

Data collection methods in field-based LDD

Friederike Lüpke
F12@soas.ac.uk



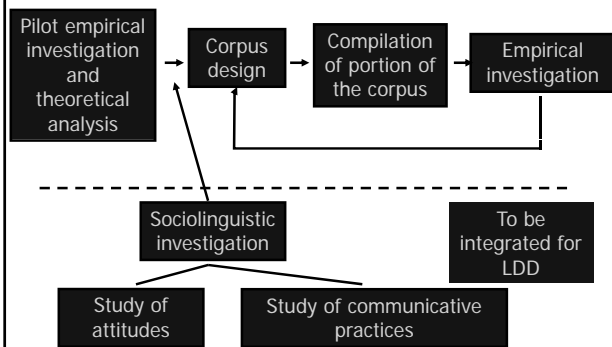
Striving for good corpora for linguistic analysis

Corpus design: chances and challenges

- It is relatively straightforward to create a representative corpus of, e.g. English fiction in the 20th century or French phone conversation.
- We know what the population is and can use statistical techniques to arrive at a stratified sample.
- We can then test the linguistic representativeness of the sample by measuring frequencies, standard deviation, etc.

But: what is the population in the case of the speech of an endangered language community?

Cycles of corpus design (Biber 1995)



Data based on different types of communicative events

Data based on 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.
- CONTRA:
 - 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 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.
- CONTRA:
 - 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'

Reference	Surface	Target	Label
ntre2 003	N ji leterna scbcxi naaxan ma, n xa a	rakolon i ra, maa o ra, n	Causative
ntre2 015	E naxa, i na n ma numero de Komptena	kolon , i marji n samba ra	
ntre2 021	I a	kolon , on lanx'ee ma moi:	
ntre2 083	nxo xa nxo booretoo, nxo nxo boore	kolon , nx'oo	Reciprocal
ntre2 083	nxo xa nxo booretoo, nxo nxo boore kolon,	kolon , nxo walesooma	
ntre2 084	Xa muxinee m' ee boore	kolon , e mun marji waleso:	
ntre2 081	Kono bai a m'an nxo	kolon ,	
ntre2 011	a m' aa	kolon e nun naaxee	Complement
ntre2 011	nxo malan, nxo xa nxo malan, nx'oo	kolon , nxo walesoo	
ntre2 029	A xili nde Damian, nan	kolon , beej i xilla na	Passive
ntre2 032	A mun ncn ma fee	kolonxi ,	
ntre2 040	luu haa e e	kolon , naaxee xaranna	Perfect
ntre2 065	I a	kolon , n konn' i menn'	
ntre2 068	fareboronden' i, kono, i a	kolon ,	
ntre2 074	na bai, i na boore	kolon ,	
ntre2 074	n mun na fala i, i na boore	kolon ,	Many transitive uses

Representativeness of a LDD corpus – Jalonke low frequency verb

Reference	Surface	Target	Label
ntre2 003	bee ma a nun saron , a nun jole, ningediinee nan fidin jee e sabaa	kolon ,	Past
ntre2 075	maa tuga, a saron , e naxe sah, moodi ,nde ji fulunxi on be?	kolon ,	
ntre2 236	xemee koo, a saron , banxee kwi, a saronx'ee banxee kwi, oyaa fan	kolon ,	
ntre2 236	banxee kwi, a saronx'ee , banxee kwi oyaa fan xa soo.	kolon ,	
ntre2 197	Geme Saron , memma fan sosonee nan na sintixi.	kolon ,	NP subject
ntre2 198	Geme Saron ,	kolon ,	Goal PP
ntre2 019	xi nan i, e nun saron , boxin'i.	kolon ,	
ntre2 112	iid en, a k'oo a saron , tanden'ii.	kolon ,	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 and lexical environments of the most frequent constructions.
 - Formulate hypotheses and prepare elicitation sessions.

