

Data Collection Management & Analysis in the Humanities

Katarzyna Klessa



UNIwersYTET
IM. ADAMA MICKIEWICZA
W POZNANIU

Who can be interested in data collection & more?

Linguists Phoneticians Psychologists
Teachers Archivists Psycholinguists
(Speech) Therapists Sociologists Students
Historians (Language) Documenters Lawyers
Speech Technology Specialists



.....

.....

.....

Diversity: challenge & chance

It is precisely in the challenge of this divergence between disciplines that one might expect to arrive at a more significant reward

– that the contributions of each might ultimately be greater than the sum of their parts.

Cox, Ch. (2011). Corpus linguistics and language documentation: challenges for collaboration. *Language and Computers-Studies in Practical Linguistics*, 73(1), 239.

Tasks: First glance

Collect

Data

Texts, audio/video recordings, “dark data”

Metadata

“Data about data”, additional information about the materials, speakers, environment...

Manage

Technology

- Hardware and software tools, data management solutions (file collections, relational databases, SQL)
- Backup copies
- Sharing options

Analyse

Description

annotation, transcription => interpretation

Processing

Data & metadata processing

Exploration

Information extraction, data mining

Workshop survey

Data Collection
Management
& Analysis *in the Humanities*

Development & usage of LRE. Workshop Participant Form, July 2017

Proszę wybrać jedną lub więcej odpowiedzi

*Required

Your experiences so far - using language resources *

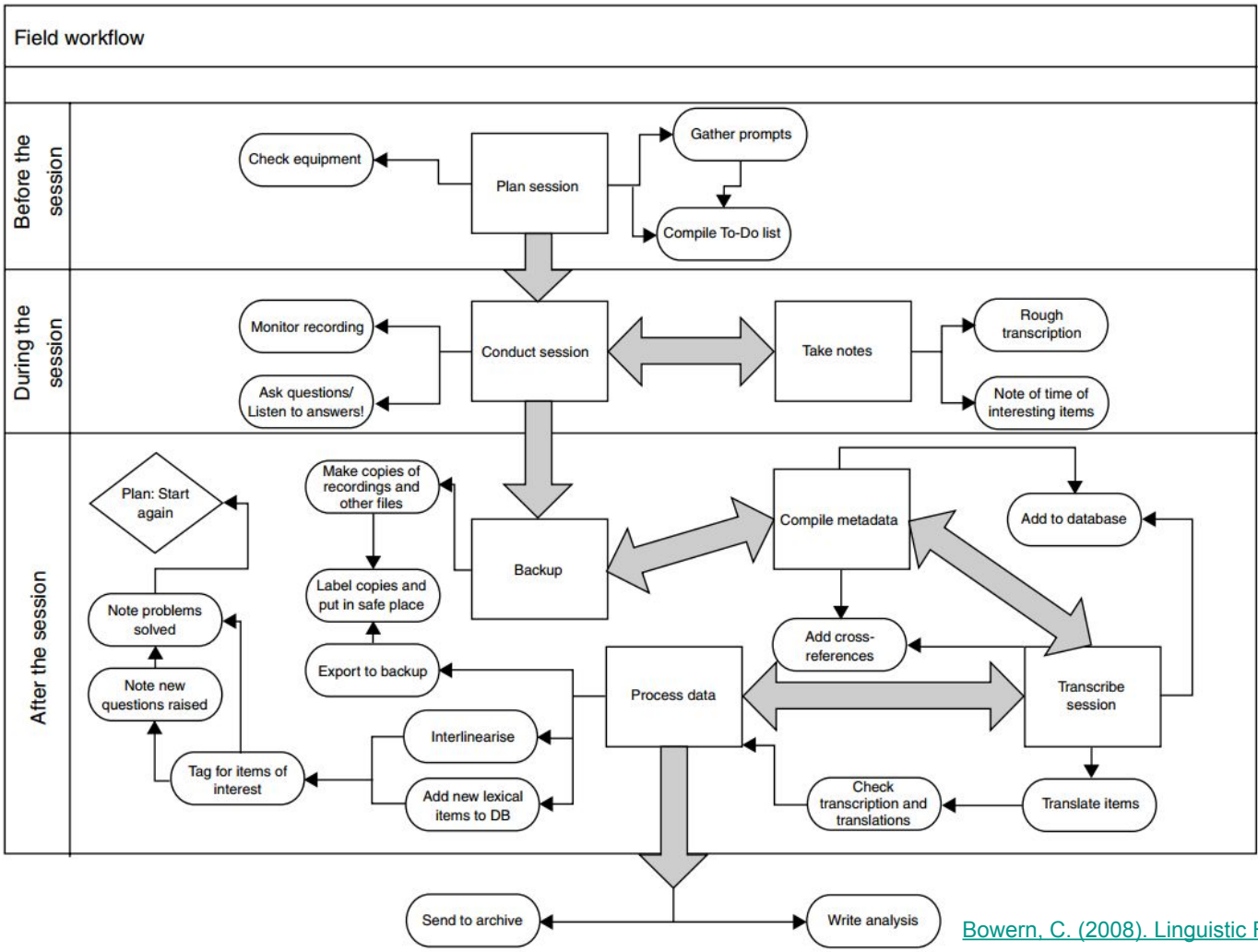
- Dictionaries, lexica, encyclopedias
- Language learning language resources

Survey link:

<https://goo.gl/forms/X2gh1ZvgSErr43VV2>

Workshop survey: not only the target data

- information search, e.g., analysis of the state of research, raw data analyses & search
- collecting information ABOUT the data to be collected (metadata)
- designing scenarios for a recording session or a perception experiment
- designing data access and data sharing rules and schemes
- the choice of data management technology
- the choice of equipment and software for data collection, e.g. recording voices
- legal aspects: licences, agreements, consulting lawyers, designing documents and forms
- recruiting participants (speakers, listeners, survey respondents)
- conducting the main experiment - speech recording (audio or video), perception test session
- managing the participant survey completion (e.g., an on-line survey for collecting participant metadata)
- relational database or server management tasks
- creating backup copies
- phonetic transcription of speech recordings
- annotation of text data, e.g. morphological glossing or other type of annotating text data
- speech segmentation (e.g., time-aligned segmenting speech signal into phones, syllables, words)
- corpus annotation management
- corpus annotation processing (e.g., converting annotations into other annotation or table formats)
- corpus annotation mining (information extraction & analysis)
- Other: _____



Steps in collecting data

Data standards

File formats, properties of the speech signals - data size, quality, minimal requirements (e.g. chapter 2 [here](#)).

