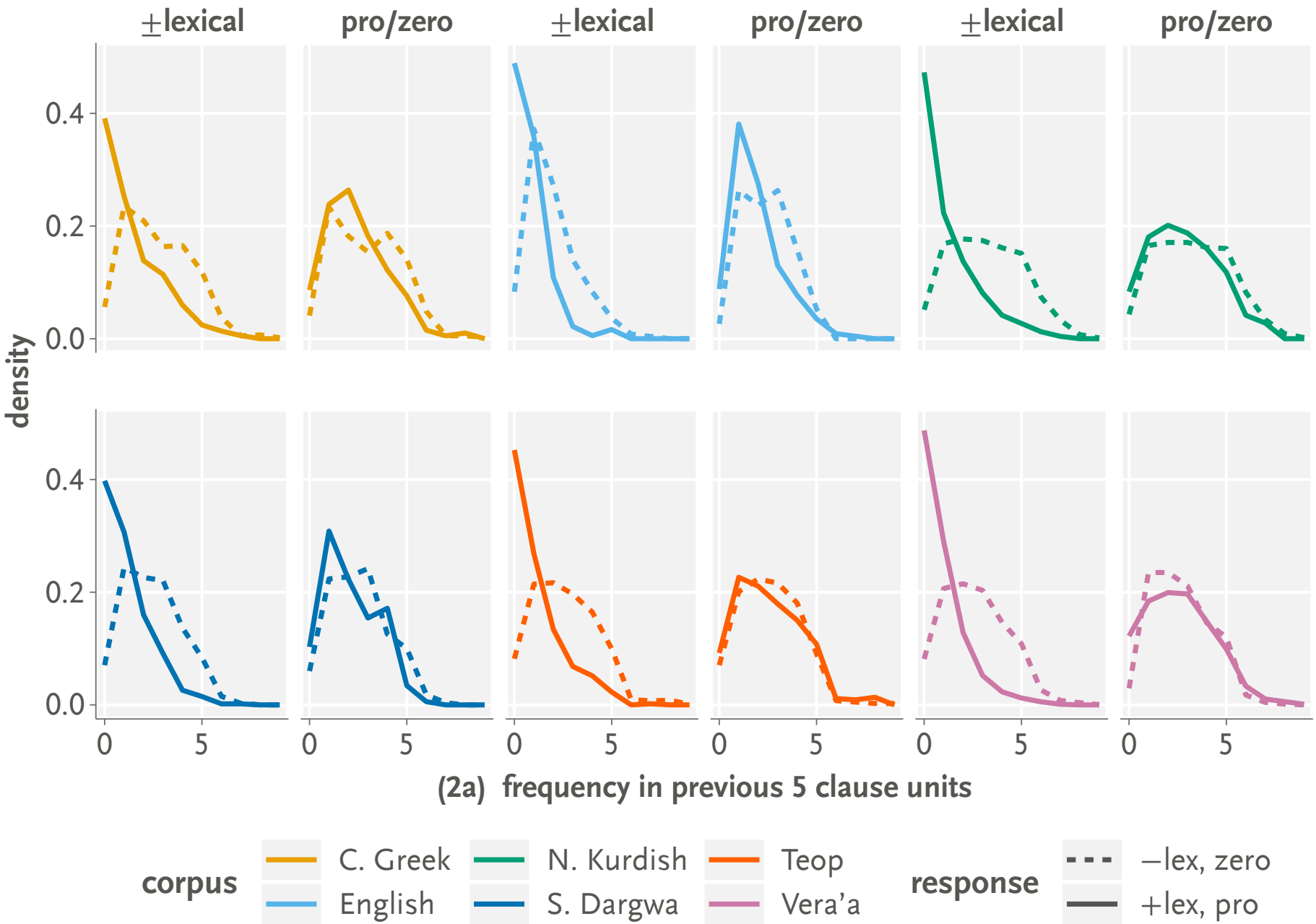
(3) expression ~ mentioned in previous clause

