

Corpus Linguistics, class 5

Katarzyna Klessa

Empirical Linguistics & Language Documentation

Corpus Annotation: Speech Corpora

Revision: What is annotation?



Revision: What is annotation?

The process / result of attaching labels (tags, etc.) to data in written and spoken language corpora (as well as in multimodal ones).

Themes
Abortion is wrong
- destroy potential (another Jesus?) (a gift from God)
- they were rich - had no reason apart from it was inconvenient

Narrator - the unborn child

TITLE
Lullaby 9.6 "For unto us a child is born,"

UNTOLD IS...
unfinished question
Made love, or had sex?
Repetition for emphasis

STRUCTURE - continuous monologue of 35 lines

Somehow at sometime
They committed themselves to me
And so, it was
Small, but firm
Tiny in shape
Lasting to live
I hung in my pulsing cave
Soon they knew of me
My mother - my father
I had no say, wrong being
I tried to run
And love
The' I couldn't think
Each part of me was saying
A silent, "Wait for me
I will bring you love!"
in an instant
Blind, not at all conscious
By the hand of one
Whose good name
Was graven on a brass plate
In Wimpoole Street.
And dropped on the sterile floor
On a floor separated plastic waste baskets
There was no Queens Counsel
To take my brief.
The cot I might have warmed
Stood in Harrods shop window.
When my passing was told
My father would
No grief filled my empty space
My death was celebrated
With two tickets to see Danny La Rue
Whom I pretended to love a woman
I like my mother was

unified question
Made love, or had sex?
Repetition for emphasis

Day of being conceived
Safety Security (Metaphor)
Feels love already

unable to fight back
Bitter at being aborted (betrayed)

no rights to live, had to hope parents wanted him

Turning Point

Good reputation but does evil things

no legal redress. A 'thing' not a 'human'

She had no more right to call herself a woman than a dog queen.

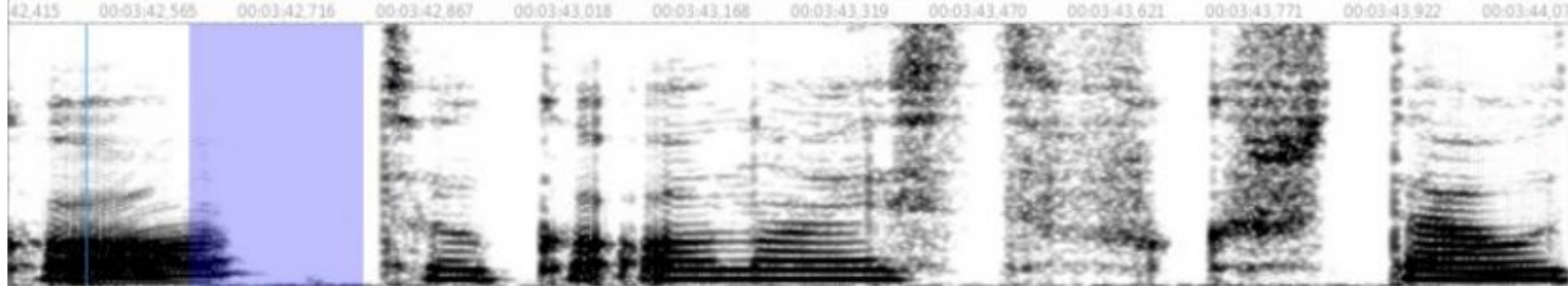
A feeling, an earnest desire, a want

It removed an inconvenience

PURPOSE - to make reader sympathetic with narrator and therefore disagree

ORGANIZATION - 35 line continuous monologue

trullenglish.weebly.com



Annotation Add New Layer Insert segment Delete Import / Export

Annotation	00:03:42.565	00:03:42.716	00:03:42.867	00:03:43.018	00:03:43.168	00:03:43.319	00:03:43.470	00:03:43.621	00:03:43.771	00:03:43.922	00:03:44.071									
Orthog			tylko ono jest																	
Syllables	ko		t'yl	ko	'o	no	j'est			fkSt'aw										
Phonemes	o		t	'y	l	k	o	'o	n	o	j	'e	s	t	f	k	S	t	'a	w
TGA			TG 33																	
			00363416583992013								377368937833572									

Why annotate speech corpora at all?



In case of some **research questions** speech corpus annotations may be the best source of information.

Annotated speech corpora are useful for development tasks, e.g.,
in **speech technology applications**.

Reading: [Phonetic Analysis of Speech Corpora](#) (Chapter 1)

What can we find in speech corpora annotations?

Time-aligned labels including:

- Orthographic or - preferably - phonetic transcription (various alphabets can be used: IPA, SAMPA, other);
- Linguistic tags, e.g.: part-of-speech (POS) or other grammatical, lexical, semantic information;
- Pragmatic information, tags related to communication, information exchange, conversation management, discourse turns...
- Paralinguistic tags: emotions, affect, attitude, voice quality...
- Non-linguistic tags: health condition, speech disorders, environment...
-

Speech annotation procedures

- Annotation tasks can be performed **manually** or using automatized procedures;
- Automatized procedures are available only for some tasks and often need to be manually verified;
- This makes the process of speech annotation a time-consuming, difficult and costly task.
- Nowadays, the need for large speech corpora is increasing. Thus, **automatization** of the procedures is extremely welcome and desired by both academia and enterprise (speech technology, large database search engines etc.).