Collection

Recording equipment, data management & storage devices, working checklists.

Verification

Data verification, processing (e.g., degrading multimedia formats for the purposes of online collaboration), formats adjustments...

Analysis

Presentation, analysis, sharing...

Steps in collecting metadata

Metadata standards

What kind of metadata is needed? A list of questions based on standards e.g., [Dublin core](#).

Collection

Acquisition & (preferably!) digital storage of answers, e.g. an electronic survey form.

Verification

Metadata verification, processing, formats adjustments...

Analysis

Presentation, analysis, sharing...

Multi-tasking, mutual dependencies & **re-usability**



Metadata - Exercise



Let's suppose that we wish to collect recordings of old songs.

What kind of information would we wish to preserve - besides the recordings themselves?

One of the earliest music recordings here:

https://upload.wikimedia.org/wikipedia/commons/8/86/Kham_Hom_-_Sweet_Words.ogg (read more)

Metadata - Exercise



One of the earliest music recordings.

Here:

https://upload.wikimedia.org/wikipedia/commons/8/86/Kham_Hom_-_Sweet_Words.ogg

What would we like to know about the recording? Where is the metadata?

[Read more](#)

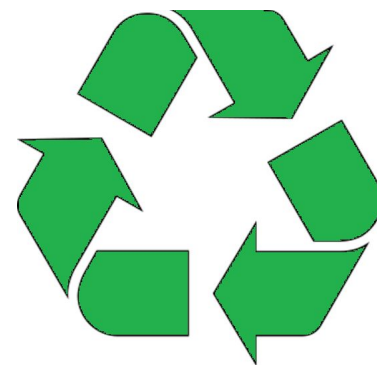
Metadata - Exercise

- title of the song
- performer/singer (name and maybe more data - in a separate subset)
- date of recording
- length/duration (seconds)
- when was it recorded?
- where was it recorded?
- category (e.g., religious, wedding)
- original key
- language/dialect/whatever
- ... ???

Note: Details for some of these metadata fields may be unavailable for each of the items included in the collection...

Metadata & **re-usability**

- An efficient metadata system may bring about more re-usability, also for the users from new other areas of interest, e.g. recordings made for phonetic studies may serve as useful resource for psycholinguists, sociologists, speech therapists, speech technology...;
- Proper description & documentation -> better usefulness;
- Sharing data vs. commercial uses -> awareness of legal issues is needed.



Data acquisition

Scenarios, issues
& example solutions

- What are the possible approaches to collecting data?
- What are the crucial notions, problems/issues?
- Is it always straightforward to collect the data strictly corresponding to our needs?
- Can we be sure what's in the data?
- Should we be afraid of “dark data”?

From raw data to a corpus

Raw data

Data collected randomly or without any strictly defined planning, not ordered.

The same set of raw data can potentially serve as a source collection for various types of analyses conducted by researchers specializing in a wide range of fields.

Corpus

A (usually structured) set of data which have been collected **on purpose** of studying certain phenomena occurring in the domain of interest. A linguistic corpus may be dedicated to enable studies of language in general, dialect or other 'sub-language'.

What do we
actually learn
based on
corpus data?



- How general can our conclusions be when derived from a certain type of data?
- We should always be aware that the choice of our data, the way we approach them can significantly influence the results obtained.

Legal issues in data acquisition

In most countries it is illegal to collect data without the consent of the copyright owner(s); e.g. speech recordings, even with respect to own own conversations with third parties.

TIP

Audio/video recording consent

The text of the consent should be clear, free from specialized terminology. Write the text in a way that by signing it the participant:

- confirms that s/he voluntarily agrees to participate in the recording session

can give separate consent for:

- all modalities of the recording (audio / video)
- the usage of the data in research studies
- publication
- archiving of the data

Consider creating 2 copies of the consent form for yourself and for the participant.