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

Data based on staged communicative events

Evoking situations

- Especially in contexts of severe language endangerment, the context of use for a number of communicative events may not exist any longer.
- Semi-speakers and rememberers may feel inhibited to simulate these communicative events.
- Evoking a speech situation may help them to recover memory and feel less shy about reproducing speech events from the past.

There are no prefabricated stimuli for this event type – it will depend on the creativity of the researcher!

Staged communicative events based on nonverbal stimuli

Static stimuli

- Picture books
 - Topological relations picture book
 - Frog story
- Photos
 - Positional verbs picture book
- Comics
 - Calvin & Hobbes
 - Tintin
 - Asterix & Obelix



Dynamic stimuli

- Acted videos:
 - Staged events
 - Cut & Break
 - Pear film
- Animated videos:
 - Fish film
 - Event triads
 - ECOM clips



Interactive stimuli

- Matching/sorting games games:
 - Basic colour terms
 - Munsell chips
 - Men and tree
 - Cluedo
- Puzzles:
 - Eisenbeiss/Matsuo puzzle
- Map tasks/route descriptions:
 - HCRC map task
 - Table top route description task



Examples for the use of static stimuli

Posture verbs in stative positions (Ameka, de Witte & Wilkins 1999)



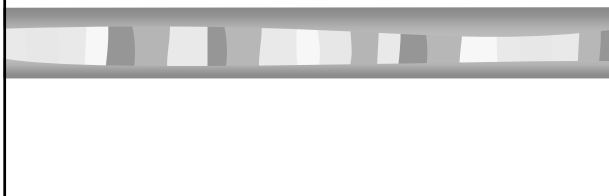
English: The bottle is standing on the rock.
 Jalonke: *Biniir-εε doo-xi gem-εε fari.*
 bottle-DEF sit-PF rock-DEF on
 'The bottle is sitting on the rock.'



Goemai: The stick is hanging on the tree trunk.
 Jalonke: *Tam-εε kiran-xi wurixuntun-na ma.*
 stick-DEF lean-PF tree trunk-DEF at
 'The stick is leaning against the tree trunk.'

19

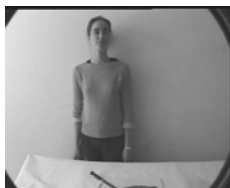
Examples for the use of dynamic stimuli



Cut & break verbs (Bohnenmeyer, Bowerman & Brown 2001)



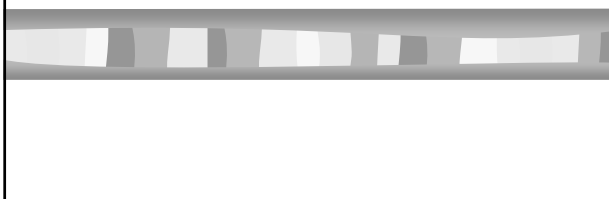
English: cut (with scissors)
 Dutch: knippen 'cut with scissors'
 Jalonke: cut-iterative (because cloth has already been cut).



English: cut (with knife)
 Dutch: snijden 'cut with a knife'
 Jalonke: cut (because fish hasn't been cut yet).

21

Examples for the use of interactive stimuli



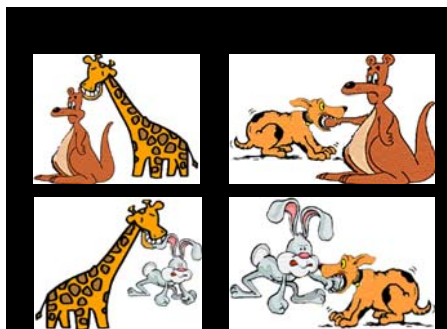
The Puzzle Task (Eisenbeiss & Matsuo 2003)



- Children have to describe puzzle pieces in order to be handed the piece to be handed to them
- The pictures are selected in order to elicit descriptions of external possession and to 'force' the children to verbalise all the relevant contrasts

