Linguistic considerations regarding field-based corpora

ELDP Training 2008
Friederike Lüpke

Why bother
- So far, the field of language documentation has focused on the shape that a language documentation should take, but not on what data should be included, how they should be collected and to whom they should be of use.
- A step toward this is a systematic investigation of the goals of language documentation, of the data collection methods associated with them, and the usability of the data resulting from them.

I-language vs. E-language
Oral vs. written

Oral vs. nonverbal
Text vs. performance

Quality vs. quantity
...

The (new?) role of data
- "For description, the main concern is the production of grammars and dictionaries whose primary audience are linguists... In these products language data serves essentially as exemplification and support for the linguist's analysis." (Austin 2006: 87)
- [...] Language documentation, on the other hand, places data at the center of its concerns." (Austin 2006: 87)
Documentation = a large, annotated corpus

Corpus (Himmelmann 1998)

- Elicitations
- Observed communicative events
- Staged communicative events

Paradigms, wordlists, results of tests...

Conversations, narratives, folk stories, proverbs...

Descriptions of picture and video stimuli, games...

---

Data types in the corpus

Corpus

- Video, audio and image data
- Written data
- Metadata

Speech in cultural context

Transcription, annotation and analysis of data

Information about the content, format and structure of data

---

But exactly what data?

- "A language documentation [...] conceived of as a lasting, multipurpose record of a language [...] should contain a large set of primary data which provide evidence for the language(s) used at a given time in a given community." (Himmelmann 2006: 7)

- "The main goal of a language documentation is to make primary data available for a broad group of users." (Himmelmann 2006: 15)

Which audience(s)?

Which community/ies? Which language(s)?
Homogeneity vs. heterogeneity

- Do we document a snapshot or the production, transmission, maintenance and change of linguistic and cultural behaviour?
- Do we aim at a monolithic record or at documenting variation (of what)?
- How do we establish representativeness?

Data for who?

- We are aware of the disciplines that also have language as a centre of interest – but do we cater for their needs?
- We want to create data relevant for the speech community/ies, but we have little evidence for the use of our electronic corpora.

- How can we create a true multipurpose record of a language?

The (new?) role of the consultant

- 
  “…some older field manuals give advice on what kind of questions to ask or not to ask, … In this manner, such manuals quite automatically assign a passive role to the speaker. If we regard fieldwork as a mutual teaching-learning event, this approach is no longer acceptable.” (Mosel 2006: 75)

- What roles do we assume for ourselves and our consultants?
The theory: cyclic corpus design

Cycles of corpus design (Biber 1995)

Pilot empirical investigation and theoretical analysis → Corpus design → Compilation of portion of the corpus → Empirical investigation

Sociolinguistic investigation

Study of attitudes

Study of communicative practices

Data resulting from observed communicative events
Data resulting from monologues

"This lecture is about the fascinating theory on..."

- **PRO:**
  - Have a high degree of ecological validity.
  - Yield phonologically, semantically and syntactically natural utterances.
  - Give insight into the culture, if thematically balanced.
  - Show high-frequency phenomena.

- **CON:**
  - Can seem natural but factually aren't because the cultural settings are not respected.
  - Can contain pragmatic oddities.
  - Are not very controlled.
  - Many features are not quantifiable because a unique performance of one speaker.
  - Don't offer negative evidence and are not good for low-frequency phenomena.

Data resulting from conversation

A: "How do you like the ELDP training so far?"
B: "All I can say is they start too early and don't give us enough breaks!"

- **PRO:**
  - Often seen as the non-plus-ultra in naturalness.
  - Yields data that are naturalistic in every respect.
  - Also gives important information about the culture.

- **CON:**
  - Is not controlled at all.
  - Is very difficult to get.
  - Is tedious and time-consuming to transcribe.
  - Is even more time-consuming to analyse.
  - Doesn't offer negative evidence and insight into low-frequency phenomena.

Representativeness of a LDD corpus – Jalonke

high frequency verb kolon 'know'

<table>
<thead>
<tr>
<th>Causative</th>
<th>Reciprocal</th>
<th>Complement</th>
<th>Perfect</th>
<th>Many transitive uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>kolon</td>
<td>kolon</td>
<td>kolon</td>
<td>kolon</td>
<td>kolon</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
<tr>
<td>a m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
<td>a mun m'aa</td>
</tr>
</tbody>
</table>
Representativeness of a LDD corpus – Jalonke
low frequency verb

<table>
<thead>
<tr>
<th>Verb</th>
<th>Mas Tugan, a</th>
<th>Tone</th>
<th>Past</th>
<th>Goal PP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n a n ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zemex ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bimex ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zemex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bimex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n a n ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zemex ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bimex ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zemex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bimex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n a n ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zemex ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bimex ci.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zemex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bimex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>!</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All uses are intransitive

Summary

- Observed communicative events that are investigated in a qualitative way allow to
  - Get a first impression of the most frequent syntactic environments of the most frequent verbs.
  - Formulate hypotheses and prepare elicitation sessions.

But: these data don’t tell us anything about markedness, about the full distributional range, about low frequency items and constructions, and about their semantic properties.

More on other data types after lunch!