Data-driven models of referential choice
Antecedent distance and beyond

Nils Norman Schiborr
University of Bamberg

7 December 2018
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.
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

“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)
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
(4) **Sanzhi Dargwa**  [sanzhi_devil_034]

\[
xun-ne-b \quad suk \quad b-ič-ib \quad k:urt:a
\]

road-**SPR-N**  Ø  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.’
(5) **Sanzhi Dargwa**  [sanzhi_devil_038]

```
k:urt:a-l  b-ič:-ib  hel-i-j  cin-na  ʋež
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

<table>
<thead>
<tr>
<th>corpus</th>
<th>clauses</th>
<th>referents*</th>
<th>mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cypriot Greek</td>
<td>1 071</td>
<td>165</td>
<td>1 653</td>
</tr>
<tr>
<td>English</td>
<td>1 244</td>
<td>261</td>
<td>1 452</td>
</tr>
<tr>
<td>Northern Kurdish</td>
<td>1 389</td>
<td>153</td>
<td>2 167</td>
</tr>
<tr>
<td>Sanzhi Dargwa</td>
<td>1 618</td>
<td>275</td>
<td>1 977</td>
</tr>
<tr>
<td>Teop</td>
<td>1 272</td>
<td>180</td>
<td>1 688</td>
</tr>
<tr>
<td>Vera’a</td>
<td>2 377</td>
<td>241</td>
<td>3 158</td>
</tr>
<tr>
<td><strong>totals</strong></td>
<td><strong>8 871</strong></td>
<td><strong>1 175</strong></td>
<td><strong>12 095</strong></td>
</tr>
</tbody>
</table>

* with \( n \geq 2 \) mentions
(A) lexical vs. non-lexical

(1) antecedent distance in clause units

predicted value

[+lex] 1.0
[−lex] 0.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(1) antecedent distance in clause units

[+lex] 1.0
[−lex] 0.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

predicted value

corpus

C. Greek
N. Kurdish
Teop
English
S. Dargwa
Vera’a
(A) lexical vs. non-lexical

(B) pronoun vs. zero

(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

predicted value

[+lex] 1.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
[−lex] 0.0

(2a) frequency in previous 5 clause units

corpus

C. Greek
N. Kurdish
Teop
English
S. Dargwa
Vera’a
(A) lexical vs. non-lexical

(B) pronoun vs. zero

(2b) frequency in previous 15 clause units

corpus

- C. Greek
- N. Kurdish
- Teop
- English
- S. Dargwa
- Vera’a
(A) lexical vs. non-lexical

(B) pronoun vs. zero

predicted value

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

corpus

C. Greek  N. Kurdish  Teop
English  S. Dargwa  Vera’a
(A) lexical vs. non-lexical

(B) pronoun vs. zero
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 corroborate** 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 Cacoulls 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
Sanzhi Dargwa
Cypriot Greek
Northern Kurdish
Persian
Coastal Balochi
Abkhaz
Hinuq
Tabasaran
Sanzhi Dargwa
Japanese
Mandarin
Jinghpaw
Burmese
Tondano
Tulil
Totoli
Sumbawa
Vera’a
Nafsan
Teop
References


References


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


References

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


References


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

Addendum A
Summaries of regression models
logistic regression model:

(1) expression $\sim$ antecedent distance

<table>
<thead>
<tr>
<th>choice</th>
<th>corpus</th>
<th>obs(0)</th>
<th>obs(1)</th>
<th>model $p$</th>
<th>C</th>
<th>coeff. $\chi^2$</th>
<th>coeff. $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pm$lex</td>
<td>cypgreek</td>
<td>587</td>
<td>368</td>
<td>$&lt;0.00001$</td>
<td>0.771</td>
<td>127.97</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$lex</td>
<td>english</td>
<td>492</td>
<td>184</td>
<td>$&lt;0.00001$</td>
<td>0.812</td>
<td>114.30</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$lex</td>
<td>nkurd</td>
<td>700</td>
<td>480</td>
<td>$&lt;0.00001$</td>
<td>0.824</td>
<td>181.99</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$lex</td>
<td>sanzhi</td>
<td>725</td>
<td>538</td>
<td>$&lt;0.00001$</td>
<td>0.751</td>
<td>161.30</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$lex</td>
<td>teop</td>
<td>844</td>
<td>484</td>
<td>$&lt;0.00001$</td>
<td>0.778</td>