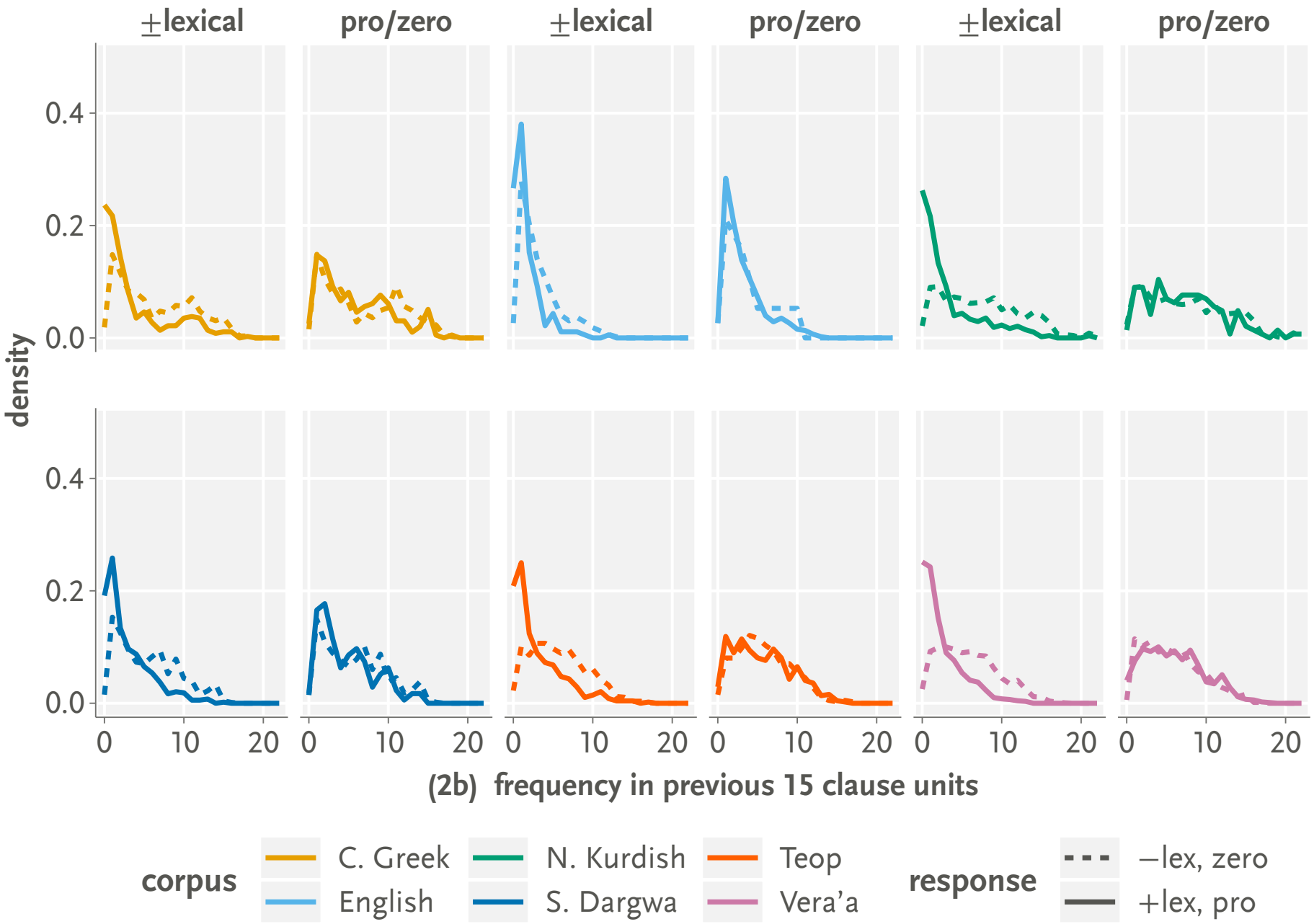
choice	corpus	obs(0)	obs(1)	model p	C	coeff. χ^2	coeff. p
±**lex**	cypgreek	587	368	<0.00001	0.723	165.40	<0.00001
±**lex**	english	492	184	<0.00001	0.717	82.94	<0.00001
±**lex**	nkurd	700	480	<0.00001	0.780	308.36	<0.00001
±**lex**	sanzhi	725	538	<0.00001	0.681	154.22	<0.00001
±**lex**	teop	844	484	<0.00001	0.727	227.84	<0.00001
±**lex**	veraa	1549	903	<0.00001	0.734	446.19	<0.00001
±**pro**	cypgreek	390	197	<0.00001	0.597	22.73	<0.00001
±**pro**	english	38	454	0.00015	0.650	10.97	0.00093
±**pro**	nkurd	556	144	0.00004	0.586	17.58	0.00003
±**pro**	sanzhi	550	175	0.17205	0.529	1.87	0.17091
±**pro**	teop	398	446	0.69791	0.506	0.15	0.69787
±**pro**	veraa	681	868	<0.00001	0.611	84.50	<0.00001