23

An Example of the contrasts involved

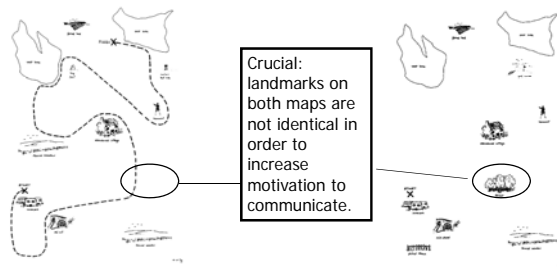


24

The HCRC map task (HCRC Edinburgh)

The instruction giver's map

The instruction follower's map



25

Ad hoc stimuli

Ad hoc stimuli

- New technologies enable fieldworkers to create stimuli 'ad hoc' in the field:
 - Digital photos
 - Video clips
 - Animations
- Although generally not usable for cross-linguistic comparison, these stimuli can yield interesting and highly relevant data difficult to get otherwise.

27

Action descriptions (Lüpke 2005, 2009)

- Videos recorded in the field that are described by consultants.
- PRO:
 - Yield fine-grained event descriptions difficult to obtain otherwise.
 - Can be used to cover semantic domains not attested so far in the corpus.
- CON:
 - Don't constitute a 'speech event' in the sense of Hymes.



28

Photos and Powerpoint animations

- Useful for ethnobotany
- Sequences of stills from digital video or ppt animations can be used to elicit stages of an event



29

Votre tour

- Lesquelles sont les méthodes que vous utilisez pour capturer les pratiques multilingues 'spontanées'?
- Avez-vous des recommandations à propos de situations, techniques, etc. spécifiques?
- Avez-vous rencontré que certaines méthodes passent dans un contexte, mais pas dans un autre?

30

Advantages and limits of SCVs

Limited ecological validity

- It is important to aim at culturally appropriate methods.
- However, total ecological validity leads to non-transferability.
- Therefore:
 - Elicitations and stimuli should replace the names of culturally unfamiliar items with more familiar ones.
 - However, replacement objects should possess the properties that are salient for the stimulus.



Different cultural settings and interpretational norms

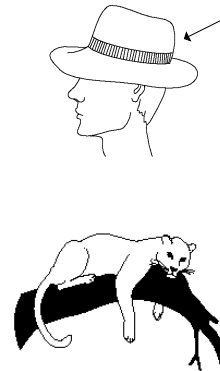
- Interactional tasks may violate culture-specific norms of interaction and should be carefully negotiated.
- Culture-specific picture-reading and other conventions should be taken into account when training consultants to work with stimuli created in Western societies with literacy in the Latin alphabet:



- There is compelling evidence that directionality in picture-reading and interpretation of picture sequences follows the orientation of the dominant writing system.
- The same holds for expectations on thematic roles of participants to the left vs. to the right of a picture

Inheritance effects

- Existing stimuli were created in order to answer specific research questions. This context should be taken into account when using them.
- TPRS: Originally created in order to investigate IN and ON topological relations.
- Later supplements cover other topological relations.



Controlled anthropomorphism: TomatoMan

- Moving images bear the strong probability of anthropomorphic interpretations of the Figures.
- In some cases (TomatoMan), this is intended, in other cases, there are no explicit guidelines on how to interpret the Figures.
- If unsure, try to guide your consultants consistently towards one interpretation only.

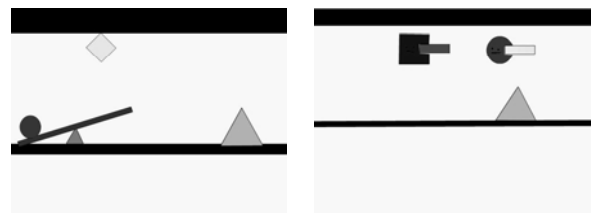


Ozuryek & Kita:
TomatoMan

Uncontrolled anthropomorphism: ECOM clips

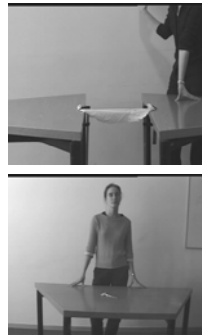
Uncontrolled

Controlled?