<td>199.89</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$lex</td>
<td>veraa</td>
<td>1549</td>
<td>903</td>
<td>$&lt;0.00001$</td>
<td>0.790</td>
<td>422.89</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$pro</td>
<td>cypgreek</td>
<td>390</td>
<td>197</td>
<td>0.00110</td>
<td>0.602</td>
<td>10.02</td>
<td>0.00155</td>
</tr>
<tr>
<td>$\pm$pro</td>
<td>english</td>
<td>38</td>
<td>454</td>
<td>0.01003</td>
<td>0.660</td>
<td>3.50</td>
<td>0.06136</td>
</tr>
<tr>
<td>$\pm$pro</td>
<td>nkurd</td>
<td>556</td>
<td>144</td>
<td>0.06474</td>
<td>0.587</td>
<td>3.68</td>
<td>0.05506</td>
</tr>
<tr>
<td>$\pm$pro</td>
<td>sanzhi</td>
<td>550</td>
<td>175</td>
<td>0.05215</td>
<td>0.543</td>
<td>3.97</td>
<td>0.04633</td>
</tr>
<tr>
<td>$\pm$pro</td>
<td>teop</td>
<td>398</td>
<td>446</td>
<td>0.16305</td>
<td>0.501</td>
<td>1.90</td>
<td>0.16769</td>
</tr>
<tr>
<td>$\pm$pro</td>
<td>veraa</td>
<td>681</td>
<td>868</td>
<td>$&lt;0.00001$</td>
<td>0.617</td>
<td>47.78</td>
<td>$&lt;0.00001$</td>
</tr>
</tbody>
</table>
logistic regression model:

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

<table>
<thead>
<tr>
<th>choice</th>
<th>corpus</th>
<th>obs(0)</th>
<th>obs(1)</th>
<th>model p</th>
<th>C</th>
<th>coeff. $\chi^2$</th>
<th>coeff. p</th>
</tr>
</thead>
<tbody>
<tr>
<td>±lex</td>
<td>cypgreek</td>
<td>587</td>
<td>368</td>
<td>&lt;0.00001</td>
<td>0.739</td>
<td>121.19</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>±lex</td>
<td>english</td>
<td>492</td>
<td>184</td>
<td>&lt;0.00001</td>
<td>0.781</td>
<td>88.69</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>±lex</td>
<td>nkurd</td>
<td>700</td>
<td>480</td>
<td>&lt;0.00001</td>
<td>0.811</td>
<td>236.23</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>±lex</td>
<td>sanzhi</td>
<td>725</td>
<td>538</td>
<td>&lt;0.00001</td>
<td>0.760</td>
<td>198.78</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>±lex</td>
<td>teop</td>
<td>844</td>
<td>484</td>
<td>&lt;0.00001</td>
<td>0.771</td>
<td>206.75</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>±lex</td>
<td>veraa</td>
<td>1549</td>
<td>903</td>
<td>&lt;0.00001</td>
<td>0.809</td>
<td>456.73</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>±pro</td>
<td>cypgreek</td>
<td>390</td>
<td>197</td>
<td>0.00044</td>
<td>0.585</td>
<td>11.81</td>
<td>0.00059</td>
</tr>
<tr>
<td>±pro</td>
<td>english</td>
<td>38</td>
<td>454</td>
<td>0.02233</td>
<td>0.626</td>
<td>5.55</td>
<td>0.01853</td>
</tr>
<tr>
<td>±pro</td>
<td>nkurd</td>
<td>556</td>
<td>144</td>
<td>0.00621</td>
<td>0.570</td>
<td>7.27</td>
<td>0.00703</td>
</tr>
<tr>
<td>±pro</td>
<td>sanzhi</td>
<td>550</td>
<td>175</td>
<td>0.00063</td>
<td>0.581</td>
<td>11.20</td>
<td>0.00082</td>
</tr>
<tr>
<td>±pro</td>
<td>teop</td>
<td>398</td>
<td>446</td>
<td>0.77061</td>
<td>0.515</td>
<td>0.09</td>
<td>0.77060</td>
</tr>
<tr>
<td>±pro</td>
<td>veraa</td>
<td>681</td>
<td>868</td>
<td>0.32722</td>
<td>0.519</td>
<td>0.96</td>
<td>0.32722</td>
</tr>
</tbody>
</table>
logistic regression model:
(2b) expression ~ recent frequency (15 clauses)

<table>
<thead>
<tr>
<th>choice</th>
<th>corpus</th>
<th>obs(0)</th>
<th>obs(1)</th>
<th>model p</th>
<th>C</th>
<th>coeff. $\chi^2$</th>
<th>coeff. p</th>
</tr>
</thead>
<tbody>
<tr>
<td>± lex</td>
<td>cypgreek</td>
<td>587</td>
<td>368</td>
<td>&lt;0.00001</td>
<td>0.703</td>
<td>75.43</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>± lex</td>
<td>english</td>
<td>492</td>
<td>184</td>
<td>&lt;0.00001</td>
<td>0.729</td>
<td>49.62</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>± lex</td>
<td>nkurd</td>
<td>700</td>
<td>480</td>
<td>&lt;0.00001</td>
<td>0.774</td>
<td>175.66</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>± lex</td>
<td>sanzhi</td>
<td>725</td>
<td>538</td>
<td>&lt;0.00001</td>
<td>0.722</td>
<td>144.24</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>± lex</td>
<td>teop</td>
<td>844</td>
<td>484</td>
<td>&lt;0.00001</td>
<td>0.738</td>
<td>151.40</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>± lex</td>
<td>veara</td>
<td>1549</td>
