

Defining documentation

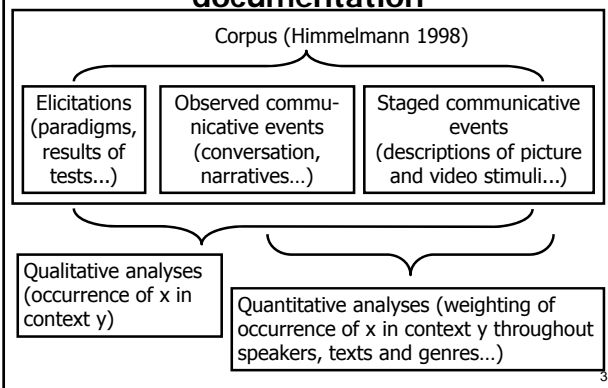
Friederike Lüpke
Endangered Languages Academic Programme
SOAS London

Your turn

- Together with your neighbour, please take five minutes to come up with a Wikipedia style definition of language documentation.
- Please share your (brief) definition with the group.

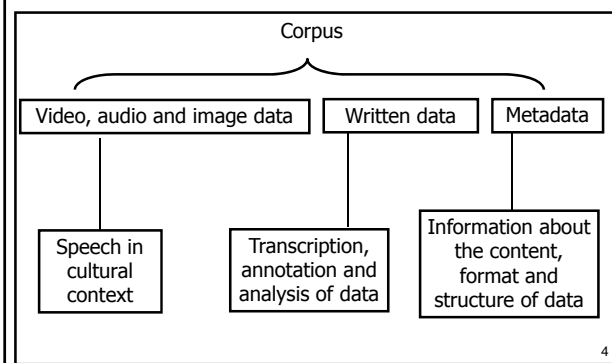
2

One view of language documentation



3

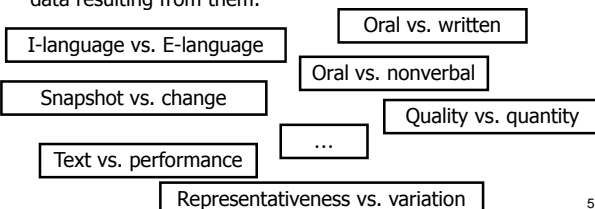
Data types in the corpus (format)



4

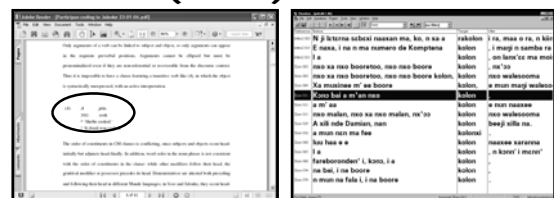
Format vs. content

- So far, the field of language documentation has focussed 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.



5

The (new?) role of data

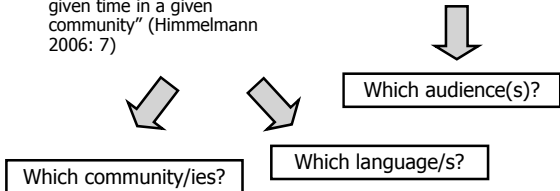


- "For description, the main concern is the production of grammars and dictionaries whose primary audience are linguists... In these products language data serves essentially as exemplification and support for the linguist's analysis." (Austin 2006: 87)
- "[...] Language documentation, on the other hand,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

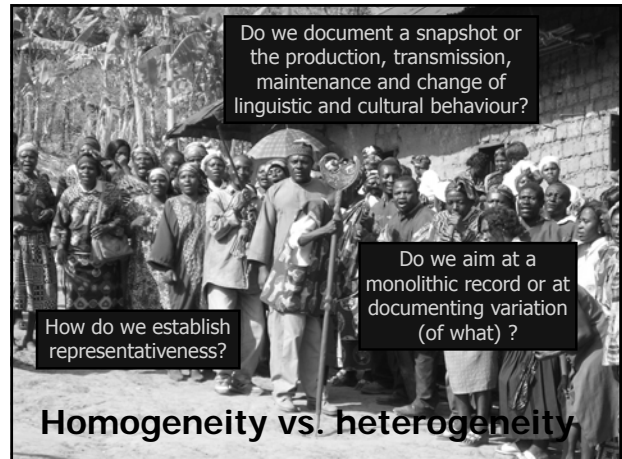
6

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)



7



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?

9

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?

10

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?

11

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: Linguistist 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?

12

Four paradoxes in language documentation

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

Why not?

- 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.
- 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.

15

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.

16

... but don't have a concept of them

Why not?

- 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.
- 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.

17

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.

18

... 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.

19

We document aspects of the natural environment and material culture

- Our lexica contain names of plants and their parts, animals, etc.
- We also collect names of items of material culture.



Photos of a leaf and an agricultural tool taken in the Bainouk language area in Senegal, 2008.