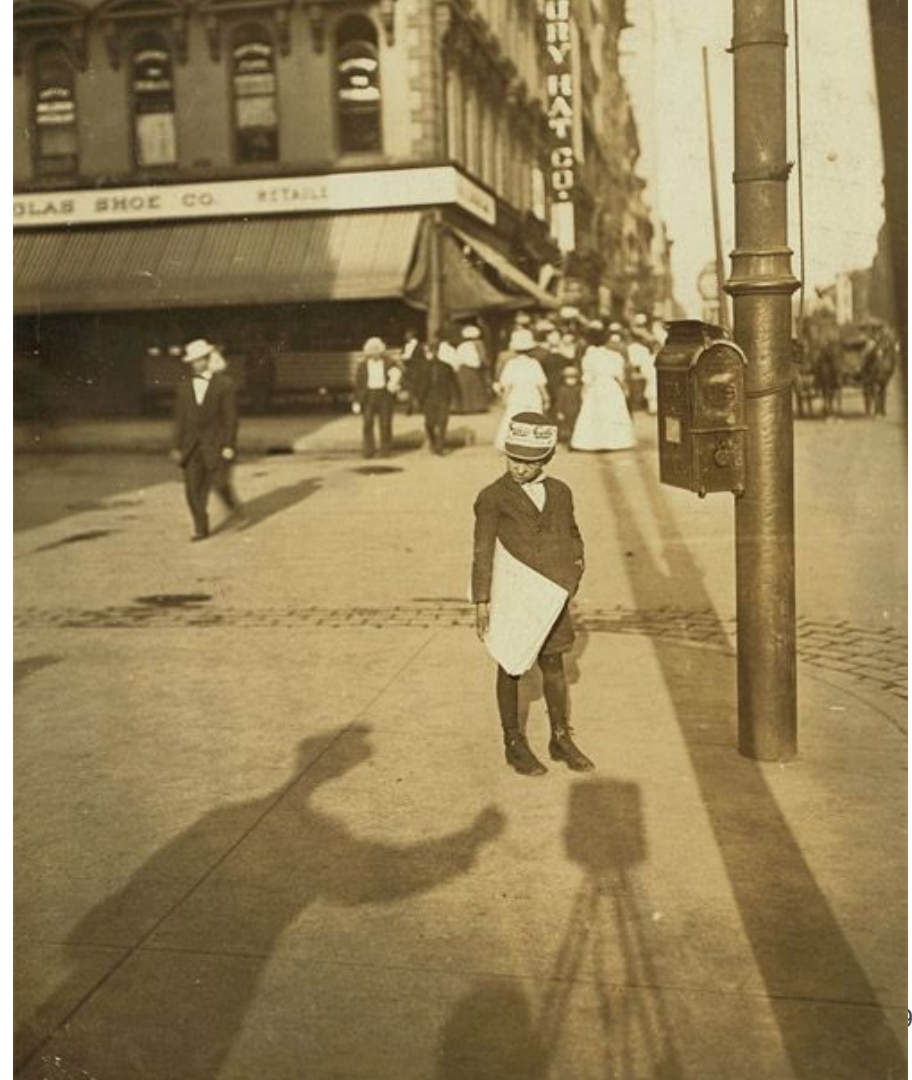
Remember the signatures!

The Observer's Paradox

The aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation.

W. Labov

Example: how to study human communication?



(Semi)spontaneous

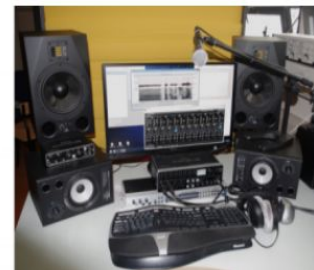
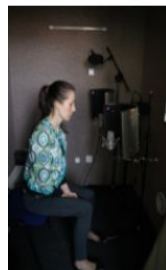
- Conversational speech, dialogues
- Narratives, story-telling
- Can be obtained in real-life situations, during fieldwork or pre-arranged conditions, also in a recording studio

Controlled

- Read speech
 - isolated utterances
 - word-lists
 - continuous texts
- Elicited dictation
 - e.g. speech (re)produced based on stimuli presented to speakers immediately before recording
- Often collected in studio

Quality

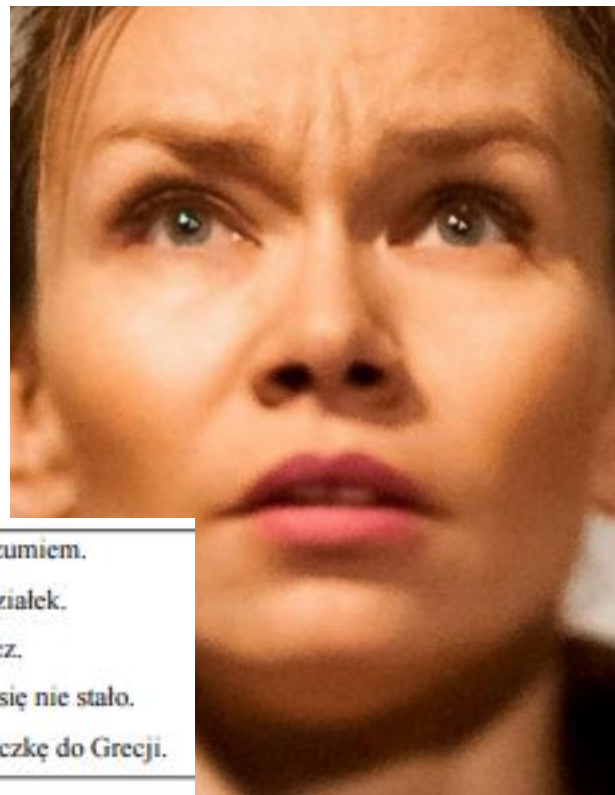
- **Studio** recordings vs. **fieldwork**
- High quality vs. spontaneous, “natural” character



Scenarios example: emotion portrayals

How to investigate emotions in speech?

- Polish *Paralingua* corpus -> Emotion portrayals, por. GEMEP Geneva corpus (Scherer et al.)
- Lexically neutral utterances produced with varying emotional load
- Basic emotions characterized by varying valence & activation



Scenarios example: authentic motherese

What are the specific features of speech directed to infants in comparison with adult directed speech?

- Mothers talking to their babies (below 1 year-old)
- Home recording conditions, during everyday activities, feeding the infants, dressing up, playing with them, etc...
- Subsequently: studio recordings, but....!



Scenarios example: task-oriented dialogues

A compromise?



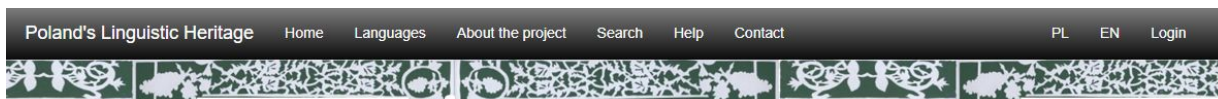
Borderland (<http://borderland.amu.edu.pl/>) a corpus of Polish & German speech:

- cooperative / competitive dialogues: let's built an imaginative tower!
- let's decide about a gift for a person (based on varying input info)!
- The focus of the speaker is **on the task rather than being recorded**, at least that's what is intended -> let the participants forget about the artificial conditions...

Another perspective - language archives

- A specific type of data collections.
- Unlike language corpora it is often characterized by **highly diversified** and **not always well-balanced** contents. The reason for this is that the contents frequently depends on availability.
- For example, in case of a repository for endangered languages, the size and complexity of the collections will be conditioned by the number of speakers, source texts available, as well as the cultural or social limitations -> some communities might be reluctant to share their heritage.

