Modelling referential choice in natural spoken discourse
Multi-CAST, GRAID, and RefIND

Nils Norman Schiborr
University of Bamberg

16 September 2019
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 1000 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**
Annotation of non-standard corpora — Referential choice and Multi-CAST — Nils Norman Schiborr — University of Bamberg

- Cypriot Greek
- Northern Kurdish
- Persian
- Coastal Balochi
- Sanzhi Dargwa
- Jinghpaw
- Sumbawa
- Arta
- Tondano
- Tulil
- Teop
- Vera'a
- Nafsan
- Mandarin
- Ainu
- Japanese
- Bora
- Aguaruna

Available in preparation corpus — available in preparation

Tabasaran
<table>
<thead>
<tr>
<th>language</th>
<th>affiliation</th>
<th>citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arta</td>
<td>Austronesian, Polynesian</td>
<td>(Kimoto 2019)</td>
</tr>
<tr>
<td>Cypriot Greek</td>
<td>I.E., Greek</td>
<td>(Hadjidas &amp; Vollmer 2015)</td>
</tr>
<tr>
<td>English</td>
<td>I.E., Germanic</td>
<td>(Schiborr 2015)</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>Persian</td>
<td>I.E., Iranian</td>
<td>(Adibifar 2016)</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>Tondano</td>
<td>Austronesian, Polynesian</td>
<td>(Brickell 2016)</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>

*every corpus in the collection is an individually citable resource*
Documentation