20

... but we lack the time and tools to analyze them

- Field-based lexica and dictionaries contain numerous entries of the kind 'plant species', 'bird species', etc.
- We often only have the time to find mnemonic labels ('hoe') for items of material culture, without having the time to dedicate ourselves to a detailed description of their use.

21

Established areas of linguistic interests in 'culture'



Ethnography of speaking/anthropological linguistics

- **S**etting (location, time, ...)
- **P**articipants
- **E**nds / purpose / function
- **A**ct sequence and content
- **K**ey / tone
- **I**nstrumentalities (means / channel)
- **N**orms of interaction and interpretation
- **G**enre / type of speech event

Hymes (1962): components of a speech event

23

Linguistic relativity

- The strong deterministic (Sapir-Whorf) hypothesis:
 - Language influences the way we think/perceive.



~~Blue~~

If languages have no separate words for blue and green, speakers don't perceive blue and green as different.



~~Green~~



Grue

24

Documentation of material culture

■ Example: Archi dictionary:

– <http://www.smg.surrey.ac.uk/Archi/Linguists/index.aspx?LE=112&WE=1680>

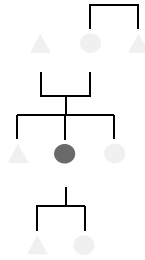
– Photos establish denotation/real-world reference of cultural objects.



k'ič̣ 'a ring and a metal "tongue" used with a padlock'

25

Kinship terms



- Important field of research in (linguistic) anthropology
- Existence of kinship terms and (arguably) the basic components of their analysis are universal
- The meanings of kinship terms differ widely from language to language
- Kinship terms often reflect behaviour/roles associated with certain kin.
- Kin relations have an impact on language use (taboos, avoidance registers, honorifics, joking relationships, etc.)

26

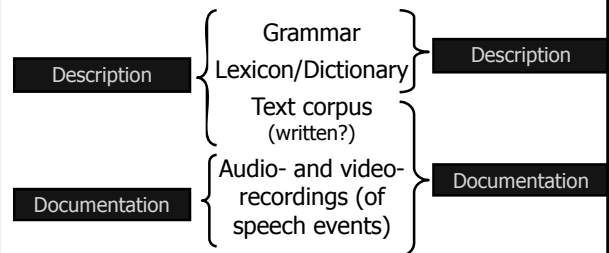
Documentation vs. description



Two views

The traditional view (cf. Krauss 1992)

Documentary Linguistics (cf. Himmelmann 1998, Woodbury 2003)



28

Importance of content

The traditional view

Importance
– useful for comparative/theoretical linguistics
– worthy of publication and academic credit
– etc.



Grammar
Lexicon/Dictionary
Text corpus (written?)
Audio- and video-recordings (of speech events)

Documentary Linguistics

Importance
– useful for the speech community
– useful for a wide range of academic disciplines
– allows for various (later) analyses and their verification

29

LDD and the Saussurean dichotomy

- Langue (language system)
 - “Competence”
 - Not amenable to direct observation
 - Enables a Speaker to produce, and a Hearer to decode individual speech events
- Parole (speech event)
 - “Performance”
 - Amenable to direct observation (documentation)
 - Enables an analyst or a language learner to draw (explicit or implicit) generalisations and conclusions as to the language system (description)
 - ... if observed in sufficient quantity and quality

30

Description vs. typological study

- “[...]the categories of language structure are language-particular” (Haspelmath 2007: 121)
- “Instead of fitting observed phenomena into the mould of currently popular categories, the linguist’s job should be to describe the phenomena in as much detail as possible, using as few presuppositions as possible.” (Haspelmath 2007: 125)

?

31

Your turn

- Please form three groups and take five minutes to discuss the following question:
 - Can you see other relationships between language documentation and description than the ones mentioned so far?
 - If yes, which are they?

32

How to represent speech



How to represent (what of) speech

- How can speech be transcribed and represented?
- There are no spaces and punctuation marks in spoken language!
- How do we decide on word boundaries and clause boundaries?

34

Transcription

Jaminjung (Mindi family, N. Australia, data from Eva Schultze-Berndt):

Orth: bugu mulurrng ganamany

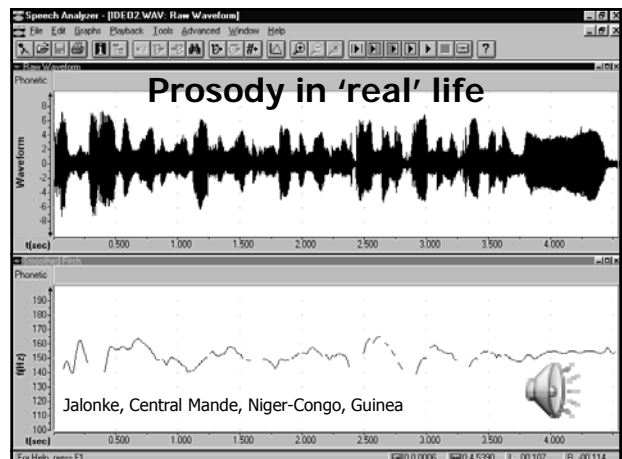
Pros: eXjX#† p XαuQ -jdqdp d[⊙]#_

bugu dibard garumany
yinthuwurlawung

eXjX#† ɟledʲ ɛjdǎxp d[⊙]#† a
nīqWxz xα-z XQ _

‘it just crashed on the ground
(and he) just came jumping over here’

35



The primacy of spoken language...