<td>903</td>
<td>&lt;0.00001</td>
<td>0.787</td>
<td>395.59</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>± pro</td>
<td>cypgreek</td>
<td>390</td>
<td>197</td>
<td>0.02848</td>
<td>0.551</td>
<td>4.71</td>
<td>0.02999</td>
</tr>
<tr>
<td>± pro</td>
<td>english</td>
<td>38</td>
<td>454</td>
<td>0.21058</td>
<td>0.561</td>
<td>1.66</td>
<td>0.19737</td>
</tr>
<tr>
<td>± pro</td>
<td>nkurd</td>
<td>556</td>
<td>144</td>
<td>0.61956</td>
<td>0.512</td>
<td>0.25</td>
<td>0.62041</td>
</tr>
<tr>
<td>± pro</td>
<td>sanzhi</td>
<td>550</td>
<td>175</td>
<td>0.00183</td>
<td>0.573</td>
<td>9.26</td>
<td>0.00234</td>
</tr>
<tr>
<td>± pro</td>
<td>teop</td>
<td>398</td>
<td>446</td>
<td>0.40993</td>
<td>0.523</td>
<td>0.68</td>
<td>0.41005</td>
</tr>
<tr>
<td>± pro</td>
<td>veara</td>
<td>681</td>
<td>868</td>
<td>0.24941</td>
<td>0.517</td>
<td>1.32</td>
<td>0.24997</td>
</tr>
</tbody>
</table>
logistic regression model:
(2c) expression ~ frequency since start of text (log scale)

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

(3) expression ~ mentioned in previous clause

<table>
<thead>
<tr>
<th>choice</th>
<th>corpus</th>
<th>obs(0)</th>
<th>obs(1)</th>
<th>model  $p$</th>
<th>C</th>
<th>coeff. $\chi^2$</th>
<th>coeff. $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pm$ lex</td>
<td>cypgreek</td>
<td>587</td>
<td>368</td>
<td>$&lt;0.00001$</td>
<td>0.723</td>
<td>165.40</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$ lex</td>
<td>english</td>
<td>492</td>
<td>184</td>
<td>$&lt;0.00001$</td>
<td>0.717</td>
<td>82.94</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$ lex</td>
<td>nkurd</td>
<td>700</td>
<td>480</td>
<td>$&lt;0.00001$</td>
<td>0.780</td>
<td>308.36</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$ lex</td>
<td>sanzhi</td>
<td>725</td>
<td>538</td>
<td>$&lt;0.00001$</td>
<td>0.681</td>
<td>154.22</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$ lex</td>
<td>teop</td>
<td>844</td>
<td>484</td>
<td>$&lt;0.00001$</td>
<td>0.727</td>
<td>227.84</td>
<td>$&lt;0.00001$</td>
</tr>
<tr>
<td>$\pm$ lex</td>
<td>veraa</td>
<td>1549</td>
<td>903</td>
<td>$&lt;0.00001$</td>
<td>0.734</td>
<td>446.19</td>
<td>$&lt;0.00001$</td>
</tr>
</tbody>
</table>

| $\pm$ pro | cypgreek  | 390   | 197    | $<0.00001$ | 0.597  | 22.73           | $<0.00001$ |
| $\pm$ pro | english   | 38    | 454    | 0.00015   | 0.650  | 10.97           | 0.00093    |
| $\pm$ pro | nkurd     | 556   | 144    | 0.00004   | 0.586  | 17.58           | 0.00003    |
| $\pm$ pro | sanzhi    | 550   | 175    | 0.17205   | 0.529  | 1.87            | 0.17091    |
| $\pm$ pro | teop      | 398   | 446    | 0.69791   | 0.506  | 0.15            | 0.69787    |
| $\pm$ pro | veraa     | 681   | 868    | $<0.00001$ | 0.611  | 84.50           | $<0.00001$ |
Addendum B
Raw data distributions
Referential choice

(1) antecedent distance in clause units

corpus

<table>
<thead>
<tr>
<th></th>
<th>C. Greek</th>
<th>N. Kurdish</th>
<th>Teop</th>
<th>S. Dargwa</th>
<th>Vera’a</th>
</tr>
</thead>
<tbody>
<tr>
<td>density</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
</tbody>
</table>

response

<table>
<thead>
<tr>
<th></th>
<th>–lex, zero</th>
<th>+lex, pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>density</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>
(2a) frequency in previous 5 clause units

corpus

- C. Greek
- English
- N. Kurdish
- S. Dargwa

response

- −lex, zero
- +lex, pro
(2b) frequency in previous 15 clause units

corpus

- C. Greek
- English
- N. Kurdish
- S. Dargwa

response

- −lex, zero
- +lex, pro
Figure 3: Graph showing the relationship between density and the mention of an element in the previous clause. The x-axis represents the density of mention in the previous clause, and the y-axis represents the density of mention in the current clause. The lines represent different corpora: English, Greek, Kurdish, S. Dargwa, Teop, and Vera’a. The responses are shown for different lexical and pro/zero choices.

Legend:
- C. Greek
- N. Kurdish
- S. Dargwa
- Teop
- Vera’a

- lex, zero
- +lex, pro