Language archives - www.inne-jezyki.amu.edu.pl



Poland's Linguistic Heritage Documentation Database for Endangered Languages

Welcome to the website of the documentation database of endangered language varieties spoken on the territory of Poland and developed in - synchronic or diachronic - language contact(s) with Polish (excluding the dialects of the Polish language itself). The focus of the present inventory is on a wide range of non-Polish languages and their non-standard varieties illustrating the richness and diversity of Poland's language landscape and the variety of its language contacts.



Chram Starowierów in Daugavpils



Daugavpils. Ms Wanda and Mr Jan



Polish gravestones

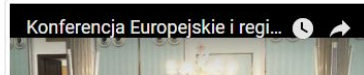
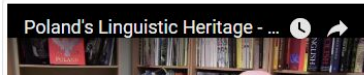


Krāsiava (Krasław) - Castle

List of languages

- Latgalian *
- (Polish) Yiddish *
- Wymysorys / Wilamowicean *
- Haknavian and Bleisko-Biała enclave *
- Armeno-Kipchak
- Belarusian dialects
- Czech dialects
- Low German

In Poland and in its neighbouring countries (once included in the territory of Poland) there are many languages spoken by small groups of speakers that have not been documented so far and they are severely endangered. These languages prove the linguistic diversity and richness of the former Republic of Poland (the Polish historical name is "Rzeczypospolita") and are an important component of the Polish national legacy. In terms of linguistic diversity, the territory of "Rzeczypospolita" is a region of contacts between various languages and communication communities. Although today the diversity is significantly impoverished as compared to earlier times, it still exists.



Data management

Selected tools & approaches

Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed do
eiusmod tempor incididunt ut
labore et dolore magna aliqua.

From file collections to relational databases

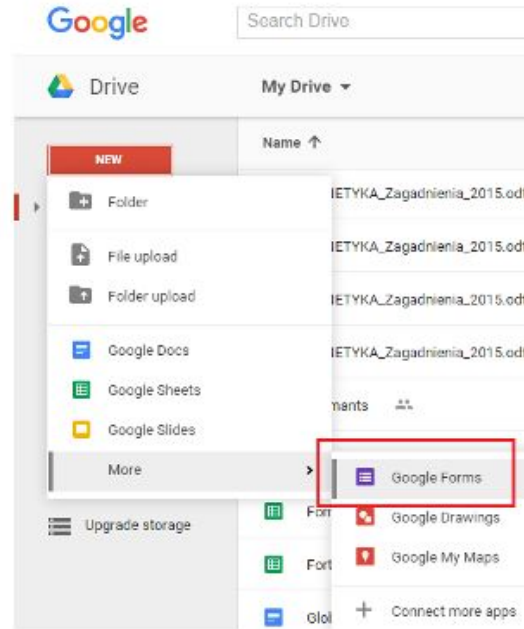
Some general purpose tools available for data management purposes.

Some management options included in annotation & analysis tools:

- file collection management,
- workspaces etc.

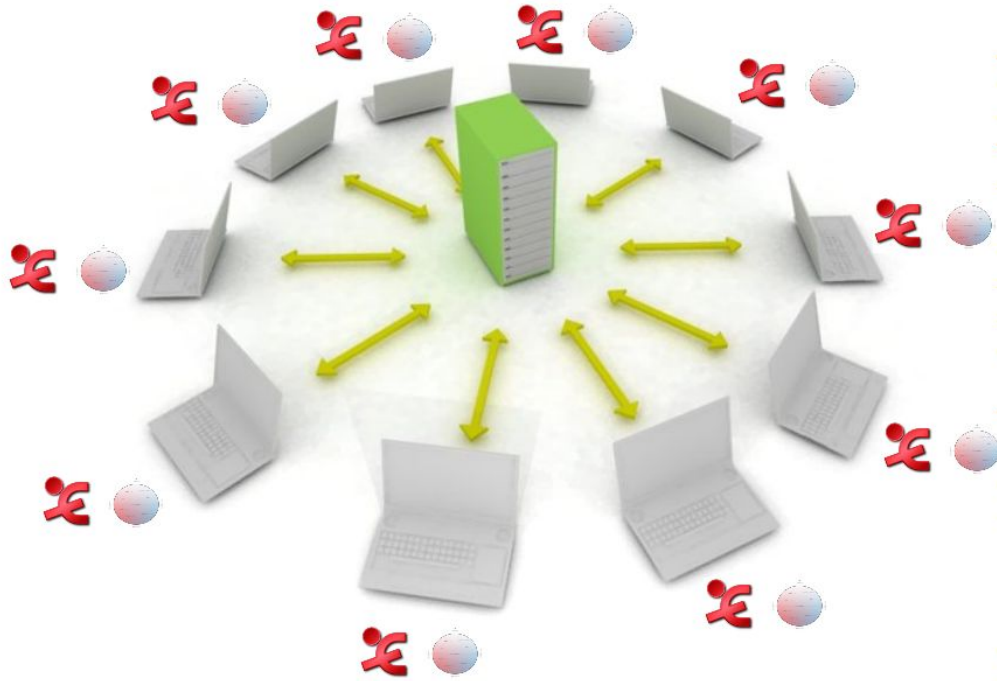
However: for **large corpora & simultaneous usage** by many people **more robust tools are needed** such as dedicated data & workflow management software + relational database solutions.

From file collections to relational databases



- Files
- Folders
- File or folder collections, calculation spreadsheets
- Relational databases

Client-server architectures: Collaboration support



SQL

/ˈɛs kjuː ˈɛl/ or /ˈsiːkwəl/,
Structured Query Language

Data analysis

Selected tools & approaches

Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed do
eiusmod tempor incididunt ut
labore et dolore magna aliqua.

Annotation, transcription, time-alignment...