What about visual information?

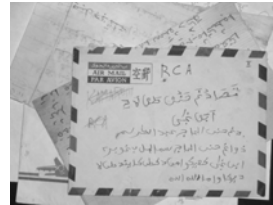


- Advantages of video recording:
 - Gestures and facial expressions can be captured
 - Position of interlocutors and referent objects can be (partly) captured
 - More interesting for speakers and non-linguists
- Disadvantages:
 - Higher cost, very large files (if digitized)
 - No software (yet) for easy transcription
 - More intrusive?

37

What about writing?

- Written communicative event if applicable (note, graffiti, sign, letter, newspaper article, essay ...)
- Transcription of unrecorded communicative event
 - Equipment failure
 - Equipment not appropriate, e.g. recording felt to be too intrusive
 - Equipment not at hand (driving, hunting, at night ...)



38

Segmentation

- For transcription of spoken language, a division into intonation units may be more appropriate than a division into clauses and sentences "as in written language"
- Intonation unit – preliminary definition:
 - coherent intonation contour
 - contains at least one pitch accent
 - delimited by boundary intonation (a rising or falling pitch movement)
 - prototypically, but not always, delimited by a pause (cf. e.g. Cruttenden 1997, Halliday 1985, Chafe 1994, Ladd 1996)

39

Translation

- Object language vs. Metalanguage(s)
- Which language(s)?
- Possibilities
 - second language of speaker(s)
 - regional language
 - national language
 - standard language (e.g. Standard German)
 - native language of compiler
 - language of academic affiliation (English in our case)
 - academic *lingua franca* (English!?)

40

Translation

- Free rendition (idiomatic in metalanguage)?
- Literal rendition (emulating object language)?
- Both?
- Solution for linguists: two layers of translation
 - Interlinear translation (glossing): word-by-word or morpheme-by-morpheme
 - Free translation

41

Data about data



42

Commentary & metadata

- Commentary:
 - Ethnographic background
 - Nonverbal context
- Metadata (data about data):
 - Cataloguing - information on language, speakers, collector, place, time, etc.
 - Descriptive - information about genre, text type, relation to other resources, etc.
 - Structural metadata - information about the structural organisation of the data (levels of annotation, etc.)
 - Technical - information on data format and quality
 - Administrative - information on tasks performed on data, access rights, etc.)

43

General information

- General information on the language
 - Genealogical affiliation (Language Family) - CAUTION
 - Number of speakers - CAUTION
 - Sound system / Orthography used
 - General grammatical information (Sketch Grammar)
 - Abbreviations used in interlinear glossing Bibliography (other references on the lg)

44

Ethnographic information

- Ethnographic background
 - Geographical location, climate, vegetation
 - Means of subsistence
 - Kinship system
 - Special speech styles
 - ...

45

Information on speakers

- Speakers (CAUTION - privacy rights)
 - Full name(s)??
 - Photo? (
 - Age, Gender, (Marital status)
 - Genealogy/relation to other speakers (*important for levels of politeness, avoidance registers, etc.!*)
 - Place of birth / Settlement
 - Linguistic background
 - all languages spoken / familiar with
 - lg(s) of parents
 - patterns of lg use
 - Education / Profession
 - Areas of special expertise
 - etc.

46

Information on the research

- Profile of documenter(s):
 - Linguistic / ethnic / educational background
 - Fieldwork experience / experience in transcription
 - Familiarity with the language (native speaker, fluent, learner, linguistic survey...)
 - Relationship / degree of acquaintance with speaker(s)
- Period of time over which documentation took place
 - Any special circumstances

47

Outlook

"... there may be imperfect consensus as to what constitutes adequate description. Description sufficient for a phonetician's purposes may not satisfy a syntactician, and a typologist's broad schematizations will not necessarily meet the requirements of a morphophonologist. All of these specialists' concerted efforts will certainly leave sociolinguists and linguistic anthropologists dissatisfied, and ethnic groups hoping to revive a more or less moribund heritage language will likewise be ill served by the narrowness of most structural accounts. (...) the investigations need to widen so as to encompass more of the social and cultural as well as the structural range that each language uniquely represents."

Dorian 2002: 138f.



Documentation

48