Flawed parameters underlying the stimulus

- Stimuli are only as good as the parameters covered by the scenes they contain.
- If these have gaps, the resulting data won't offer a complete picture of the semantic/functional domain covered by the stimulus.
- Likewise, if the stimulus only features unusual scenarios for the domain in question, the data may be inconclusive for its



37

General advantages and disadvantages

- Advantages of stimuli:
 - Are highly controlled, quantifiable and comparable.
 - Yield phonologically, semantically and syntactically accurate data.
 - Are free from linguistic interference of the metalanguage and from misunderstandings of context.
- Disadvantages:
 - Cross-cultural applicability can be limited.
 - Use is limited to visually depictable scenes.
 - Do not allow a semasiological approach (investigation the different uses of a form), but only an onomasiological approach (studying the formal expression of a given function).

38

Conclusion

- Like other methods for data collection, the use of stimuli is no 'free and easy' solution but requires careful consideration of factors regarding:
 - The design of the stimulus and its appropriateness to cover the domain in question in the specific cultural setting
 - The training needed for consultants (and researchers) to become familiar and at ease with the task.
 - The instructions and procedure necessary to obtain the desired results.
 - The ways in which the data can be analysed.

Like all all kinds of data, stimuli –based should be complemented with other data types in order to get the complete picture!

39

Data based on elicitation



Data resulting from translational equivalent elicitation of single words

"How do you say 'bee' in Gunyaamolo?"

- PRO:
 - Are easy when starting work on an unknown language.
 - Give good data to work on phoneme inventory, basic lexicon, and for lexical comparison.
 - Are quantifiable and highly controlled.
 - Offer negative evidence.
- CONTRA:
 - Yield phonologically odd utterances.
 - Can easily lead to misunderstandings due to the lack of context.
 - Give wrong ideas on the extension and intention of elicited words.
 - Impose taxonomies of the metalanguage.
 - Translatable items are limited in number.
 - Hyper-cooperative consultants may create neologisms and produce calques to be helpful.

41

Better: word lists as a result (Mosel 2004, 2006)

- There are suggestions to view wordlists as a result of elicitation rather than as an elicitation tool.
- Mosel (2004, 2006):
 - Collect lexical data organised in semantic, often usage-based domains (i.e building a canoe, farming, ...)
 - Let consultants lead the sessions and create the relevant taxonomies rather than imposing yours on them.
 - At an advanced stage of the research, run community workshops that at the same time work on standardisation, orthography, etc.

42

Data resulting from sentence translation

"How do you say 'Point Sud is a great place' in Bamanan?"

- PRO:
 - Sentence translation offers an easy way to see if something can be said, to help language learning and to prepare elicitation sessions
- CONTRA:
 - The contexts for and the felicity conditions of sentences are often not taken into account.
 - Often, translation equivalents are mixed up with acceptability judgments, creating uncontrollable parameters.

43

Recommendations for sentence translations (Matthewson 2004)

- Provide a discourse context for the sentence prior to eliciting its translation.
- Ask for translations of complete sentences only.
- Try to make the source string a grammatical sentence.
- Assume that the result string is a grammatical sentence.
- Take sentence translations as cues about felicity conditions rather than as an absolute truth.

44

Data resulting from acceptability judgements

"Can I say 'this book' when the book is lying over there?"

- PRO:
 - Are controlled and quantifiable.
 - Can give results for domains that are difficult to cover otherwise.
 - Give comparable results for many fields.
 - Offer negative evidence.
- CONTRA:
 - Very often do not test acceptability of the utterance, but rather of the context provided for it.
 - Can therefore very often be contradicted by the same and/or different speakers.
 - Often have other hidden factors like nature of instructions, order of presentations, frequency, training of consultants, etc., that influence the judgment.