Dialekt mazowiecki - Mazury

Tekst gwarowy – Biała Piska 1

PDF PRINT EMAIL

Justyna Garczyńska

Nagranie: Julia Pikacz, Anna Godziuk

Przepisanie: Monika Kresa

Opracowanie: Justyna Garczyńska

Informator: Helena Born, zamieszkała w Białej Piskiej, ur. 20.12.1933 r. w Kaliszkach. Rodzice pochodzili z Kaliszek, pracowali w majątku dziedzica, rodzina mieszkała w czworakach dworskich. Ukończyła sześć klas szkoły podstawowej niemieckiej. Pracowała w Pisku w fabryce drewna, a następnie w przedszkolu jako pomoc.

O wojnie



żeńską formą liczebnikową dwie zastąpiona przez formę męską dwa

Urodziłam się w Kaliszkach i ta... lat. No ja tam mieszkałam do czasu wojny. I po wojn..., jak w... to nazad z pegieerzu z Kalisz... to nam ten pan kazał take woz... przyszykować drabiniaste, nakładli słomy, tygo



Annotation, transcription, time-alignment

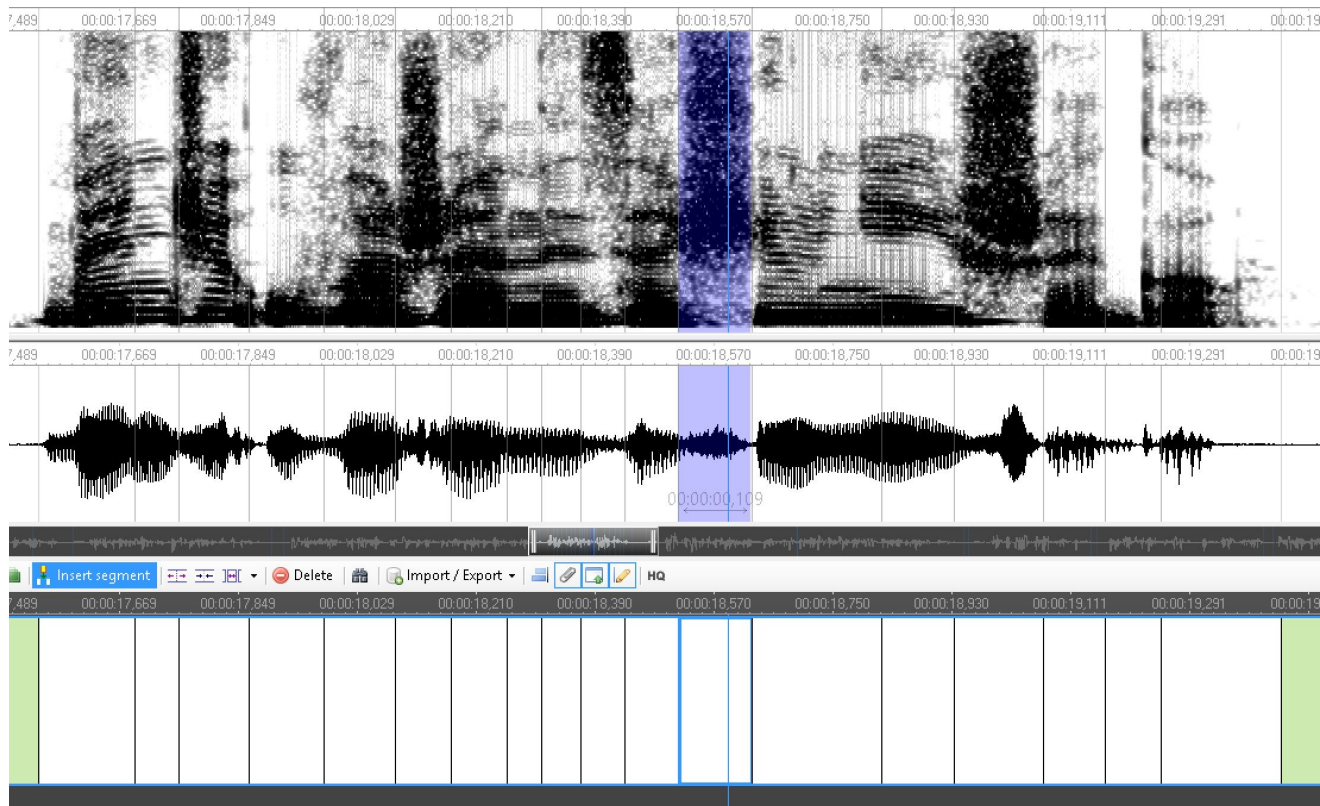
pe`vnego ,razu || puw`notsni °vjatr i,swoptse || spje>tfałice || `ktoznix
jest cil,nejfi || `vwaęne pje°xodziw °drogoŵ jacię >tjwovjek || ovi`penti
°ftęepwi ,pwaftj || `umu ,vilię °vjents || zetenznix °kturi `pjerfi zmuęi
pfexodzon,tsego || abi`zdjow o`kritę || `beņdze uva°zani zaęilnej,sego ||
puw`notsni °vjatr °zatjow od`razu °doptę s`tsawej ,ęiwi || aleim`motsnej
>dow || tim`ęilnej po°druzni o°tulawęę ,fpwaftj || `vrestęę puw°notsni
°vjatr daw,spokuj || `ftedi °swoptse za`tfewo pfi,gzevaę, || af`xfile
'puznej po'druzni zdjow,pwaftj || 'ften >sposup || puw`notsni °vjatr `muęaw
>pfiznatę || ze`swoptse jestęil°nejje od, nego

