# Data collection methods in field-based LDD

Friederike Lüpke
Fl2@soas.ac.uk

THE HANS RAUSING
Endangered Languages Project

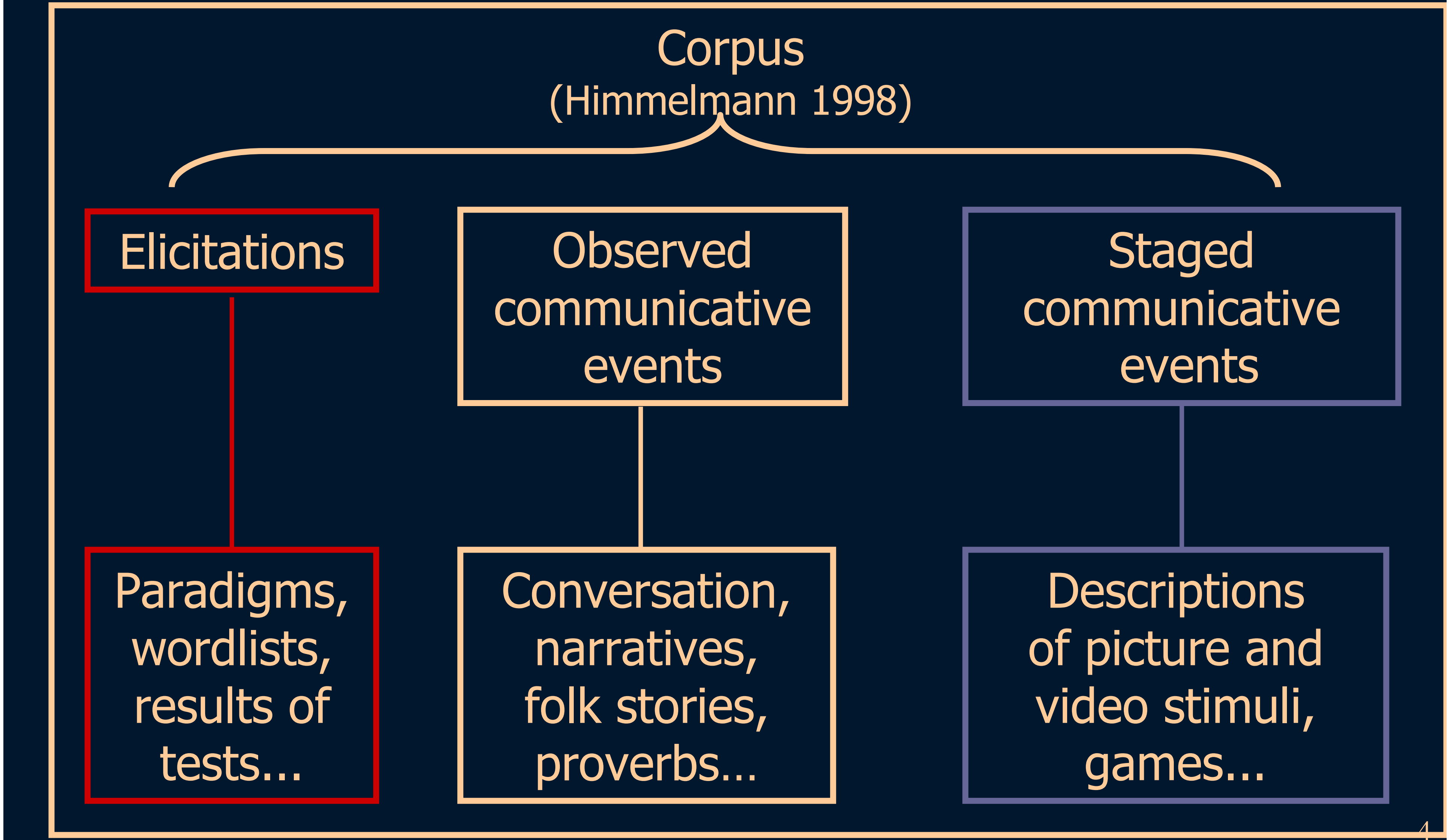
SOAS
University of London

# Structure of the talk

- The role of data in language documentation and description
- Three paradoxes of LDD
- Different data and different methods for linguistically and situationally balanced corpus:
  - Cyclic corpus design
  - Different types of communicative events
  - Awareness of non-linguistic aspects of communicative events
- A set of principles for linguistic data collection in LDD

# The role of data in LDD

## Documentation = a large, annotated corpus



Corpus

(Himmelmann 1998)

Elicitations

Paradigms, wordlists, results of tests...

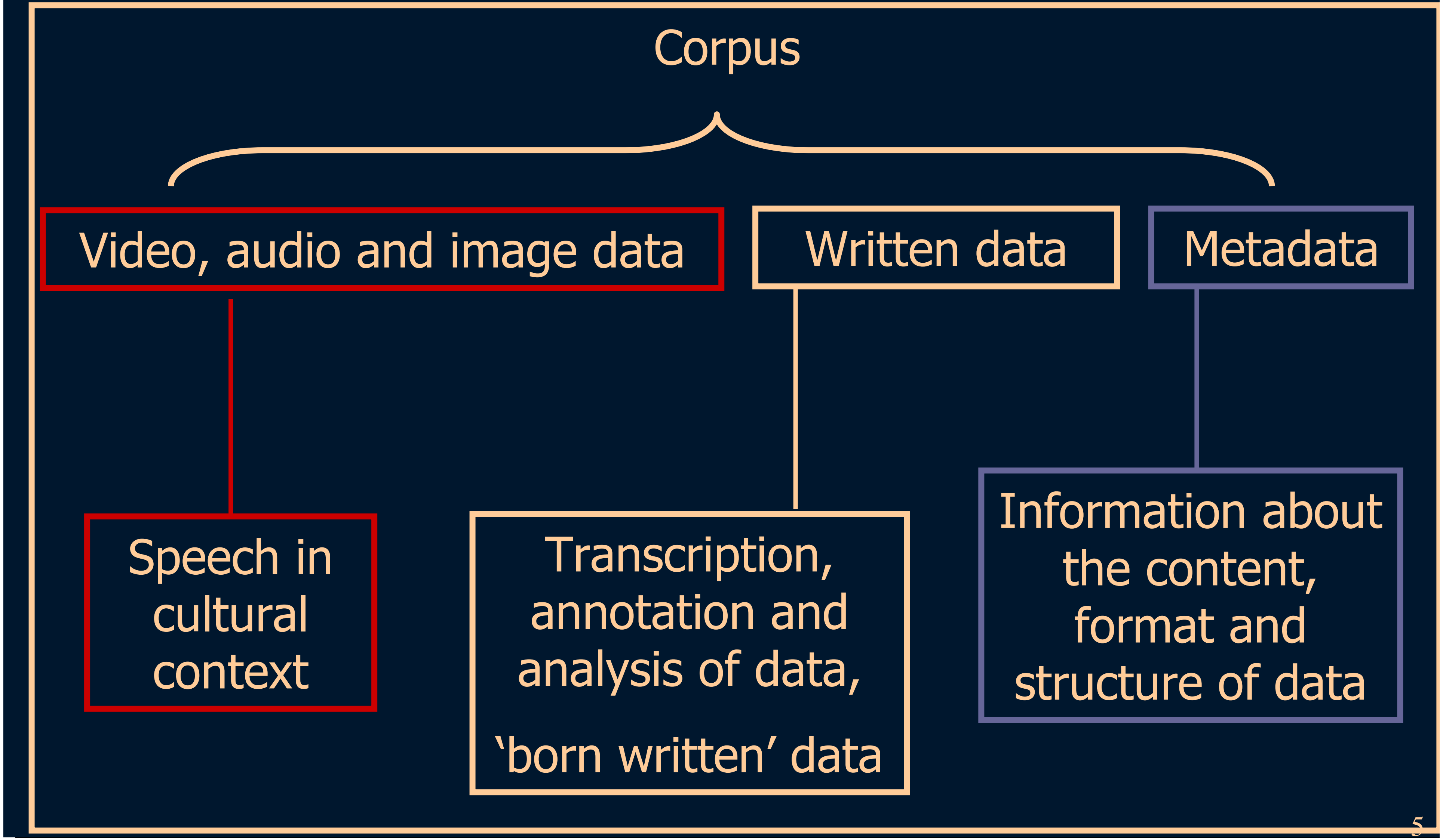
Observed communicative events

Conversation, narratives, folk stories, proverbs…

Staged communicative events

Descriptions of picture and video stimuli, games...

