

DOCUMENTARY LINGUISTICS I

prof. Nicole Nau, UAM winter 2018/2019

Third lecture
16 October 2018

TOPICS OF THE DAY

- ❖ Basics of language documentation: review and expansion
- ❖ Briefly about metadata (more will come later)
- ❖ A look into an archive
- ❖ Your first task

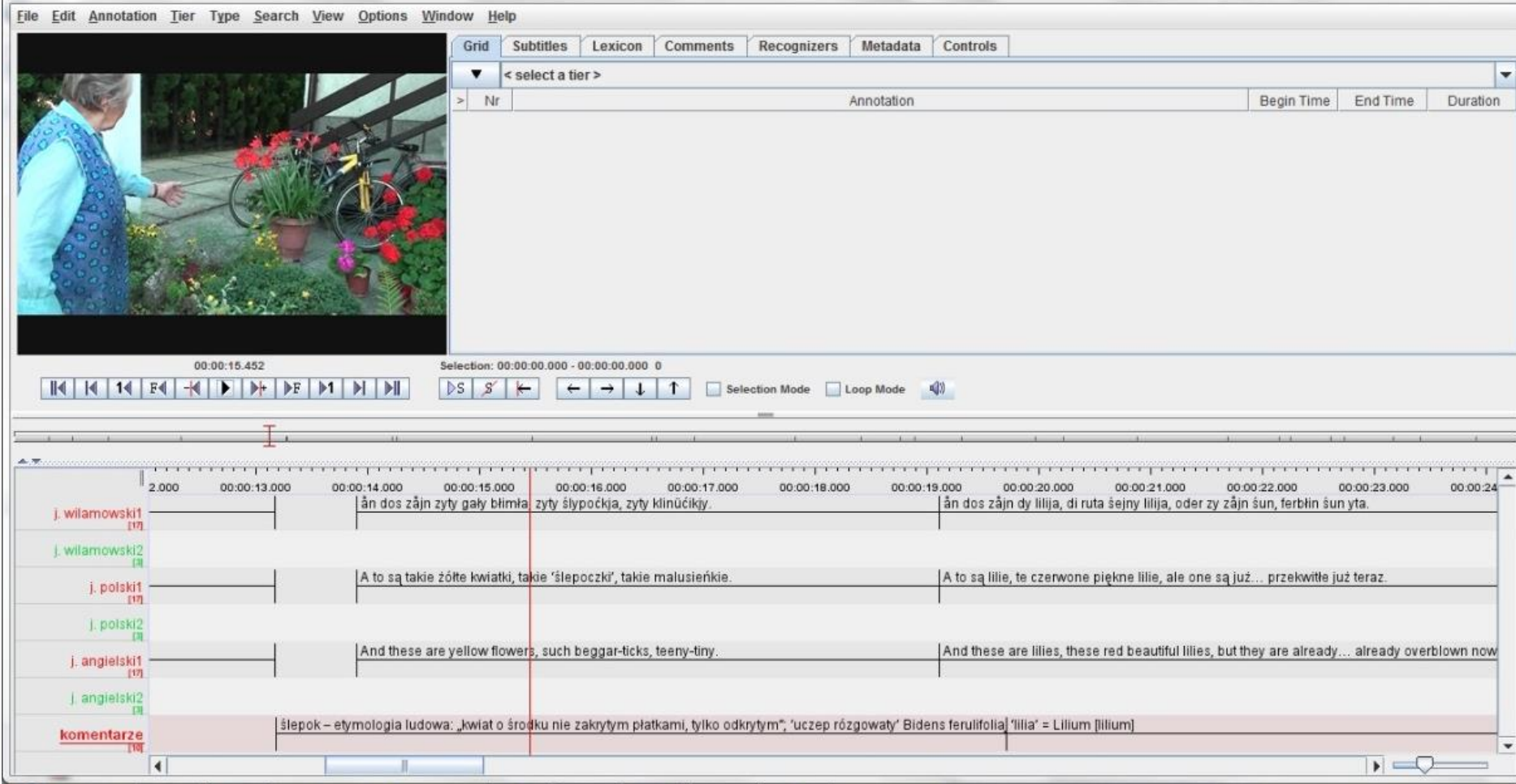
LANGUAGE
DOCUMENTATION IN
THE PAST
(but still useful today)

Private collection of Latgalian
plant names



LANGUAGE DOCUMENTATION TODAY

Video on plant names in Wilamowicean with annotation in ELAN (downloadable at: <http://inne-jezyki.amu.edu.pl/Frontend/TextSource/Details/138>)



The screenshot displays the ELAN software interface. At the top left is a video window showing an elderly woman in a blue patterned vest standing in a garden with various flowers and a bicycle. Below the video is a control bar with playback buttons and a selection bar. The main area is a grid for annotations, with columns for 'Nr', 'Annotation', 'Begin Time', 'End Time', and 'Duration'. The grid is divided into several tiers:

- j. wilamowski1**: Contains two lines of Wilamowicean text: "ńń dos zńjn zyty galy blimńa zyty ślypockńja, zyty klinuńciky." and "ńń dos zńjn dy lilija, di ruta śejny lilija, oder zy zńjn śun, ferblń śun yta."
- j. wilamowski2**: Empty.
- j. polski1**: Contains two lines of Polish text: "A to sń takie zńite kwiatki, takie 'ślepockńki', takie malusieńkie." and "A to sń lilie, te czerwone piķne lilie, ale one sń juź... przekwitte juź teraz."
- j. polski2**: Empty.
- j. angielski1**: Contains two lines of English text: "And these are yellow flowers, such beggar-ticks, teeny-tiny." and "And these are lilies, these red beautiful lilies, but they are already... already overblown now"
- j. angielski2**: Empty.
- komentarze**: Contains a line of commentary: "ślepok – etymologia ludowa: 'kwiat o środku nie zakrytym płatkami, tylko odkrytym', 'uczep rńzgowaty' Bidens ferulifolia | 'lilia' = Liliu[m] | liliu[m]"

WHAT DO WE DOCUMENT WHEN DOCUMENTING A LANGUAGE? WHAT DOES A LANGUAGE DOCUMENTATION CONTAIN?

❖ Primary data

- ❖ Observable linguistic behavior, speech events, communicative activities = how people use their language
- ❖ Metalinguistic knowledge = what people know about their language and its use

❖ Metadata

❖ Annotation

-
- ❖ Description (results of analysis: grammar, dictionary)

DID YOU UNDERSTAND HIMMELMANN'S TEXT?

1) WHAT IS "**ELICITATION**"? QUOTES:

"The primary data which constitute the core of a language documentation include audio or video recordings of a communicative event (a narrative, a conversation, etc.), but also the notes taken in **an elicitation session**, or a genealogy written down by a literate native speaker."

"But very often documenting metalinguistic knowledge will involve the use of a broad array of **elicitation strategies**, guided by current theories about different kinds of metalinguistic knowledge and their structure."

"One very important type of **elicited evidence** are monolingual definitions of word meanings provided by native speakers."

"it is now standard practice to make (video) recordings of observable linguistic behavior, while for **the elicitation of metalinguistic knowledge** it is still more common simply to take written notes."

DID YOU UNDERSTAND HIMMELMANN'S TEXT?

2) WHAT IS "GRAMMATICALITY"? QUOTE:

"The documentation of metalinguistic knowledge as understood here includes much of the basic information that is needed for writing descriptive grammars and dictionaries. In particular, it includes all kinds of elicited data regarding **the grammaticality or acceptability** of phonological or morphosyntactic structures and the meaning, use, and relatedness of lexical items."

"One aspect of "a language" that is not, or at least not easily, accessible by analyzing observable linguistic behavior is the tacit knowledge speakers have about their language. This is also known as **metalinguistic knowledge** and refers to the ability of native speakers to provide interpretations and systematizations for linguistic units and events."

DID YOU UNDERSTAND HIMMELMANN'S TEXT?

3) WHAT IS "ETHNOBOTANY"? QUOTE:

"Documentary work that aims at a truly comprehensive record of a language also has to engage with **ethnobotany**, musicology, human geography, oral history, and so on."

Why?

FURTHER QUESTIONS (FROM THE HOMEWORK AND LAST LECTURE)

- ❖ For whom and for which purposes are languages documented?
- ❖ Why should native speakers take an active part in documenting a language?
- ❖ Why is it not possible to record all communicative events in a given speech community?
- ❖ Why is it important to store primary data in open archives?

DID YOU UNDERSTAND HIMMELMANN'S TEXT?

4) WHAT ARE "FALSIFIABILITY" AND "REPLICABILITY"?

"Finally, establishing open archives for primary data is also in the interest of making analyses accountable. Many claims and analyses related to languages and speech communities for which no documentation is available remain unverifiable as long as substantial parts of the primary data on which the analyses are based remain inaccessible to further scrutiny. Accountability here is intended to include all kinds of practical checks and methodological tests with regard to the empirical basis of an analysis or theory, including **replicability** and **falsifiability**."

FURTHER QUESTIONS (FROM THE HOMEWORK AND LAST LECTURE)

- ❖ What is the relationship between documentation and description?
- ❖ Why are grammars and dictionaries NOT “lasting, multipurpose records” of a language? But why are they important for a language documentation (in a broader sense)?

DID YOU UNDERSTAND HIMMELMANN'S TEXT?

5) WHAT IS "NEGATIVE EVIDENCE"?

"The major counterargument against this position would be the claim that actually producing a descriptive grammar is a necessary part of a language documentation because otherwise, essential aspects of the language system would be left undocumented. The evaluation of this claim rests on the question of whether there is some kind of important evidence for grammatical structure which, as a matter of principle, cannot be extracted from a sufficiently large and varied corpus of primary data as sketched in Section 3 above. As far as I am aware, there is especially one type of evidence of this kind, i.e. **negative evidence**. Obviously, illicit structures cannot be attested even in the largest and most comprehensive corpora."

DID YOU UNDERSTAND HIMMELMANN'S TEXT?

6) WHAT IS "STRUCTURALIST LINGUISTICS"?