Annotation, transcription, time-alignment

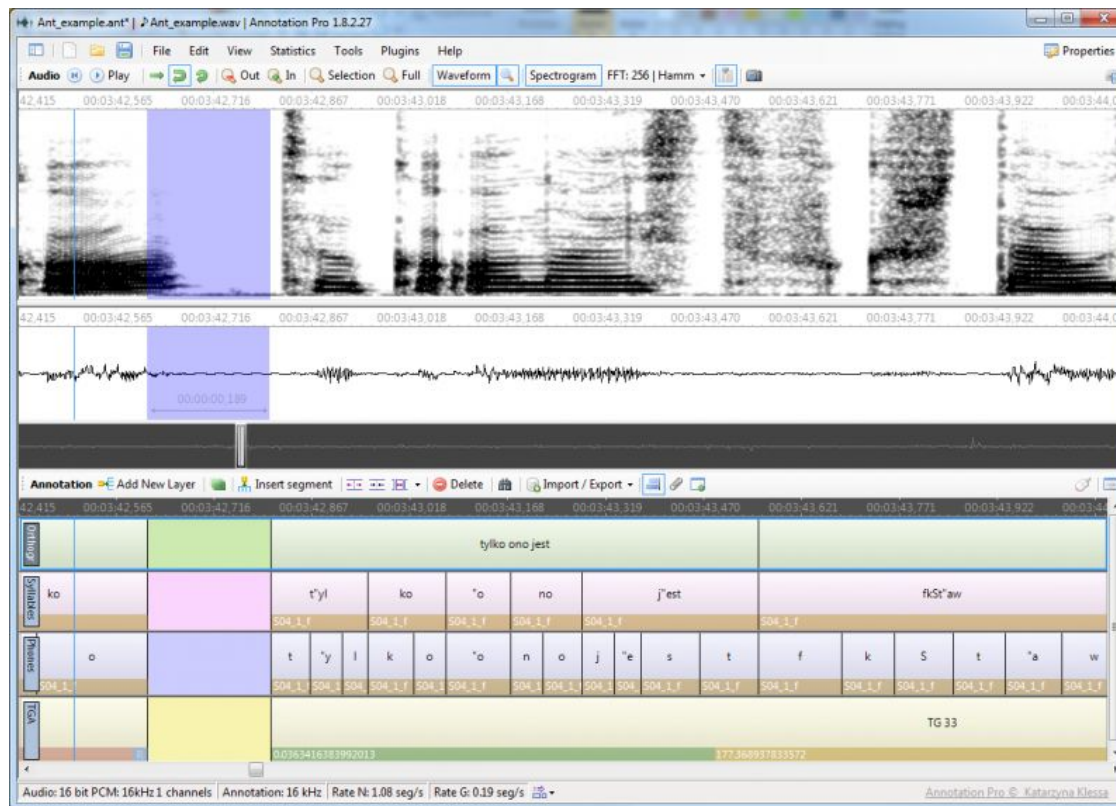
Transliteration Orthographic Polish Translation Translation EN Morphology Phonetic Comment

Position	Transliteration	Orthographic	Translation EN
1	dos iz a štikelc genumen fun canins „šlof ništ mameši”, a bux fun noveln.	דאָס איז אַ שטיקעלע גענומען פֿון צאַנינס "שלאָף נישט מאַמעשי", אַ בוך פֿון נאַוועלן	Here is an excerpt [taken] from the " Shlof Nisht Mameshi " by Canin, a collection of short stories.
2	far junge un eltere kinder cu gedenken.	פֿאַר יונגע און עלטערע קינדער צו געדענקען	For younger and older children, in memory.
3	mameši, mameši, du šlufst?	מאַמעשי, מאַמעשי, די שלופסט	Mum, mum, are you asleep?
4	šluf ništ mameši.	שלוף נישט, מאַמעשי	Do not sleep, mum.
5	jax hob gehat a malejer, der far bin jax gekumen špet, mameši.	יאָך האָב געהאַט אַ מאַלייער, דערפֿאַר. בין יאָך געקומען שפּעט, מאַמעשי	Misfortune happened to me, because I came so late, mum.
6	vi nox jax bin ariber af jeno zat, hot mio gepuct a šmalcovnik.	ווי נאָך יאָך בין אַריבער אָפּ יענאַ זאַט, האָט מיאָ געפּוצט אַ. שמאַלצאַוויניק	As soon as I walked to the other side, a smuggler grabbed me.
7	jax zol im gebn gelt, hot er gevolt.	יאָך זאָל אים געבן געלט, האָט ער געוואָלט.	He wanted me to give him money.
8	hob jax im farčaket, az ix hob man gelt ba a pžekupke afn kerceljak.	האַב יאָך אים פֿאַרטשאַקעט, אַז איך האָב מאַן געלט באַ אַ. פֿשעקופּקע אָפּן קערצעליאַק	I told him that I had got some money from a street vendor on Karcelak.

Annotation, transcription, **time-alignment**



Multilayer annotation of speech, gesture & text



Annotation Pro:

- Speech annotation
- Annotation mining
- Perception tests

www.annotationpro.org

Multilayer annotation of speech, gesture & text

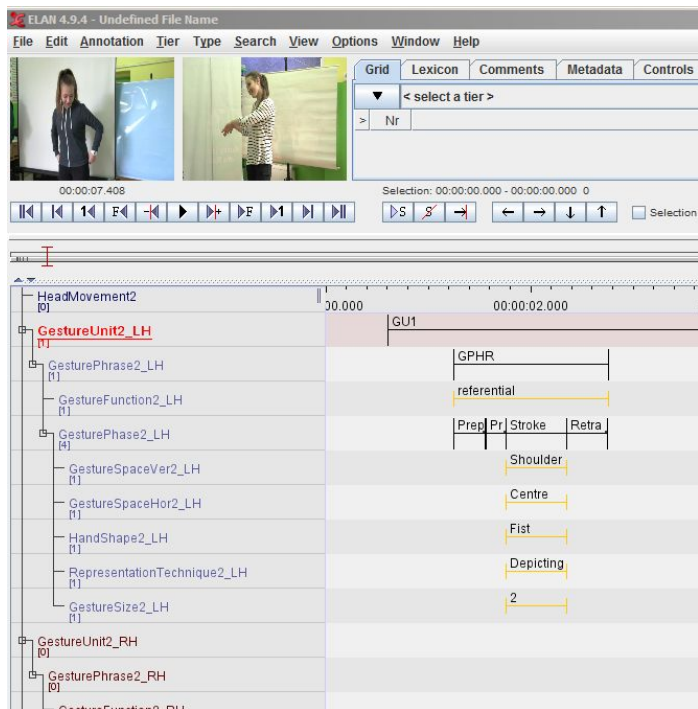


Image: *Borderland* project, M. Karpiński

ELAN <https://tla.mpi.nl/tools/tla-tools/elan/> image: *Borderland* project

Multilayer **annotation** of speech, gesture & text

Save FTrans 1 FTrans 2 CParam Base Meaning Gloss POS

Phrase: : han håper på å komme

Free translation 1: he hopes to be able to come

Free translation 2:

Constr. params: Change NP+PP[INF:equiSBJ]-propositionalAttitude-----

Word:	han	håper		på	å	komme	
Morph:	han	håp	er	på	å	kom	e
Baseform:	han	håpe	er	på	å	komme	e
Meaning:	he	hope			to	come	
Gloss tags:	SBJ.3.SG.NOM		PRES	OBL	INF		INF
POS:	PN	Vitr		PREP	COMP	Vitr	

TypeCraft annotation & text glossing tool <https://typecraft.org/>

Annotation data export, import & processing

The screenshot displays the Annotation Pro 2.5.7.0 interface. On the left, a workspace table lists audio files (k3_ang_1.ant to k3_joy_5.ant) with columns for Name, Open, Edit, Listen, and Acc. The main window shows a spectrogram and waveform of an audio file. Below these is an annotation layer with a timeline and a table of segments. The table contains phonetic segments like #t'e, #s, #y, #t, #k, #r, #o, #z, #f'e, #ras, #fs'ys, #tko, #ro, #z'u. A circular emotion wheel is visible on the right, with segments for: Irritated angry, Interested involved, Proud satisfied, Acutely in joy, Experiencing pleasure, Bored weary, Ashamed embarrassed, In despair sad, and Anxious fearful. The bottom right shows a properties panel for a selected layer, including options for color, locking, and font settings.

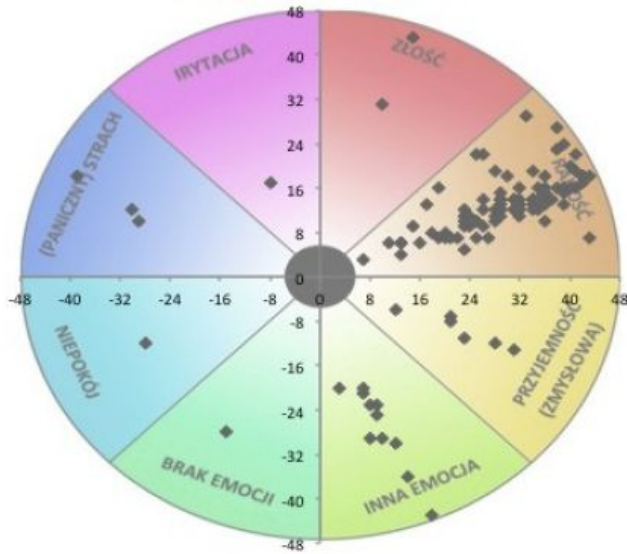
Name	Open	Edit	Listen	Acc
Ungruiped	1	0	0	--
k3_ang_1.ant	0	0	0	--
k3_ang_2.ant	0	0	0	--
k3_ang_3.ant	0	0	0	--
k3_ang_4.ant	0	0	0	--
k3_ang_5.ant	0	0	0	--
k3_joy_1.ant	0	0	0	--
k3_joy_2.ant	0	0	0	--
k3_joy_3.ant	0	0	0	--
k3_joy_4.ant	0	0	0	--
k3_joy_5.ant	0	0	0	--

Layer	Segment	Audio Files	Project	Segment List				
L1	#t'e	#s	#y	#t	#k	#r	#o	#z
L2	#f'e	#ras	#fs'ys	#tko	#ro	#z'u		
L3	#f'e.ras	#fs'ys.tko						
L4	#f'e	#ras	#fs'ys	#tko	#ro	#z'u		
L5								TG 0