# Data types in the corpus



Corpus

Video, audio and image data

Written data

Metadata

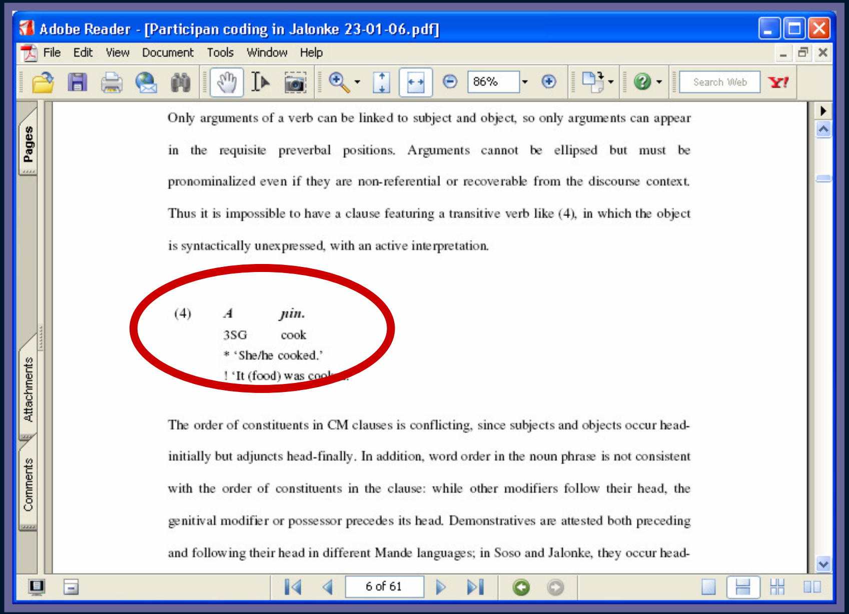
Speech in cultural context

Transcription, annotation and analysis of data,

'born written' data

Information about the content, format and structure of data

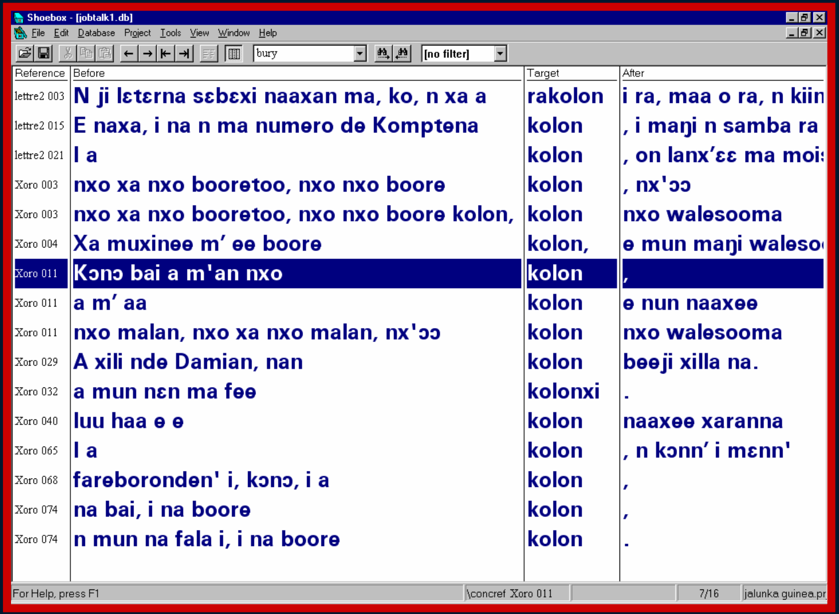
# Description vs. documentation


Adobe Reader - [Participan coding in Jalonke 23-01-06.pdf]
File Edit View Document Tools Window Help
86%
Search Web
Pages
Attachments
Comments
6 of 61

Only arguments of a verb can be linked to subject and object, so only arguments can appear in the requisite preverbal positions. Arguments cannot be ellipsed but must be pronominalized even if they are non-referential or recoverable from the discourse context. Thus it is impossible to have a clause featuring a transitive verb like (4), in which the object is syntactically unexpressed, with an active interpretation.

(4) *A* *ɲin.*
3SG cook
* 'She/he cooked.'
! 'It (food) was cooked.'

The order of constituents in CM clauses is conflicting, since subjects and objects occur head-initially but adjuncts head-finally. In addition, word order in the noun phrase is not consistent with the order of constituents in the clause: while other modifiers follow their head, the genitival modifier or possessor precedes its head. Demonstratives are attested both preceding and following their head in different Mande languages; in Soso and Jalonke, they occur head-



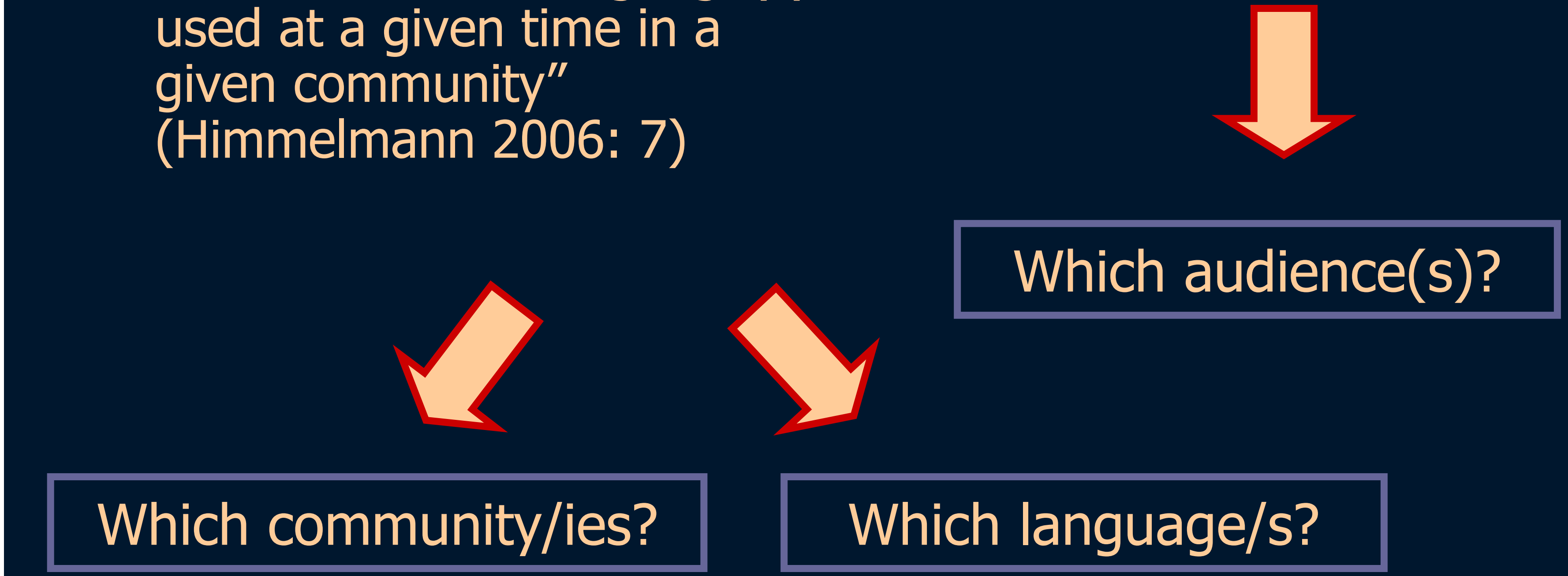
| Reference | Before | Target | After |
|---|---|---|---|
| lettre2 003 | N ji lɛtɛrna sɛbɛxi naaxan ma, ko, n xa a | rakolon | i ra, maa o ra, n kiin |
| lettre2 015 | E naxa, i na n ma numero de Komptena | kolon | , i maɲi n samba ra |
| lettre2 021 | I a | kolon | , on lanx'ɛɛ ma moi: |
| Xoro 003 | nxo xa nxo booretoo, nxo nxo boore | kolon | , nx'ɔɔ |
| Xoro 003 | nxo xa nxo booretoo, nxo nxo boore kolon, | kolon | nxo walesooma |
| Xoro 004 | Xa muxinee m' ee boore | kolon, | e mun maɲi waleso |
| Xoro 011 | Kɔnɔ bai a m'an nxo | kolon | , |
| Xoro 011 | a m' aa | kolon | e nun naaxee |
| Xoro 011 | nxo malan, nxo xa nxo malan, nx'ɔɔ | kolon | nxo walesooma |
| Xoro 029 | A xili nde Damian, nan | kolon | beeji xilla na. |
| Xoro 032 | a mun nɛn ma fee | kolonxi | . |
| Xoro 040 | luu haa e e | kolon | naaxee xaranna |
| Xoro 065 | I a | kolon | , n kɔnn' i mɛnn' |
| Xoro 068 | fareboronden' i, kɔnɔ, i a | kolon | , |
| Xoro 074 | na bai, i na boore | kolon | , |
| Xoro 074 | n mun na fala i, i na boore | kolon | . |

