Quantitative models of referential choice
Lexical anaphora in English

Nils Norman Schiborr
University of Bamberg

26 September 2019
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

- 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

<table>
<thead>
<tr>
<th>corpus</th>
<th>clause units</th>
<th>unique referents</th>
<th>sampled subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>4184</td>
<td>509</td>
<td>1345</td>
</tr>
<tr>
<td>Cypriot Greek</td>
<td>1071</td>
<td>99</td>
<td>441</td>
</tr>
<tr>
<td>Mandarin</td>
<td>1197</td>
<td>109</td>
<td>715</td>
</tr>
<tr>
<td>Nafsan</td>
<td>1012</td>
<td>118</td>
<td>692</td>
</tr>
<tr>
<td>Northern Kurdish</td>
<td>1359</td>
<td>120</td>
<td>642</td>
</tr>
<tr>
<td>Sanzhi Dargwa</td>
<td>1066</td>
<td>103</td>
<td>475</td>
</tr>
<tr>
<td>Teop</td>
<td>1302</td>
<td>101</td>
<td>771</td>
</tr>
<tr>
<td>Tulil</td>
<td>1264</td>
<td>148</td>
<td>590</td>
</tr>
<tr>
<td>Vera’a</td>
<td>3608</td>
<td>293</td>
<td>2422</td>
</tr>
<tr>
<td><strong>totals</strong></td>
<td><strong>14866</strong></td>
<td><strong>1600</strong></td>
<td><strong>8093</strong></td>
</tr>
</tbody>
</table>
n = 3762

n = 2685

n = 1646

% of subjects

referring expression

zero anaphora

free pronoun

full NP

English

C. Greek

Nafsan

S. Dargwa

Tulil

Mandarin

N. Kurdish

Teop

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
transitivity of predicate

- English
- C. Greek
- Naksan
- Mandarin
- N. Kurdish
- S. Dargwa
- Tulil
- Teop
- Vera’a

n = 2890
n = 5203

% expressed lexically
The figure shows the percentage of referents expressed lexically across different languages and the number of other referents mentioned in the last 3 clauses. The languages included are English, C. Greek, Nafsan, Mandarin, N. Kurdish, S. Dargwa, Tulil, Teop, and Vera’a. The chart indicates that the percentage of referents expressed lexically decreases as the number of other referents increases, with English consistently showing the highest percentage across all clause counts.
### Correlation coefficients

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

<table>
<thead>
<tr>
<th>factor</th>
<th>8 corpora</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>referent is human</td>
<td>0.887</td>
<td>0.275</td>
</tr>
<tr>
<td>clause is transitive</td>
<td>0.870</td>
<td>0.348</td>
</tr>
<tr>
<td>antecedent distance</td>
<td>0.843</td>
<td>0.885</td>
</tr>
<tr>
<td>information pressure</td>
<td>0.557</td>
<td>0.749</td>
</tr>
<tr>
<td>antecedent is lexical</td>
<td>0.894</td>
<td>0.747</td>
</tr>
<tr>
<td>antecedent is subject</td>
<td>-0.840</td>
<td>-0.453</td>
</tr>
</tbody>
</table>
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:
  - pronouns 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 idiosyncrasies

- 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?
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

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


References


References