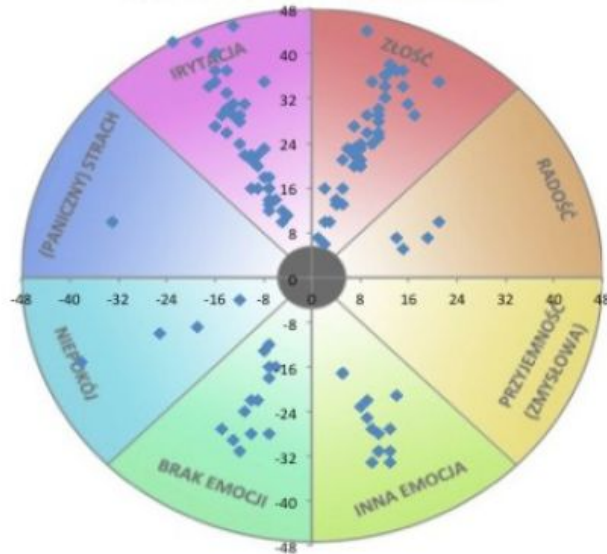


Approaching emotions & affective states

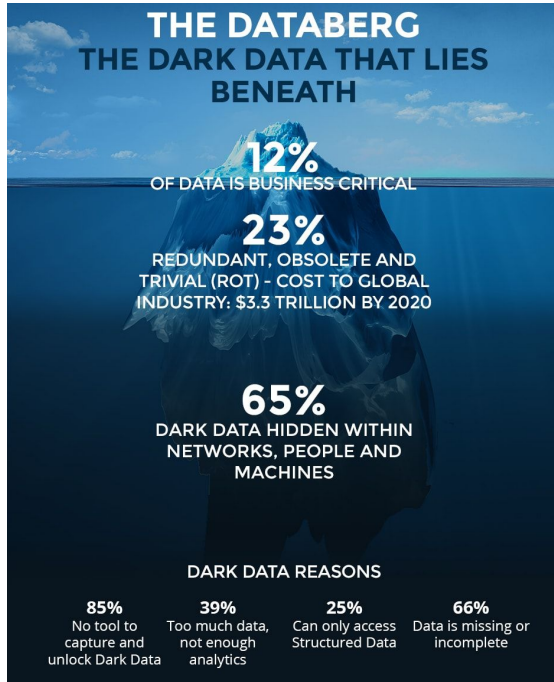
INTENDED JOY



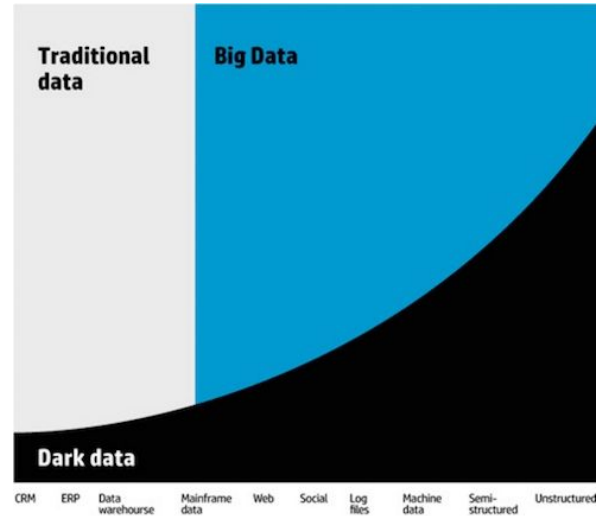
INTENDED ANGER



Data mining: what can we find there?



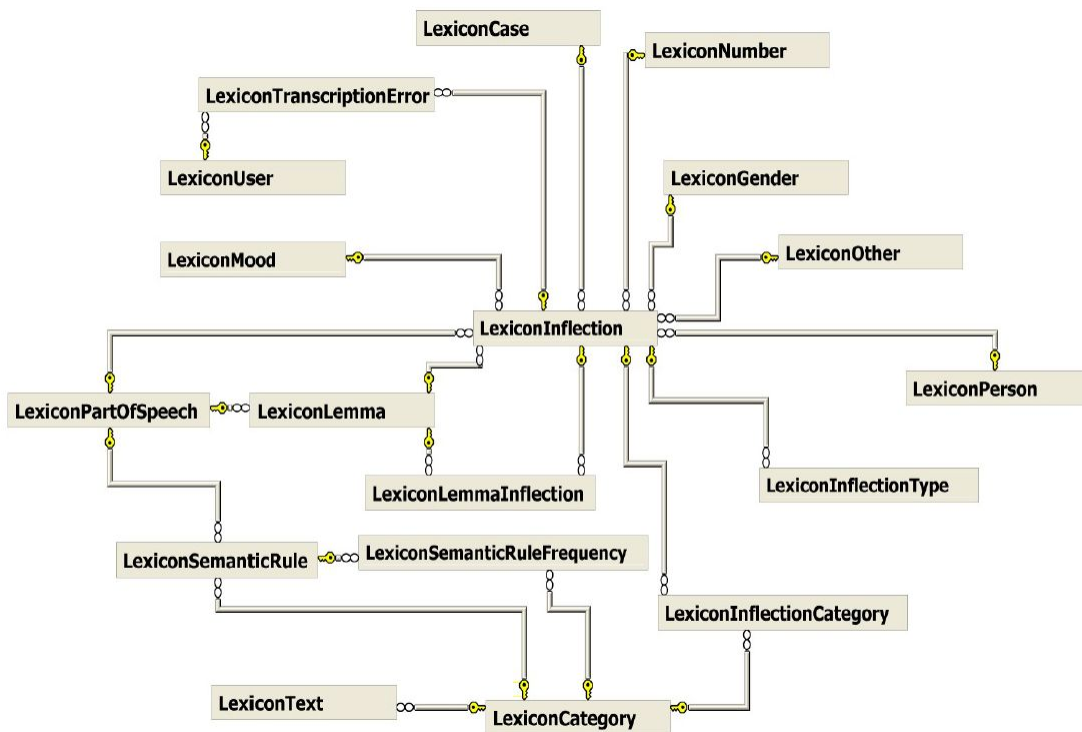
Picture: <https://datumize.com/evolution-dark-data/>



Picture:

<http://www.kdnuggets.com/2015/11/importance-dark-data-big-data-world.html>

Data mining: what can we find there?



- Discover relationships, dependencies, patterns, rules
- In order to do that we often need to combine multiple types of information within one workspace

Data formats & tools

interoperability

- International standards for metadata formats: IMDI, Dublin Core, OLAC
- Best practices for file formats, e.g., XML-based
- It is always worthwhile to check to import/export option of tools we plan to utilize: the more the better

Summary

- **Data & metadata** collection, management & analysis are subsequent steps in the process of dealing with resources but only to a certain extent.
- The three steps are mutually related, and sometimes they **overlap** and need to be re-defined depending on a particular application.
- According to the contemporary best practices for digital humanities, data collections need to be **re-usable**, and should be stored with the use of data formats enabling **interoperability**.



MA Studies @ AMU Poznań: www.eldo.amu.edu.pl

Rekrutacja 2017

ELLD_o

Empirical Linguistics and
Language Documentation

MA programme
in English

LANGUAGES IN DANGER

EN | PL | DE | HU | NL

home interactive map book of knowledge
teaching materials what can you do about

FEATURED LANGUAGE

miyako

Miyako belongs to the Ryukyuan group of languages which together with their "big sister" Japanese make up the Japonic language...

[more](#)

INTERACTIVE MAP



Thank you!



Katarzyna Klessa

E-mail: klessa@amu.edu.pl

More about me:

www.katarzyna.klessa.pl