- "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)

# But exactly what data?

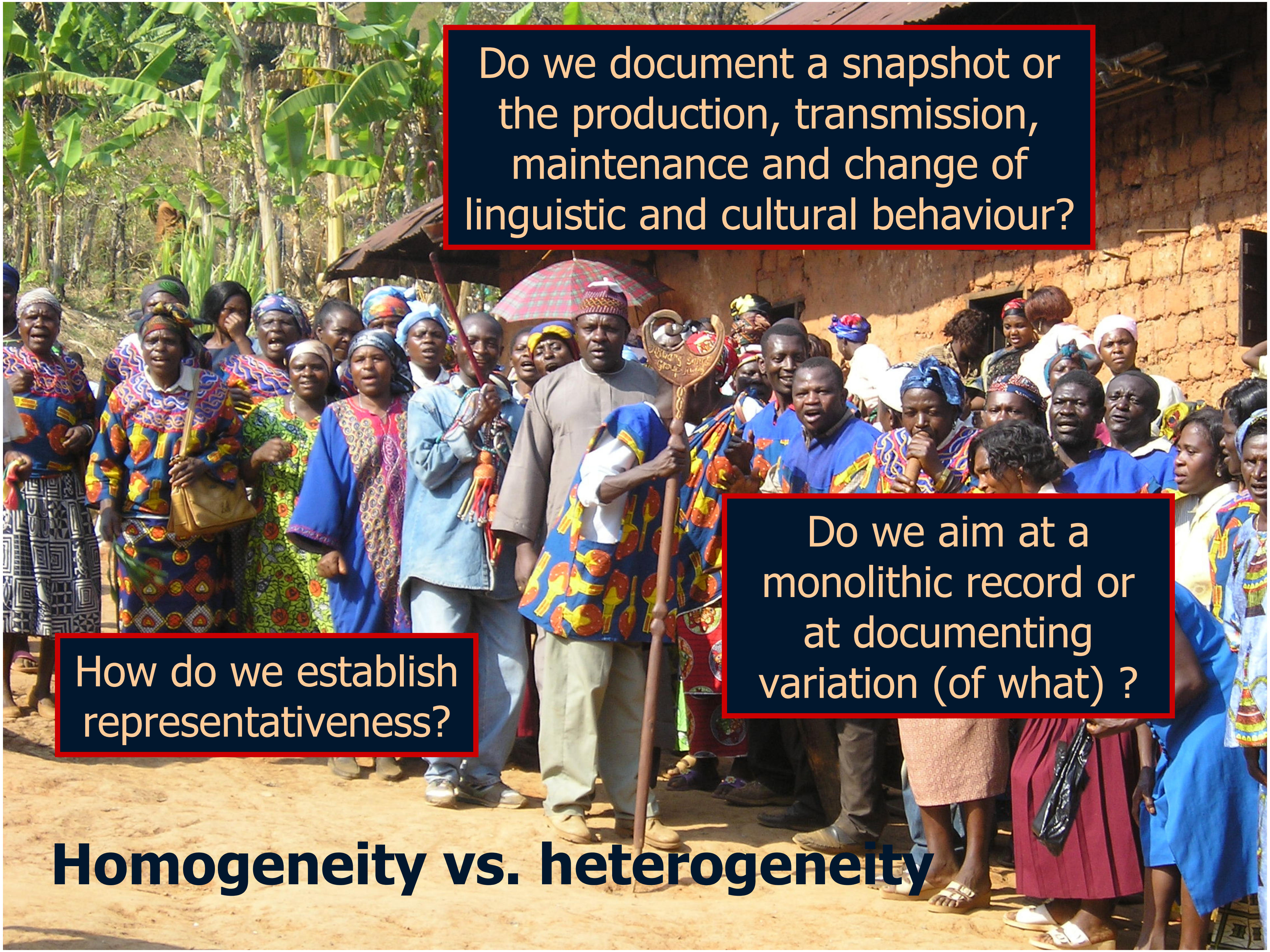
- “A language documentation […] conceived of as a lasting, multipurpose record of a language [… ] should contain a large set of primary data which provide evidence for the language(s) used at a given time in a given community” (Himmelmann 2006: 7)
- “The main goal of a language documentation is to make primary data available for a broad group of users.” (Himmelmann 2006: 15)



Which audience(s)?

Which community/ies?

Which language/s?



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?

# Homogeneity vs. heterogeneity

# 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?

# 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 and impact of our electronic corpora.

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

# Data and methodology

- “The major discovery of post-1957 “syntactic theory” is not “theoretical”, but methodological: That a huge amount of generalizations can best be found by adopting an “experimental” approach…What remains of the published body of research is the empirical part. So all the papers that are neatly divided into a “data/generalizations” part and an “analysis” part have a good chance of continuing to be useful”. (Haspelmath 2006: Linguistlist 17.2304)

If its data that is central, how can we assure that our data are, and will be, relevant? ?

How can we reach maximal transparency and explicitness in providing information about how and why we collected our data ?

# What status for negative evidence?

- "With regard to the usual way of obtaining negative evidence (i.e. asking one or two speakers whether examples x, y, z, are "okay"), it is doubtful whether this really makes a difference in quality compared to evidence provided by the fact that the structure in question is not attested in a large corpus. Elicited evidence is only superior here if it is very carefully elicited, paying adequate attention to the sample of speakers interviewed, potential biases in presenting the material, and the like." (Himmelmann 2006: 23)

How much methodological and theoretical awareness can we expect in language documentation?

Which methods are robust and widely accepted?