"The **structuralist idea of language as an abstract system** has been articulated in a variety of oppositions including the well-known Saussurean distinction of *langue* vs. *langage* vs. *parole* and the Chomskyan distinction of *competence* vs. *performance*."

"In line with **the structuralist conception of the language system**, grammars and dictionaries contain abstractions based on a variety of analytical procedures. With the data contained in grammars and dictionaries, most aspects of the analyses underlying the abstractions are not verifiable or replicable. There is no way of knowing whether fundamental mistakes have been made unless the primary data on which the analyses build are made available *in toto* as well."

"As pointed out in particular by Andrew Pawley (1985, 1993, and elsewhere), there is a large variety of linguistic structures often subsumed under the heading of *speech formulas* which do not really fit **the structuralist idea of a clean divide between grammar and dictionary** and thus more often than not are not adequately documented in these formats."

LET'S GO ON: WHAT IS METADATA, AND WHAT IS IT GOOD FOR?

Metadata in the broad sense / Metadata vs. annotation

Annotation of primary data may contain

- transcription / transliteration
- translation(s)
- grammatical annotation («glossing», «tagging»)
- comments
- ...

Tools for annotating language data: ELAN, AnnotationPro, Flex, Toolbox, ...

METADATA: MINIMUM ACCORDING TO JOHNSON (2004) – TEXT FOR NEXT WEEK

- ❖ creators' full names
- ❖ name of the language
- ❖ date of creation
- ❖ place of creation
- ❖ access restrictions
- ❖ genre keyword

TYPES
OF
METADAT
A WITH
EXAMPLE

(AUSTIN
2006)

Table 1. Different types of metadata associated with a computer file

Cataloguing	Title: Sasak.dic; Collector: Peter K Austin; Speakers: Yon Mahyuni, Lalu Hasbollah; Language code: SAS
Descriptive	Trilingual Sasak-Indonesian-English dictionary, linked to finderlists, morpheme forms link to Sasak text collection
Structural	Dictionary entries with headword, part of speech, gloss in Bahasa Indonesia and English, cross-references for semantic relations; SIL FOSF record format
Technical	Shoebox 5.0 ASCII text file
Administrative	Open access to all; Last edited version dated 2004-06-25; backup 2004-06-20 on DVD 012

WHAT ARE METADATA FOR?

- ❖ For users of the archive: finding and selecting records

«Metadata, or catalogue information, is what makes discovery possible.»
(Johnson 2004)

- ❖ For later generations: have maximal information about the record

«Metadata catalogue information is especially vital for digital materials, because they are not amenable to direct inspection, as is a book or other printed matter.» (Johnson 2004)

- ❖ For documentators: keep your collection in order!

- ❖ For archivers: structure the archive in a logical way

METADATA ARE NEEDED ON VARIOUS LEVELS!

For each recording, metadata usually contain information on:

- participants (speakers and bystanders, their roles)
- time and location
- recorded by, recorded with
- ...

Speaker metadata: age, sex, ..., dialect

Metadata for the whole documentation: information on the language

Table 2. Extended format for a language documentation

Himmelman (2006)

Primary data	Apparatus	
recordings/records of observable linguistic behavior and metalinguistic knowledge	Per session	For documentation as a whole
	<p><i>Metadata</i></p> <p><i>Annotations</i></p> <ul style="list-style-type: none"> – transcription – translation – further linguistic and ethnographic glossing and commentary 	<p><i>Metadata</i></p> <p><i>General access resources</i></p> <ul style="list-style-type: none"> – introduction – orthographical conventions – glossing conventions – indices – links to other resources ...
		<p><i>Descriptive analysis</i></p> <ul style="list-style-type: none"> – ethnography – descriptive grammar – dictionary

A) YOUR FIRST TASK FOR GRADING => ELECTRONIC HANDOUT

B) Reading for the next two weeks (23. + 30.10.)

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. *Language Documentation and Description*, ed. Peter K. Austin, vol. 2, 140-153. London: SOAS. Available at:
<http://www.ejournals.org/PID/026>.

Mosel, Ulrike. 2006. Fieldwork and community language work (in *Essentials of Language Documentation*)

LANGUAGE ARCHIVES

<http://dobes.mpi.nl/> (DOBES = **D**okumentation **bed**rohter **S**prachen)

<https://elar.soas.ac.uk/> (ELAR = **E**ndangered **L**anguages **A**rchive)

<https://www.ailla.utexas.org/> (**A**ILLA is a digital archive of recordings and texts in and about the indigenous languages of Latin America)

<http://catalog.paradisec.org.au/> (**PARADISEC** = Pacific And Regional Archive for Digital Sources in Endangered Cultures)

http://lacito.vjf.cnrs.fr/pangloss/index_en.htm (**PANGLOSS** collection)

<http://siberian-lang.srcc.msu.ru/> Siberian Lang (МАЛЫЕ ЯЗЫКИ СИБИРИ: НАШЕ КУЛЬТУРНОЕ НАСЛЕДИЕ)

<http://inne-jezyki.amu.edu.pl/Frontend/> Poland's Linguistic Heritage