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).


Annotation
Add New Layer
Insert segment
Delete
Import / Export
Ortogr
tylko ono jest
Syllables
ko
t"yl
ko
"o
no
j"est
fkSt"aw
Phones
o
t
"y
l
k
o
"o
n
o
j
"e
s
t
f
k
S
t
"a
w
TGA
TG 33
Audio: 16 bit PCM: 16kHz 1 channels
Annotation: 16 kHz
Rate N: 1.08 seg/s
Rate G: 0.19 seg/s
Annotation Pro © Katarzyna Klessa

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…
- ….

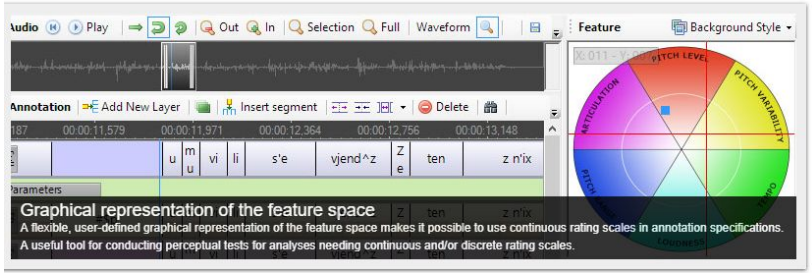
Speech annotation procedures

- 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.).

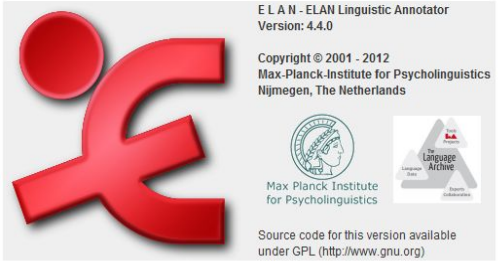
Speech annotation tools

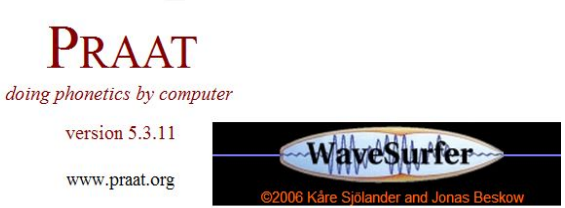
[Annotation Pro]
[Praat]
[Elan]
[Wavesurfer]

My Solution in Annotation



Annotation Pro





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:

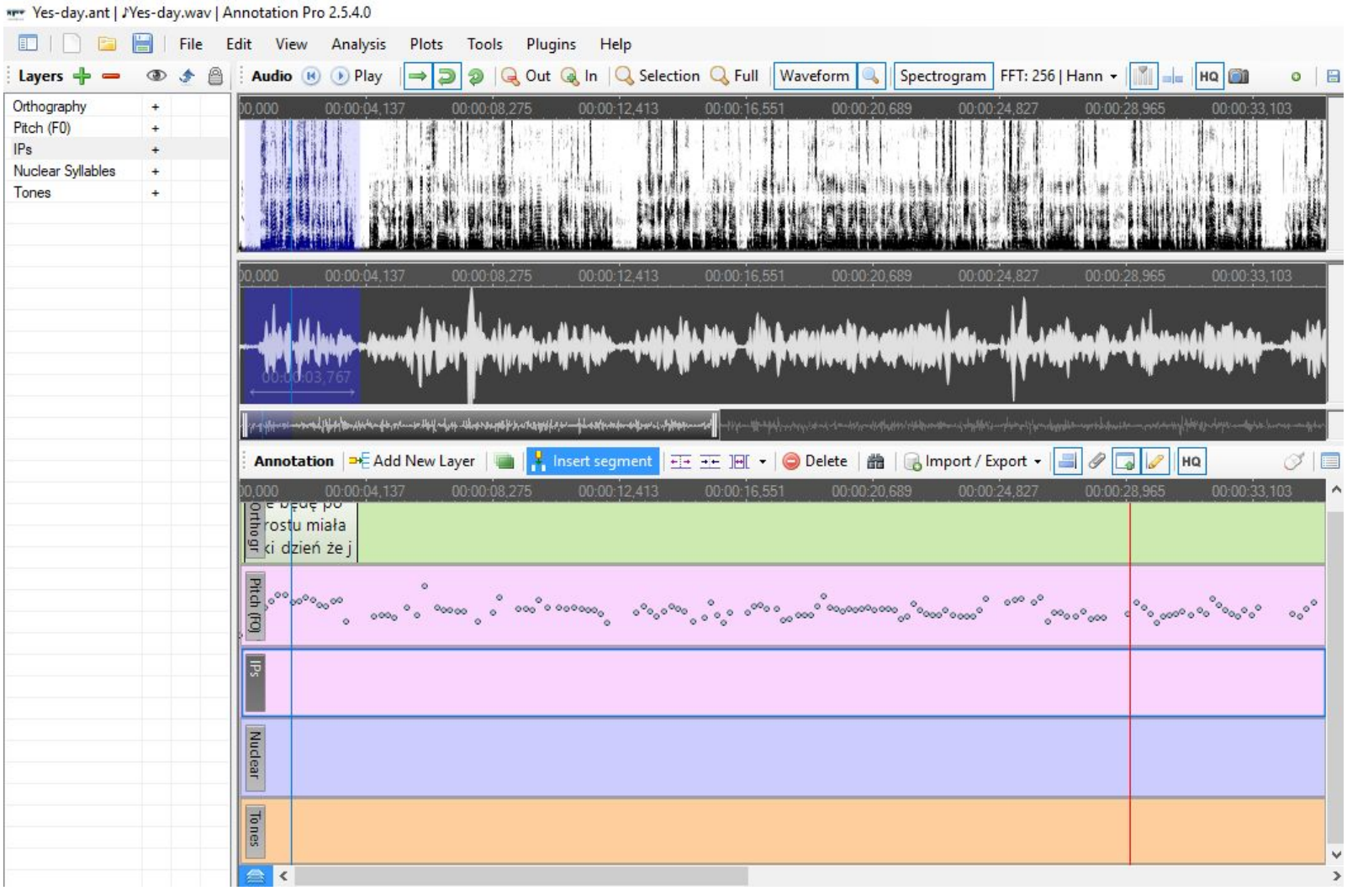
- The nuclear syllable occurs in its standard position, i.e. on the **penultimate** syllable of the last word in the IP (pre-penultimate in some words of foreign origin and ultimate in stressed monosyllables).
- The nuclear syllable is intentionally **shifted** towards the beginning of the IP in order 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.


Yes-day.ant | ♪Yes-day.wav | Annotation Pro 2.5.4.0
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
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
00:00:03,767
Annotation
Add New Layer
Insert segment
Delete
Import / Export
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.