# Three paradoxes

# We create corpora…

The structure of the Jalonke corpus (Lüpke 2005)

| Communicative event | Genre | | Rec. time (min) |
|---|---|---|---|
| Observed | Narrative | Historical | 118 |
| | | Personal | 226 |
| | | Story | 127 |
| | Conversation | | 259 |
| | Other (speeches, songs, proverbs, procedural texts, etc.) | | 318 |
| Staged | Action descriptions | | 235 |
| Total recording time | | | 1283 (ca. 21 h) |

# … but do not systematically explore them

- We don’t use the computational approaches developed by corpus linguistics.
- We don’t engage in genre and register studies.
- We don’t engage in Conversation or Discourse Analysis.

Why not?

- Computational tools and methods haven’t been adapted yet to small field-based corpora.
- Detailed genre and register studies are beyond the scope of first documentations.
- The notation systems developed by CA and DA are too time-consuming to apply to field-based data.

# We collect performances…

A Jalonke song recorded in Herikoo, Guinea,  in 2001.

- This song was recorded ‘accidentally’ during a visit to a Jalonke village.
- The purpose of the visit was to distribute a Jalonke primer.

# … but don’t have a concept of them

- We take video recordings of performances, but are mainly interested in the speech, not in the visual information, musical structure, etc. present in them.
- We don’t systematically record different performances, analyse, or compare them.
- We don’t  try to establish of what genres they are instances of.

Why not?

- We are more interested in the linguistic aspects than in the artistic, interactional, and rhetoric characteristics of performances.
- We come across performances in a very unsystematic way.
- A first classification of genres and registers is a huge task already.

# We document parts of oral history and literature…

- Most field linguists collect stories, integrate them into their corpus and use them for linguistic analysis and the creation of literacy materials.

A Jalonke story recorded in Herikoo, Guinea, in 2001.

# … but are not really interested in them

- Most field linguists don’t study the literary genres they collect in their own right.
- Especially folk tales are often used for the creation of literacy materials for speech communities, but without any prior reflection on:
  - The differences between oral and written discourse.
  - The impact of writing down a specific performance.
  - The impact of editing (or not) the spoken text for the purpose of writing it.
  - The creation of a de facto standard in terms of orthographical, grammatical and stylistic patterns.
  - The creation of a de facto standard in terms of content and ‘authorized’ version.

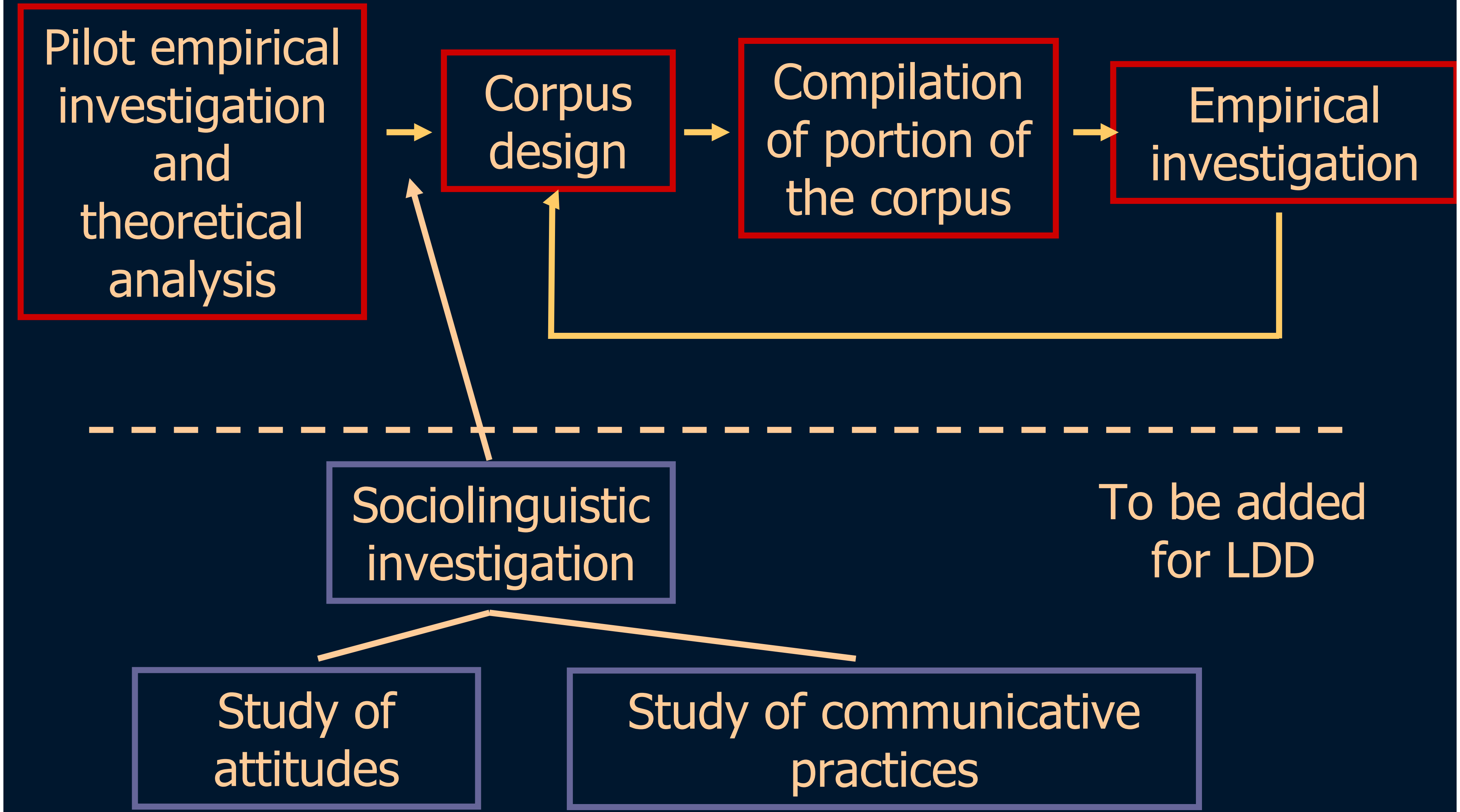
# Striving for good corpora for linguistic analysis

# Corpus design: chances and challenges

- It is relatively straightfoward 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)