45

Recommendations for judgment tasks (Lüpke, 2009, Schütze 1996, 2005)

- Create detailed instructions for the rating of sentences.
- Develop a clear scale for ratings.
- Provide your consultants with some example sentences and your ratings of them before the task.
- Conduct some training tasks before the actual task.
- Document demographic details of the consultants and try to aim for a homogenous group in terms of education, literacy, handedness, et.

46

Summary

- Elicited data that are inspected in a qualitative way allow to
 - Get the full distributional range of a given item/construction.
 - Test the semantic properties of that item/construction.
 - Provide negative evidence, i.e. information on unattested structures/uses, ungrammaticality, etc.

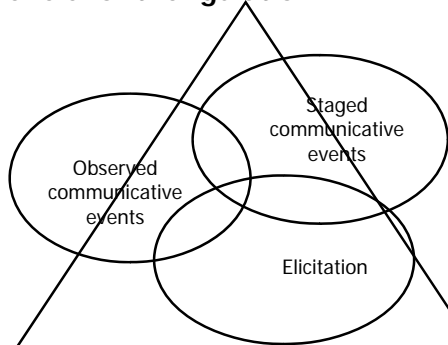
But: these data are often influenced by the metalanguage/elicitation method and not naturalistic at all.

47

My conclusion



A circle for triangulation



Some principles of data collection

Labov's four principles (Labov 1975)

- I. The Consensus Principle: if there is no reason to think otherwise, assume that the judgments of any native speaker are characteristic of all speakers of the language.
- II. The Experimenter Principle: if there is any disagreement on introspective judgments, the judgments of those who are familiar with the theoretical issues may not be counted as evidence.
- III. The Clear Case Principle: disputed judgments should be shown to include at least one consistent pattern in the speech community or be abandoned. If differing judgments are said to represent different dialects, enough investigation of each dialect should be carried out to show that each judgment is a clear case in that dialect.
- IV. The Principle of Validity: when the use of language is shown to be more consistent than introspective judgments, a valid description of the language will agree with that use rather than introspections.

(Labov 1975: 40)

... complemented by mine

(Lüpke 2009)

- V. The Principle of Explicitness. Analytical choices and decisions should be made explicit, i.e. the reasons to select a particular data collection method, to include or exclude a particular set of data, to work with a specific (group of) consultant(s) should be documented in metadata descriptions and annotations of primary data.

- VI. The Principle of Transparency. Abbreviations, symbols, labels, meanings of tiers used in transcriptions, numeric variables in spreadsheets, etc., should be explained in metadata and annotations of primary data.

VII. The Principle of Saliency. For the analysis of a particular research question, the most salient method for collection and analysis should be selected. For instance, descriptions of visual scenes rather than translation equivalents should serve as the basis for the analysis of spatial language.

55

VIII. The Principle of Triangulation. Wherever possible, analysis should be verified through triangulation, that is, through different methods of data collection, data from more than one consultant, different types of analysis, and comparison of data with those collected by other researchers, etc.

56

IX. The Principle of Longevity. Efforts should be made to make data valid beyond the scope of the individual research by not just seeking the data necessary to answer specific research questions or relating to one particular area of language use. So, for instance, when collecting data on topological relation markers, one should not limit oneself to stimuli-based data but complement them with observed discourse, etc.

57

Useful links



- MPI Nijmegen Language & Cognition and Acquisition Groups:
 - Large number of stimuli on a range of topics; stimuli and manuals upon request:
<http://www.mpi.nl>
- The MPI EVA Leipzig links to field tools:
 - <http://lingweb.eva.mpg.de/fieldtools/tools.htm>
- Russ Tomlin's Fish Film:
 - Stimulus designed to uncover the motivation for voice contrasts, topicality, etc.
http://logos.uoregon.edu/tomlin/research_fishfilm.html
- Wallace Chafe's Pear Film
 - Designed to compare narrative structure
<http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm>
- Phillip Wolff's animations on causality (upon request?)
 - Aimed at testing Talmy's force dynamics model of causation
<http://userwww.service.emory.edu/~pwolff/CLSLab.htm>

59