1 Corpus typology

Basic concepts

**text**  any piece of natural speech
(story, proverb, conversation, sermon, ...)

**corpus**  “collection of sampled texts, written or spoken,
in machine readable form
which may be annotated with various forms
of linguistic information”  (McEnery et al. 2006:4)

**sampling frame**  “specifies how samples are to be chosen
from the population of text”  
(McEnery & Hardie 2012:250)
**Against the notion of representativeness**
Teubert & Čermáková (2007:61, 69)

To claim “that a given corpus is representative of a discourse” in principle presupposes “access to all the texts the discourse consists of.”

“But if this utopia came true, we could well do without the corpus. ... it does not make much sense to talk about representativeness.”

---

**Classification of texts**
(Biber & Conrad 2009:Ch. 1 & 2)

Linguistic variation is systematic
Selection of linguistic features depends on non-linguistic factors.

Certain linguistic features are more frequent in some speech situations than in others.

**Register analysis**
1. Identify the situational characteristics of the texts.
2. Identify typical (pervasive) linguistic features.
3. Interpret the relationship between situational characteristics and pervasive linguistic features.

---

**2 Corpus building in language documentation**

<table>
<thead>
<tr>
<th>Typical corpus</th>
<th>LD corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>well-researched European language</td>
</tr>
<tr>
<td>texts</td>
<td>monolingual printed</td>
</tr>
<tr>
<td>corpus builder</td>
<td>team of native speakers</td>
</tr>
<tr>
<td>size</td>
<td>millions of words</td>
</tr>
<tr>
<td>purpose</td>
<td>linguistic research</td>
</tr>
<tr>
<td>compilation</td>
<td>specified by a sampling frame</td>
</tr>
</tbody>
</table>

---

**Situalional characteristics of text varieties**
(abbrev. from Biber & Conrad 2009:40)

**Participants** and their social characteristics
**Mode** (speech /writing), **Medium** (taped, radio, handwritten ...)
**Production** circumstances (real time, planned, edited, ...)
**Setting** (private / public; ...)
Communicative **purposes** (narrate, report, describe, ...)
**Topic**

The situational characteristics should be documented in the **metadata** of each text.
Corpus building with the Teop people

Language: Teop

genetic affiliation: Austronesian, Oceanic
location: Bougainville, Papua New Guinea

Sociolinguistic background:
- number of speakers: ca 6000
- subsistence farming
- many literate people
- civil war 1988-1999

Project:
- 1994 first recordings by Ruth Saovana Spriggs (funded by Australian Research Council)
- DOBES Project 2000-2007
- DFG Project 2009-2011
- 2011- funded by private resources

Collaborative language documentation

The Teop (opportunistic) corpus

structure and access needs to be revised as soon as possible!

http://corpus1.mpi.nl/ds/imdi_browser/
Revised corpus structure of the Teop Language Corpus
(on my computer)

- 00_ELAN_Teop_Corpus
  - 01_legends R
  - 02_legends E
  - 03_legends W
  - 04_personal_narratives R
  - 05_personal_narratives E
  - 06 Personal_narratives W
  - 07_descriptions R
  - 08_descriptions E
  - 09_description W
  - 10 unclassified R

Metadata also need revision

3. Corpora for language maintenance and linguistics

The aims of the LD are set by the linguist & indigenous documenters

- create something useful for the speech community
  (Pear-stories and Frog-stories are not useful!)

- trust that this will be useful for linguistics

Production of educational materials

- substantial part of the language documentation
  (Mosel 2012b)

New genres:

1. edited versions of legends
2. autobiographical narratives
3. procedural texts
4. encyclopedic descriptions of plants, animals, artefacts

"mini-dictionaries" (Mosel 2011)
3.1 Different modes: recorded speech vs. edited texts

1. original recordings with annotations
2. edited versions of the recordings

The comparative corpus
- gives a fuller picture of the expressive potential of the language;
- shows alternative ways of expressing the same content
- provides a new type of data for research on what speakers actually do when they put an oral text into writing

(Mosel 2015)

3.2 Different genres

3.2.1 Legends
- may contain archaic expressions
- are situated in imaginary worlds where animals talk and things can change into living beings
  > interesting for noun classification
- contain direct speech interjections, swearing ...

3.2.2. Encyclopedic descriptions

non-verbal clauses in definitions

(1) SUBJ.NP  PRED.NP  QUALIFICATIVE ATTRIBUTIVE AP
The bokua  (is) a fish  a big (one)

(2) SUBJ.NP  PRED.NP  POSSESSIVE AP
The booboo  (is) a fish  (with) a strong skin

(3) SUBJ.NP  PRED.NP  RELATIVE CLAUSE
The shelf  (is) a thing  that we put things (on).