Pilot empirical investigation and theoretical analysis

Corpus design

Compilation of portion of the corpus

Empirical investigation

Sociolinguistic investigation

To be added for LDD

Study of attitudes

Study of communicative practices

# 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

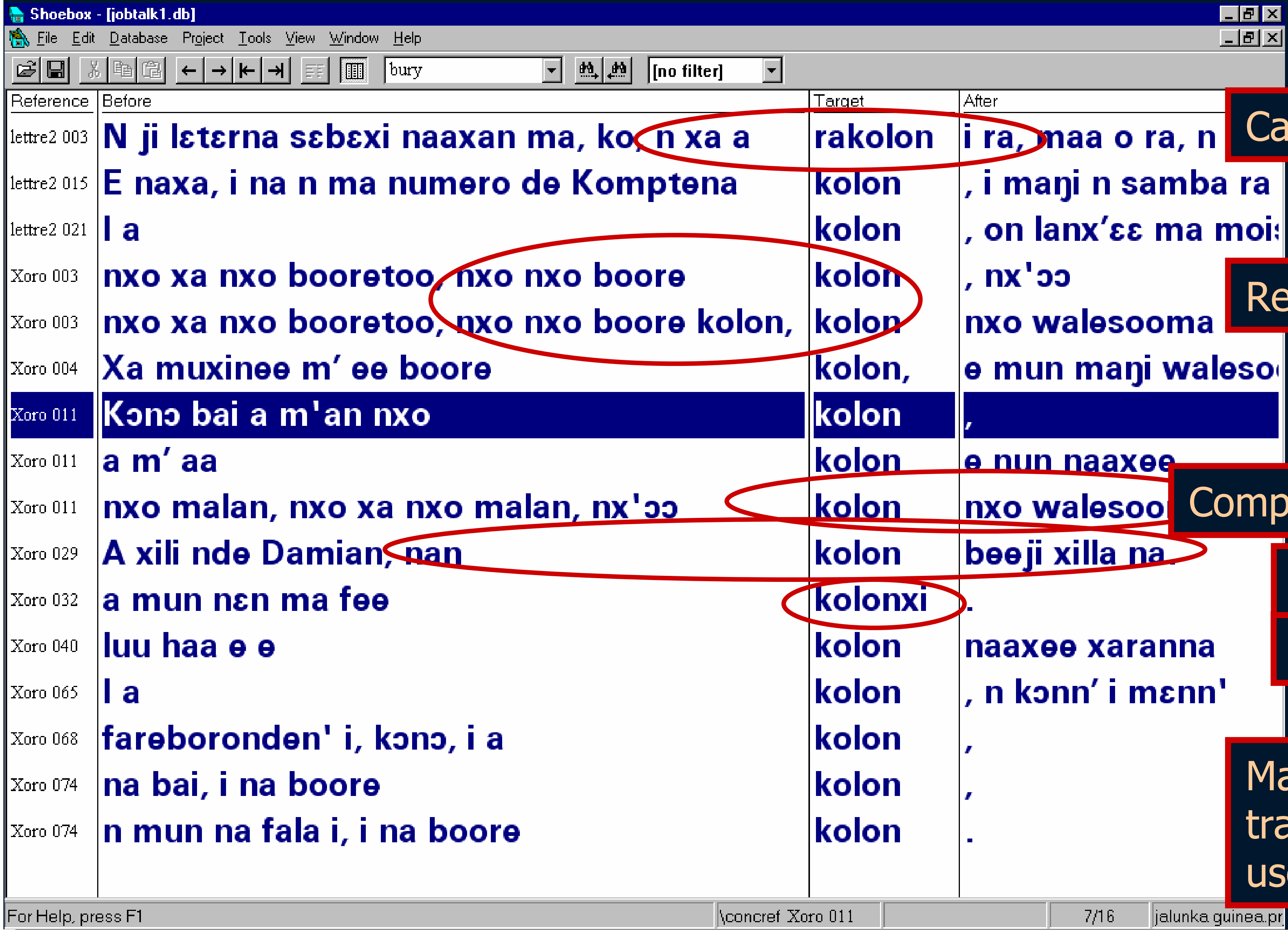
A: “How do you like the summerschool so far?”

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 | Before | Target | After |
|---|---|---|---|
| lettre2 003 | N ji lɛtɛrna sɛbɛxi naaxan ma, ko n xa a | rakolon | i ra, maa o ra, n |
| lettre2 015 | E naxa, i na n ma numero de Komptena | kolon | , i maŋi n samba ra |
| lettre2 021 | I a | kolon | , on lanx'ɛɛ ma moi |
| Xoro 003 | nxo xa nxo booretoo, nxo nxo boore | kolon | , nx'ɔɔ |
| Xoro 003 | nxo xa nxo booretoo, nxo nxo boore kolon, | kolon | nxo walesooma |
| Xoro 004 | Xa muxinee m' ee boore | kolon, | e mun maŋi waleso |
| Xoro 011 | Kɔnɔ bai a m'an nxo | kolon | , |
| Xoro 011 | a m' aa | kolon | e nun naaxee |
| Xoro 011 | nxo malan, nxo xa nxo malan, nx'ɔɔ | kolon | nxo walesoo |
| Xoro 029 | A xili nde Damian, nan | kolon | beeji xilla na |
| Xoro 032 | a mun nɛn ma fee | kolonxi | . |
| Xoro 040 | luu haa e e | kolon | naaxee xaranna |
| Xoro 065 | I a | kolon | , n kɔnn' i mɛnn' |
| Xoro 068 | fareboronden' i, kɔnɔ, i a | kolon | , |
| Xoro 074 | na bai, i na boore | kolon | , |
| Xoro 074 | n mun na fala i, i na boore | kolon | . |