- **extensively documented:**
  - what’s in it?
  - how’s it structured?
  - what’s changed?
    - \( \rightarrow \) *collection overview*
  - how’s it annotated?
    - \( \rightarrow \) *guidelines* for basic schemes
    - \( \rightarrow \) *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**
(1) **Nafsan** (Austronesian, Oceanic)

```
kineu  a=  pam  nataŋmol  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

\langle \text{np} . \ h : \ p \rangle

1  2  3

1  full noun phrase  (form)
2  human, third person  (animacy)
3  direct object  (function)
\langle \text{pro} . 1 : a \rangle

1. free definite pronoun (form)
2. human, first person (animacy)
3. subject of a transitive clause (function)
- 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. \langle pro \rangle \rightarrow \langle dem_pro \rangle; \langle s \rangle \rightarrow \langle s_ds \rangle)
- 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)

\[
\begin{align*}
kineu & \quad a= \quad pam \quad nata\text{m}ol & \quad i= \quad tol \quad su \\
1\text{SG} & \quad 1\text{SG.RS}= \quad \text{eat} \quad \text{person} \quad 3\text{SG.RS}= \quad \text{three} \quad \text{PF} \\
\#ds & \quad \text{pro.1:a} \quad =lv \quad \text{v:pred} \quad \text{np.h:p} \quad =rn \quad \text{rn} \quad \text{rv} \\
0026 & \quad 0048
\end{align*}
\]

‘[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
<table>
<thead>
<tr>
<th>corpus</th>
<th>text</th>
<th>uid</th>
<th>gword</th>
<th>gloss</th>
<th>graid</th>
<th>refind</th>
</tr>
</thead>
<tbody>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>#</td>
<td>#</td>
<td>##neg</td>
<td>other</td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>and</td>
<td>and</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>the</td>
<td>the</td>
<td>ln_det</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>house</td>
<td>house</td>
<td>np:dt</td>
<td>0015</td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>#</td>
<td>#</td>
<td>#rc</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>0</td>
<td>0_house</td>
<td>rel_0:g</td>
<td>0015</td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>we</td>
<td>1PL</td>
<td>pro.1:s</td>
<td>0014</td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>come</td>
<td>come.pst</td>
<td>v:pred</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>in</td>
<td>in</td>
<td>adp</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>first</td>
<td>first</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>%</td>
<td>%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>we</td>
<td>1PL</td>
<td>pro.1:s</td>
<td>0014</td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>did-n’t</td>
<td>do.pst-NEG</td>
<td>lv_aux</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>stop</td>
<td>stop.inf</td>
<td>v:pred</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>kent01</td>
<td>0164</td>
<td>long</td>
<td>long</td>
<td>other</td>
<td></td>
</tr>
</tbody>
</table>
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:**
  ???
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.
Referential choice

- referring expressions differ in **informativity** and **specificity** (e.g. zero vs. full NPs)
- speakers need 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

<table>
<thead>
<tr>
<th>corpus</th>
<th>clause units</th>
<th>unique referents</th>
<th>mentions as subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cypriot Greek</td>
<td>1 071</td>
<td>99</td>
<td>441</td>
</tr>
<tr>
<td>English</td>
<td>4 184</td>
<td>509</td>
<td>1 343</td>
</tr>
<tr>
<td>Mandarin*</td>
<td>1 197</td>
<td>109</td>
<td>715</td>
</tr>
<tr>
<td>Nafsan</td>
<td>1 012</td>
<td>118</td>
<td>692</td>
</tr>
<tr>
<td>Northern Kurdish</td>
<td>1 359</td>
<td>120</td>
<td>642</td>
</tr>
<tr>
<td>Sanzhi Dargwa</td>
<td>1 066</td>
<td>103</td>
<td>475</td>
</tr>
<tr>
<td>Teop</td>
<td>1 302</td>
<td>101</td>
<td>771</td>
</tr>
<tr>
<td>Tulil</td>
<td>1 264</td>
<td>148</td>
<td>590</td>
</tr>
<tr>
<td>Vera’a</td>
<td>3 608</td>
<td>293</td>
<td>2 422</td>
</tr>
<tr>
<td><strong>totals</strong></td>
<td><strong>14 866</strong></td>
<td><strong>1 600</strong></td>
<td><strong>8 091</strong></td>
</tr>
</tbody>
</table>

* (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
Humanness of mentions

- **two values:**
  - human or
  - non-human

- regular expressions matching the GRAID annotations:
  - \h \(\rightarrow\) human (+ third person)
  - (non-human third person is unmarked)
  - then filter for first/second person mentions, \h [12]
Annotation of non-standard corpora

Referential choice and Multi-CAST

— Nils Norman Schiborr — University of Bamberg

n = 6736

n = 1355

% mentioned in form

human

non-human

humanness of mentions

- zero
- pronoun
- full NP

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

<table>
<thead>
<tr>
<th>gword</th>
<th>graid</th>
<th>refund</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>pro.1:s</td>
<td>0000</td>
</tr>
<tr>
<td>went</td>
<td>v:pred</td>
<td></td>
</tr>
<tr>
<td>with</td>
<td>adp</td>
<td></td>
</tr>
<tr>
<td>this</td>
<td>ln</td>
<td></td>
</tr>
<tr>
<td>man</td>
<td>np.h:obl</td>
<td>0036</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bla</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td>bla</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he</td>
<td>pro.h:a</td>
<td>0036</td>
</tr>
<tr>
<td>had</td>
<td>v:pred</td>
<td></td>
</tr>
<tr>
<td>horses</td>
<td>np:p</td>
<td>0042</td>
</tr>
</tbody>
</table>
### Antecedent distance

<table>
<thead>
<tr>
<th>gword</th>
<th>graid</th>
<th>refind</th>
<th>clause index</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>pro.1:s</td>
<td>0000</td>
<td>1 ← clause boundary</td>
</tr>
<tr>
<td>went</td>
<td>v:pred</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with</td>
<td>adp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>this</td>
<td>ln</td>
<td></td>
<td></td>
</tr>
<tr>
<td>man</td>
<td>np.h:obl</td>
<td>0036</td>
<td></td>
</tr>
<tr>
<td></td>
<td>##</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bla</td>
<td>other</td>
<td></td>
<td>2 ← clause boundary</td>
</tr>
<tr>
<td>bla</td>
<td>other</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>##</td>
<td></td>
<td></td>
</tr>
<tr>
<td>he</td>
<td>pro.h:a</td>
<td>0036</td>
<td></td>
</tr>
<tr>
<td>had</td>
<td>v:pred</td>
<td></td>
<td></td>
</tr>
<tr>
<td>horses</td>
<td>np:p</td>
<td>0042</td>
<td></td>
</tr>
</tbody>
</table>
### Antecedent distance

<table>
<thead>
<tr>
<th>gword</th>
<th>graid</th>
<th>refind</th>
<th>clause index</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>l</code></td>
<td><code>pro.1:s</code></td>
<td><code>0000</code></td>
<td><code>1</code> ← clause boundary</td>
</tr>
<tr>
<td><code>went</code></td>
<td><code>v:pred</code></td>
<td></td>
<td><code>1</code></td>
</tr>
<tr>
<td><code>with</code></td>
<td><code>adp</code></td>
<td></td>
<td><code>1</code></td>
</tr>
<tr>
<td><code>this</code></td>
<td><code>ln</code></td>
<td></td>
<td><code>1</code></td>
</tr>
<tr>
<td><code>man</code></td>
<td><code>np.h:obl</code></td>
<td><code>0036</code></td>
<td><code>1</code></td>
</tr>
<tr>
<td><code>bla</code></td>
<td><code>other</code></td>
<td></td>
<td><code>2</code></td>
</tr>
<tr>
<td><code>bla</code></td>
<td><code>other</code></td>
<td></td>
<td><code>2</code></td>
</tr>
<tr>
<td><code>he</code></td>
<td><code>pro.h:a</code></td>
<td><code>0036</code></td>
<td><code>3</code> ← clause boundary</td>
</tr>
<tr>
<td><code>had</code></td>
<td><code>v:pred</code></td>
<td></td>
<td><code>3</code></td>
</tr>
<tr>
<td><code>horses</code></td>
<td><code>np:p</code></td>
<td><code>0042</code></td>
<td><code>3</code></td>
</tr>
</tbody>
</table>
## Antecedent distance

<table>
<thead>
<tr>
<th>gword</th>
<th>graid</th>
<th>refind</th>
<th>clause index</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>pro.1:s</td>
<td>0000</td>
<td>1</td>
</tr>
<tr>
<td>went</td>
<td>v:pred</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>with</td>
<td>adp</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>this</td>
<td>ln</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>man</td>
<td>np.h:obl</td>
<td>0036</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>antecedent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>clause boundary</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>bla</td>
<td>other</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>clause boundary</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>he</td>
<td>pro.h:a</td>
<td>0036</td>
<td>3</td>
</tr>
<tr>
<td>had</td>
<td>v:pred</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>horses</td>
<td>np:p</td>
<td>0042</td>
<td>3</td>
</tr>
</tbody>
</table>

- `## 1` 
- `## 2` 
- `## 3` 

### Notes

- `anteecedent`: The antecedent is the word that a pronoun refers back to. In this example, the pronoun `he` refers back to the noun phrase `this man`. 
- `clause boundary`: The clause boundary is the beginning of a new clause. 
- `anaphor @ 3 − 1 = 2 clauses distance`: An anaphor is a word that refers back to a word in the text. In this example, the word `horses` is an anaphor, referring back to the noun phrase `this man`. The distance between the clause containing the anaphor and the clause containing the antecedent is 2 clauses.
% mentioned in form

distance from antecedent in clause units

- zero
- pronoun
- full NP

n = 4955
n = 1080
n = 498
n = 297
n = 200
n = 1061
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**


Hadjidas, Harris & Vollmer, Maria C. 2015. Multi-CAST Cypriot Greek. 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.*

References