Speech annotation tools

Annotation Pro

Praat

Elan

Wavesurfer

My Solution in Annotation

Annotation Pro interface showing audio waveform, annotation layer, and feature space representation.

Graphical representation of the feature space
A flexible, user-defined graphical representation of the feature space makes it possible to use continuous rating scales in annotation specifications. A useful tool for conducting perceptual tests for analyses needing continuous and/or discrete rating scales.

Annotation Pro

E L A N - ELAN Linguistic Annotator
Version: 4.4.0

Copyright © 2001 - 2012
Max-Planck-Institute for Psycholinguistics
Nijmegen, The Netherlands

Max Planck Institute
for Psycholinguistics

Language
Archive

Source code for this version available
under GPL (<http://www.gnu.org>)

PRAAT

doing phonetics by computer

version 5.3.11

www.praat.org

©2006 Kåre Sjølander and Jonas Beskow

Annotation of intonation / prosody

- *Intonation describes how the voice rises and falls in speech. The three main patterns of intonation in English are: falling intonation, rising intonation and fall-rise intonation ([Cambridge Dictionary](#)).*
- A popular specification of annotation useful in the context of describing intonation (and other prosodic features) is for example ToBi - Tone and Break Indices annotation (read about a variant of ToBi [HERE](#)).
- The general ideas of the specification may be true for many languages, however, some language-dependent issues occur and thus specifications are adjusted to particular languages or language families.

Intonation labelling in PoInt corpus

First the utterances needed to be transcribed orthographically. Then, the orthographic script was converted to Polish [SAMPA](#) (extended).

Next, the following steps were performed to annotated prosodic features:

Step 1: Dividing a portion of signal into intonational phrases (IPs)

Step 2: Finding nuclear syllables

Step 3: Labeling tones

Reading & **more details**: [HERE](#)

Intonation labelling in PoInt corpus

Step 2: Finding nuclear syllables



During the labeling procedure, two basic situations were distinguished:

- The nuclear syllable occurs in its standard position, i.e. on the **penultimate** syllable of the last word in the IP (pre-penultimate in some words of foreign origin and ultimate in stressed monosyllables).
- The nuclear syllable is intentionally **shifted** towards the beginning of the IP in order to mark
- the focus, express an emotional attitude or for some other reasons.

Intonation labelling in PoInt corpus

Step 2: Finding nuclear syllables



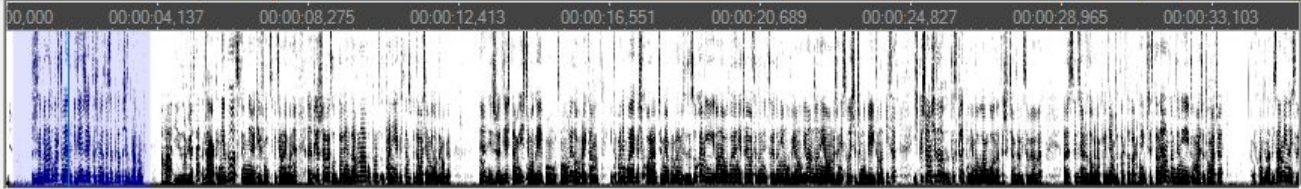

During the labeling procedure, two basic situations were distinguished:

- The nuclear syllable occurs in its standard position, i.e. on the **penultimate** syllable of the last word in the IP (pre-penultimate in some words of foreign origin and ultimate in stressed monosyllables).
- The nuclear syllable is intentionally **shifted** towards the beginning of the IP in order to mark
- the focus, express an emotional attitude or for some other reasons.

File Edit View Analysis Plots Tools Plugins Help

Layers + - Audio Play Out In Selection Full Waveform Spectrogram FFT: 256 | Hann HQ

Orthography +
Pitch (F0) +
IPs +
Nuclear Syllables +
Tones +

Annotation Add New Layer Insert segment Delete Import / Export HQ

00:000 00:00:04.137 00:00:08.275 00:00:12.413 00:00:16.551 00:00:20.689 00:00:24.827 00:00:28.965 00:00:33.103

Orthogr e beę po
rostru miała
ci dzień że j

Pitch (F0)

IPs

Nuclear

Tones

Exploring speech corpus annotations

Speech annotation files are in fact various kinds of text files:

- often structured according to XML-based formats
 - Elan .EAF files
 - Annotation Pro .ANT / .ANTx files)
- or other:
 - Praat TextGrid files
 - various outputs from Wavesurfer, e.g., based on .CSV table formats

These files can be processed / explored also using text analysis tools.

tbc.