For Help, press F1 | \concref Xoro 011 | 7/16 | jalunka guinea.pr

Causative

Reciprocal

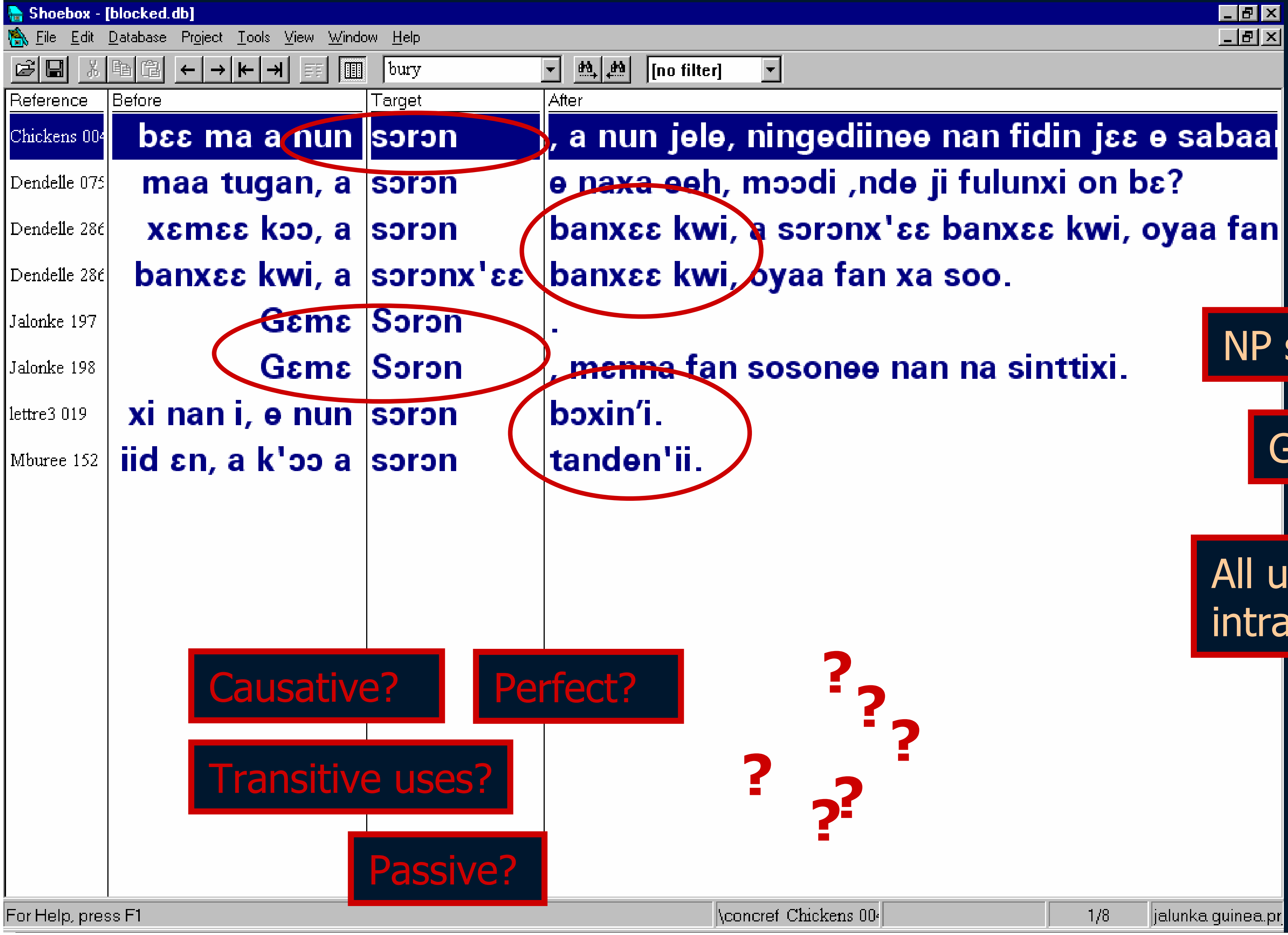
Complement

Passive

Perfect

Many transitive uses

# Representativeness of a LDD corpus – Jalonke low frequency verb



| Reference | Before | Target | After |
|---|---|---|---|
| Chickens 004 | bɛɛ ma a nun | sɔrɔn | , a nun jele, ningediinee nan fidin jɛɛ e sabaa |
| Dendelle 075 | maa tugan, a | sɔrɔn | e naxa eeh, mɔɔdi ,nde ji fulunxi on bɛ? |
| Dendelle 286 | xɛmɛɛ kɔɔ, a | sɔrɔn | banxɛɛ kwi, a sɔrɔnx'ɛɛ banxɛɛ kwi, oyaa fan |
| Dendelle 286 | banxɛɛ kwi, a | sɔrɔnx'ɛɛ | banxɛɛ kwi, oyaa fan xa soo. |
| Jalonke 197 | Gɛmɛ | Sɔrɔn | . |
| Jalonke 198 | Gɛmɛ | Sɔrɔn | , menna fan sosonee nan na sinttixi. |
| lettre3 019 | xi nan i, e nun | sɔrɔn | bɔxin'i. |
| Mburee 152 | iid ɛn, a k'ɔɔ a | sɔrɔn | tanden'ii. |

Causative?

Perfect?

Transitive uses?

Passive?

? ? ? ? ??