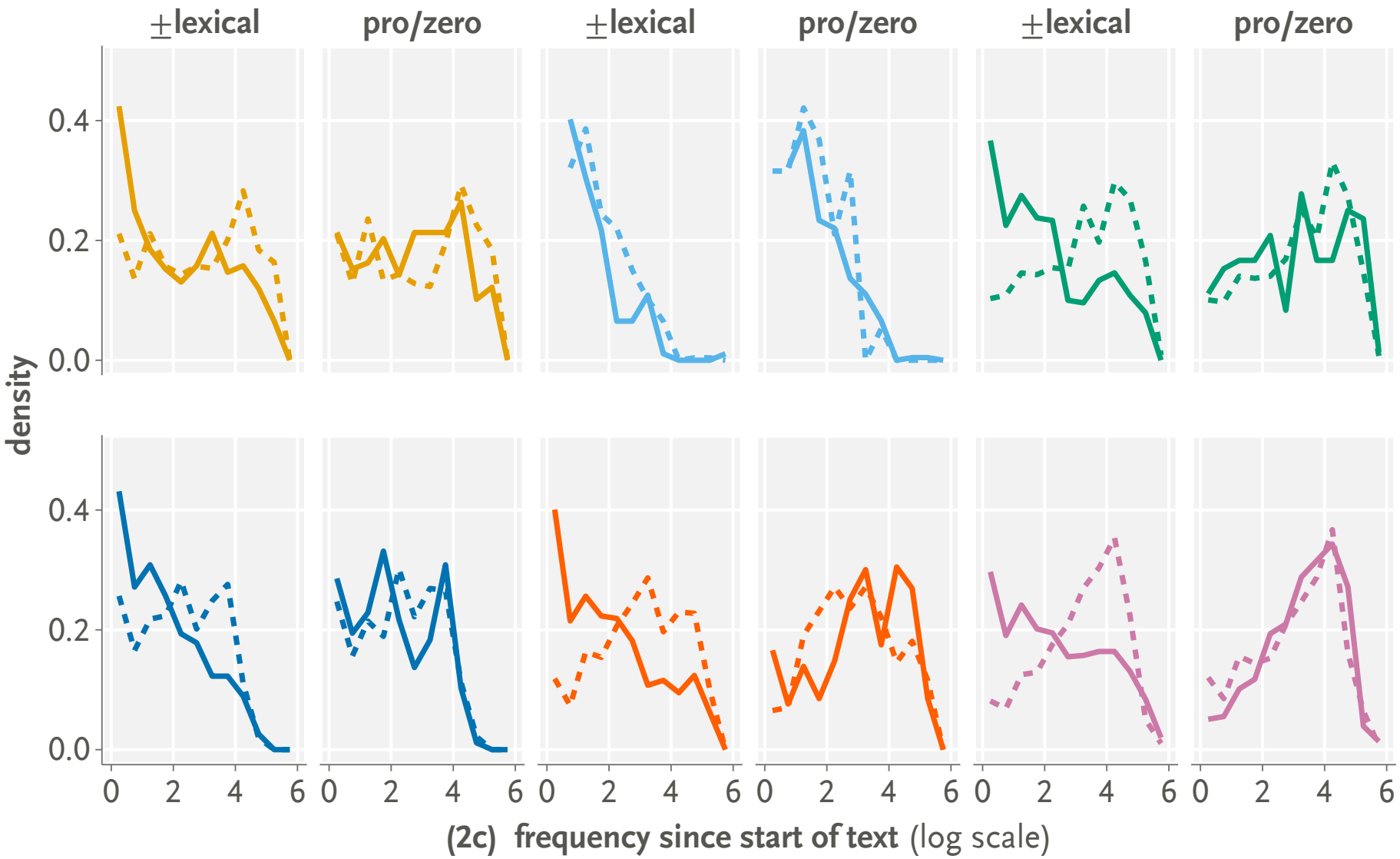
Addendum B

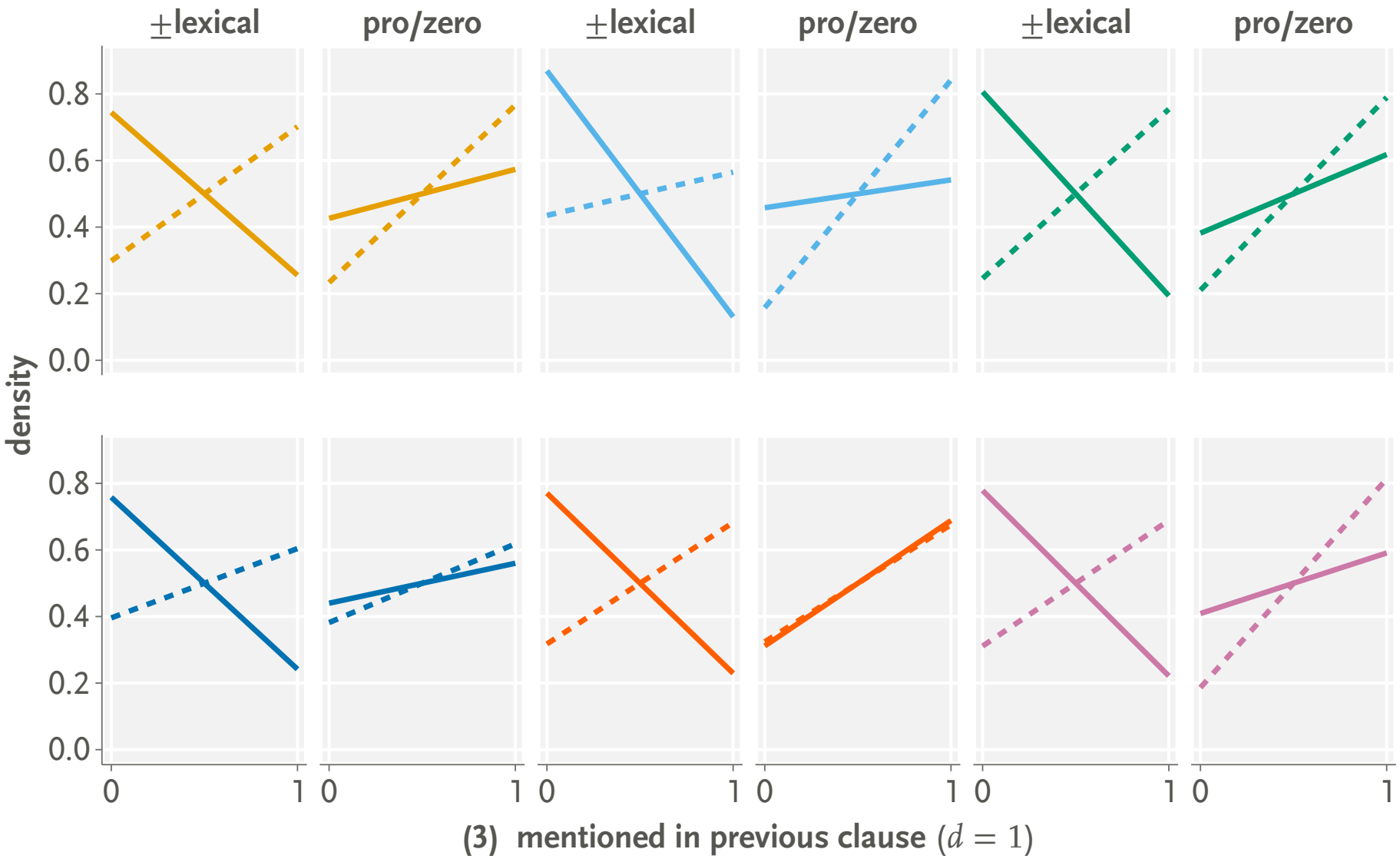
Raw data distributions


±lexical
pro/zero
±lexical
pro/zero
±lexical
pro/zero
density
0.8
0.6
0.4
0.2
0.0
1
5
10
15
(1) antecedent distance in clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a
response
–lex, zero
+lex, pro


±lexical
pro/zero
±lexical
pro/zero
±lexical
pro/zero
density
0.4
0.2
0.0
0
5
(2a) frequency in previous 5 clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a
response
–lex, zero
+lex, pro


±lexical
pro/zero
±lexical
pro/zero
±lexical
pro/zero
density
0.4
0.2
0.0
0
10
20
(2b) frequency in previous 15 clause units
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a
response
–lex, zero
+lex, pro


±lexical
pro/zero
±lexical
pro/zero
±lexical
pro/zero
density
0.4
0.2
0.0
0 2 4 6
(2c) frequency since start of text (log scale)
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a
response
–lex, zero
+lex, pro


±lexical
pro/zero
±lexical
pro/zero
±lexical
pro/zero
density
0.0
0.2
0.4
0.6
0.8
0
1
(3) mentioned in previous clause ($d = 1$)
corpus
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a
response
–lex, zero
+lex, pro