Definitions of “thing”-words

supply excellent examples for:

1. non-verbal clauses
2. topicalisation
3. various kinds of modifiers
   (1) adjectival phrases (‘big’)
   (2) possessive adjectival phrases
       (‘having a thick skin’)
   (3) relative clauses
Definitions of “action” - words

NMLZ    DEM    COMPLEMENT CLAUSE
‘The tearing... this (is) when we remove ...’

A siri atovo
ART  tear  sago.palm.leaf
‘The tearing of the sago palm leaf,
.

ei be- ara gono kahi o paka
DEM  when- 1PL.INCL  get  from  ART  leaf
this (is) when we get from the leaf.’

bono sikiri nae.
ART  midrib  3SG.POSS
the midrib

3.2.3 Narratives vs procedural texts

Narratives
Paratactic clauses
Coordinate clauses
Sequence of past events

Procedural texts
Adverbial clause constructions: ‘when ..., then...’
Regular fixed order of actions

> create a corpus of comparative narrative and procedural texts 
minimise variables

Buy a chicken and ...

Make series of photographs and use them as stimuli for
1. the description of how to butcher a chicken
2. the narrative of how the twins helped their father butchering a chicken

procedural text: 40 clauses, 12 adverbial clause constr.
narrative text: 53 clauses, no adverbial clauses 13 paratactic clauses
3.3 Different themes – different grammatical phenomena

3.3.1 Tropical fishes are colourful

Do colour-words behave like beera ‘big’ and mataa ‘good’?
- do they have the same morphology?
- do they have the same syntactic functions?
- do they enter comparative constructions?

3.3.2 What trees are good for

Teop clause structure:
- intransitive: SUBJ VC
- transitive: SUBJ VC OBJ
  - agent patient/recipient/theme
- ditransitive: SUBJ VC OBJ₁ OBJ₂
  - agent recipient theme
  - agent patient instrument

Translation equivalents of Teop ditransitive clauses:
The man gave the child a coconut.
The man made the canoe from wood.
What is the dominant word order in Teop?

The more tree descriptions, the more clauses with OBJ₁, VC SUBJ OBJ₂ wordorder!

It makes only sense to speak of a dominant constituent order with respect to a particular type of genre.

>> quantitative analysis of constituent orders in different genres
>> for each clause you need to annotate the constituent order

General problem of LD corpora: annotation is very time-consuming.

ELAN: corpus building and exploitation tool
facilitates:
1. time aligned annotations of audio and video recordings
2. annotations on several tiers
3. simultaneous searches on several tiers
4. searches with the query language Regular Expressions
5. concordances and statistics of search results
6. export to Toolbox, Praat, printable files

4.1 Annotation

( McEnery et al. 2006:30-45, 75)

1. gives additional information on the form of the texts;
2. “records a linguistic analysis explicitly “;
3. “imposes a linguistic analysis upon a corpus user” but
4. this record of an explicit analysis is “open for scrutiny”
Annotation in LD corpora in ELAN

Research independent annotation
1. minimal annotation (DoBes)
2. gramm_units (Erfurt Referentiality Project)
3. interlinear morphological glossing (Leipzig glossing rules)

Research dependent annotation
1. GRAID (Haig & Schnell 2011)
2. Coreference annotation (McEnery et al. 2006:38-40)

4.2 Searches in ELAN
structured search on multiple files – single tier
concordance view – keyword in context - kwic

4.2 Searches in ELAN : collocations
Is old the only collocation of woman?
Examine 840 examples???
Search with regular Expressions

Find

Search for: a/an/the
followed by a word that does not start with o
followed by woman/women

If the protagonist is introduced by a NP with peha in the first utterance/clause of the text, how is it referred to in the second sentence?

peha   indefinite marker
bona   anaphoric determiner

If the protagonist is introduced by a NP with peha in the first utterance/clause of the text, how is it referred to in the second sentence?

Alignment view
5 Corpus size and grammatical analysis

Frequency counts are not sufficient for describing grammar, however. They point to interesting phenomena that deserve further investigation and interpretation (Conrad 2010:228).

Does the corpus supply enough data for the research question? How big is big enough?

The smaller the corpus, the more restricted are the research topics, and the more restricted the situational variables must be. So that the corpus presumably covers the full range of variability of the researched linguistic item.

6.1 Recommendations for corpus compilation

Focus on a few registers/genres.

The more diversified the corpus is, the smaller are the subcorpora, and the smaller the probability that you can adequately identify regular patterns of language usage.

6.2 Recommendations for corpus exploitation

# Use ELAN or a similar tool with an implemented powerful query language.

# Keep in mind that extended annotation is very time consuming. Don’t get lost in annotations.

# Document your annotation rules and your search methods.

# Make your corpus and the metadata accessible.

# Aim at scientific research that is replicable and falsifiable.

References


Moesl, Ulrike. 2014a. Searches in ELAN corpora with regular expressions.