Past

NP subject

Goal PP

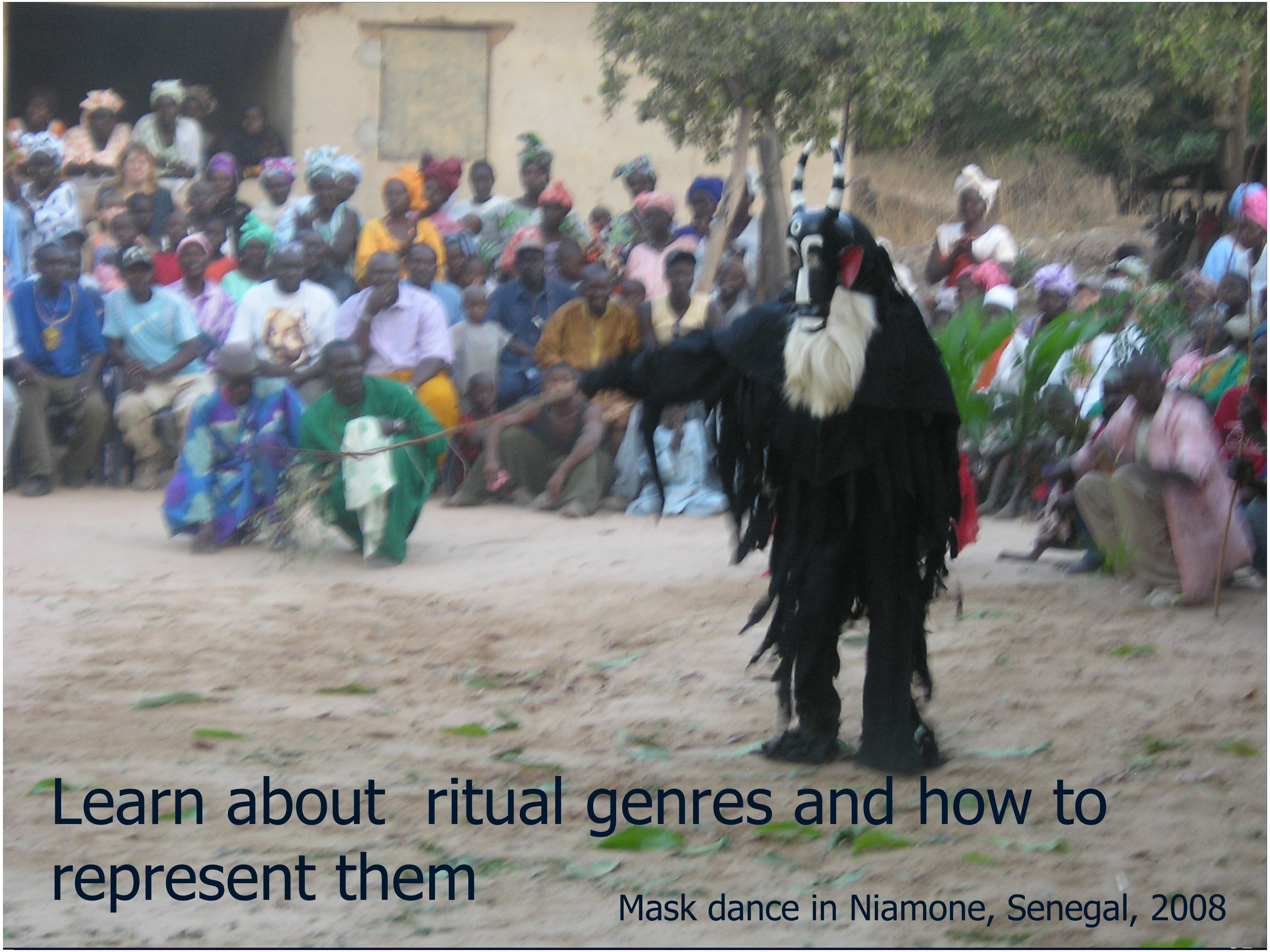
All uses are intransitive

# 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.

# OCVs: more avenues to pursue

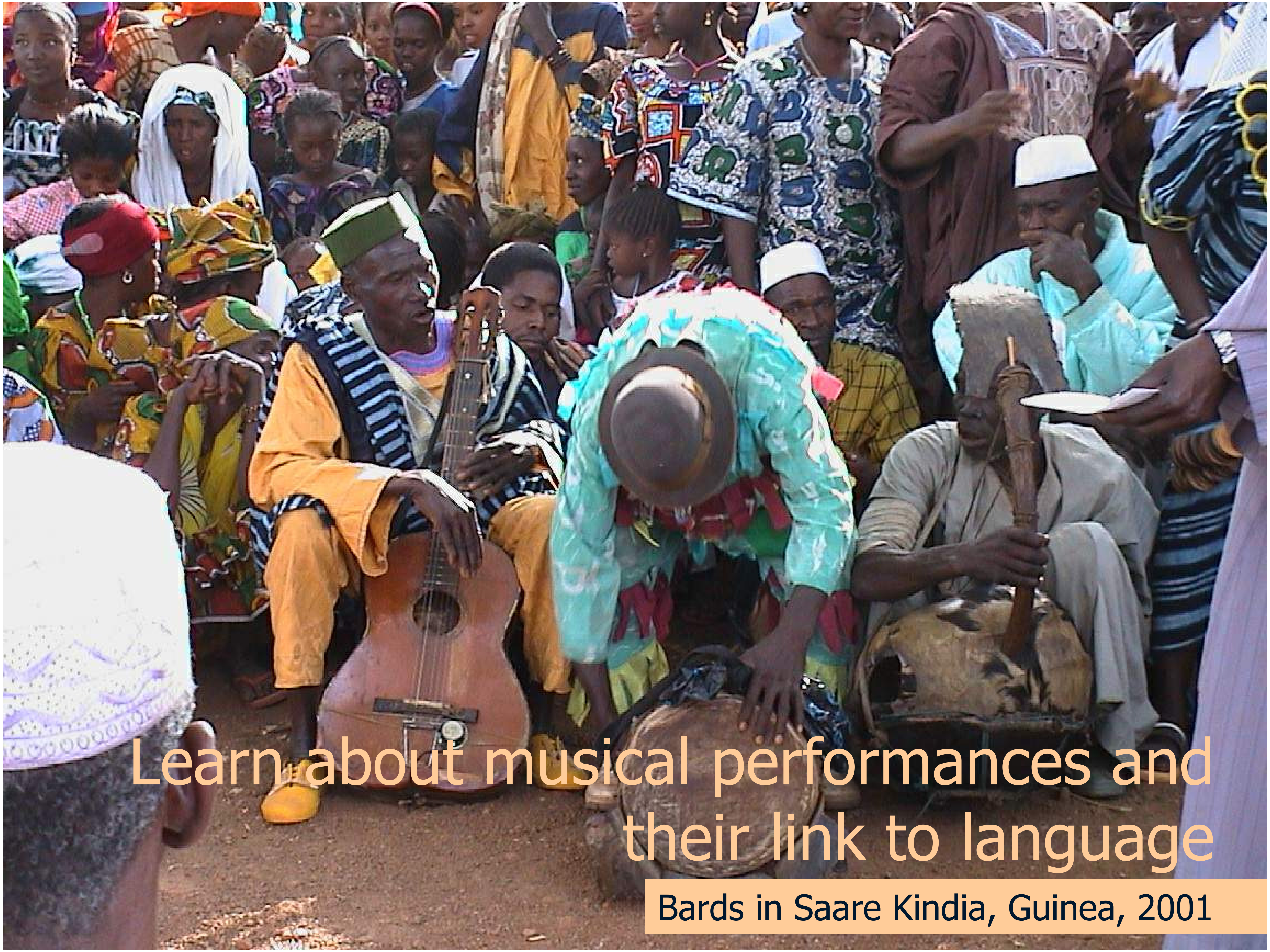


# Learn about ritual genres and how to represent them

Mask dance in Niamone, Senegal, 2008

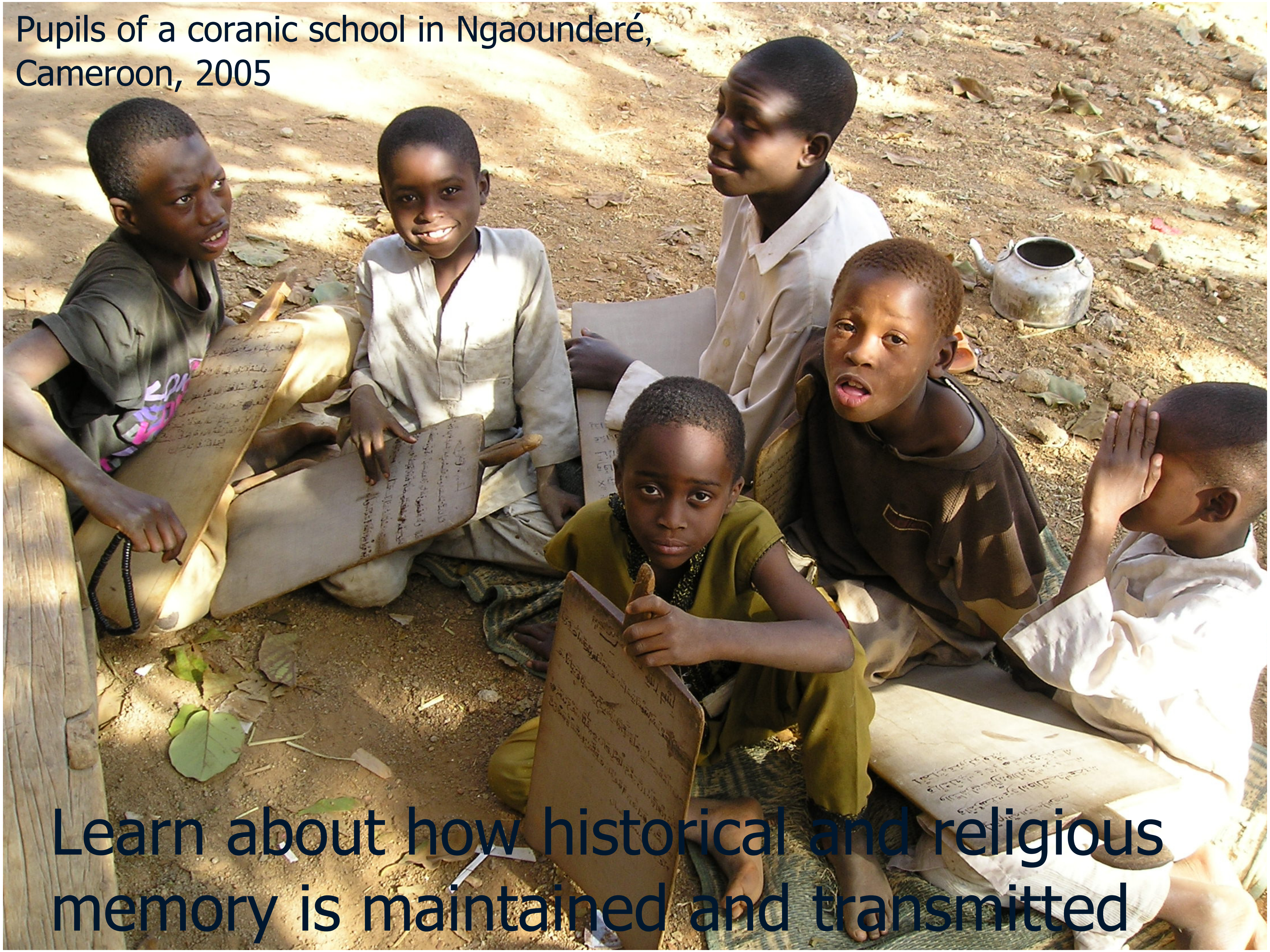
# Learn about material culture and its role for performance, memory and identity

Artefacts in Seleki, Senegal, 2008



# Learn about musical performances and their link to language

Bards in Saare Kindia, Guinea, 2001



Pupils of a coranic school in Ngaounderé, Cameroon, 2005

Learn about how historical and religious memory is maintained and transmitted

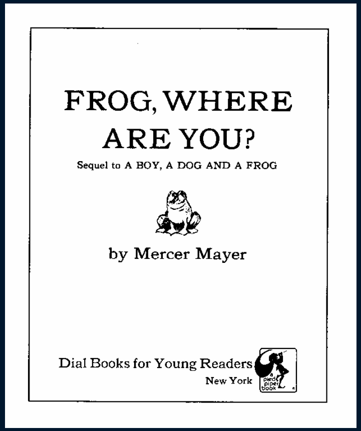
**… not in order to analyse all these aspects of culture as expressed through languages, but to create the record necessary to investigate them in the future…**

# Data based on staged communicative events

(Very briefly, more detail in the afternoon classes!)

# Types of stimuli

- Static stimuli:
  - Comics
  - Picture books
  - Photos
- Dynamic stimuli:
  - Acted videos
  - Animated videos
  - Staged life events
- Interactive stimuli:
  - Puzzle tasks
  - Map tasks
  - Matching games

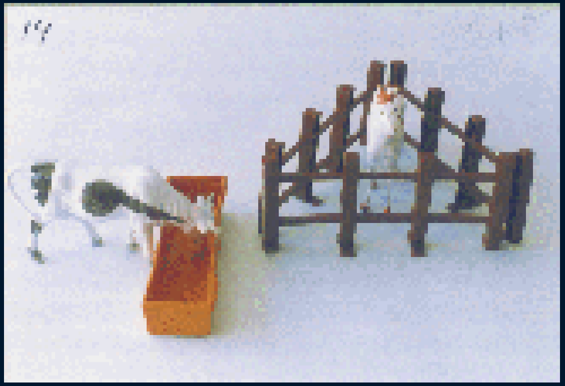


FROG, WHERE ARE YOU?

Sequel to A BOY, A DOG AND A FROG

by Mercer Mayer

Dial Books for Young Readers
New York



# General advantages and limits of stimuli

- PRO:
  - 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.
- CONTRA:
  - Cross-cultural applicability can be limited.
  - Use is limited to visually depictable scenes.
  - Do not allow a semasiologial approach (investigation the different uses of a form), but only an onomasiological approach (studying the formal expression of a given function)

# 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.

# 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.

# Data resulting form sentence translation

“How do you say ‘This is a nice city’ in Gunyaamolo?”

- 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.

# 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.

# 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.

# Recommendations for jugdment tasks (Lüpke, ms. Schütze 1996, 2005)

- Create detailled 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.

# 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.

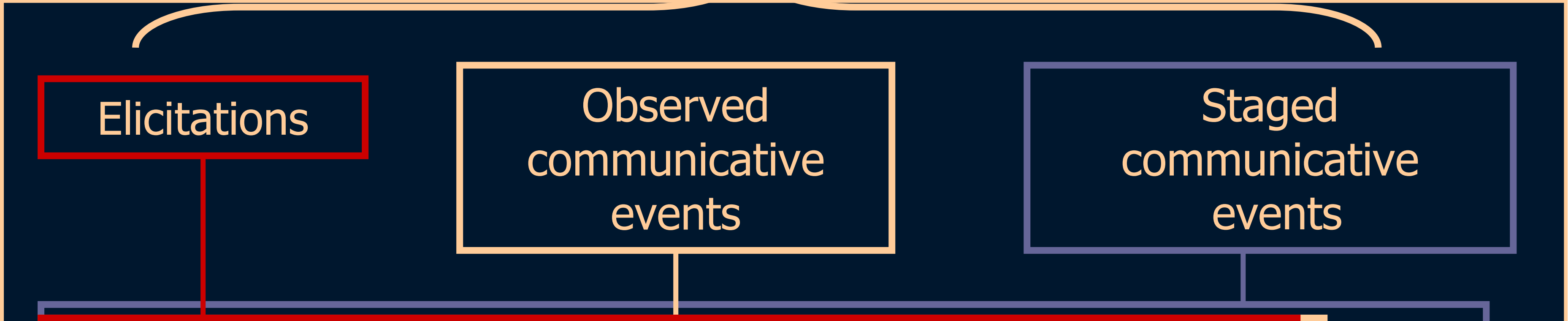
But: these data are often influenced by the metalanguage/elicitation method and not naturalistic at all.

# My conclusion

# Why all kinds of data?

Field-based corpora are relatively small. Thus:

- They don’t show the full distributional range of a given item.
- They don’t offer negative evidence.



Elicitations

Observed communicative events

Staged communicative events

Are highly controlled, but linguistically prompted. Thus:

- They are likely to be influenced by the linguistic input.
- Their ‘naturalness’ cannot be assessed.

hus:

ion.

cross

languages.

# What kinds of data for what kinds of findings?

# 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

V. The Principle of Expliciteness. 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.

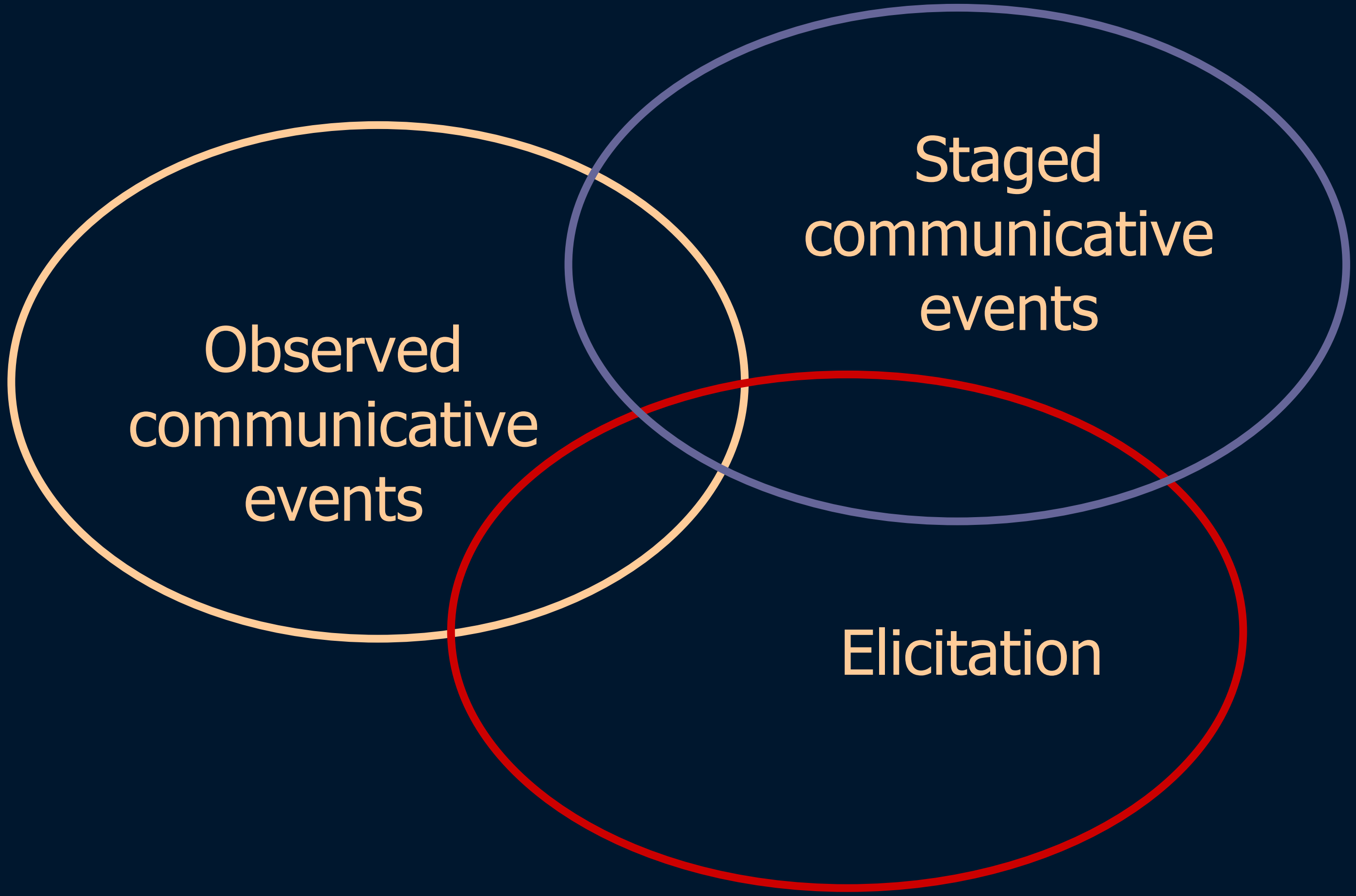
VI. The Principle of Transparency. Abbreviations, symbols, labels, meanings of tiers used in transcriptions, numeric variables in spreadsheets, etc., should be explained in metadata and annotations of primary data.

VII. The Principle of Salience. 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.

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.

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.

# A circle


Observed communicative events
Staged communicative events
Elicitation