Towards Spoken Latvian Corpus: Current Situation, Methodology and Development

Ilze Auziņa

Institute of Mathematics and Computer Science, University of Latvia

Background (1)

- The development of the Latvian National Corpus was initiated by the State Language Commission in 2004
- The Latvian Language Corpus Conception, 2005
- During last six years several text corpora have been developed at IMCS, UL
- Financial support:
  - the State Language Agency
  - the Latvian Council of Science

Background (2)

www.korpuss.lv

Name	Characterization	Running words	Morpho-logical annotation
miljons-2.0	The *Balanced Corpus of Modern Latvian*, 2009, created in IMCS	3.5 millions	no
miljons–2.0m	The *Balanced Corpus of Modern Latvian*, 2009, created in IMCS	3.5 millions	yes
Saeima-2.0	*Corpus of the Transcripts of the Saeima's (Parliament of Latvia) sittings*	32.5 millions	no
timeklis-1.0	The *Web corpus*, created in IMCS	100 millions	partial

- **only written texts**
- **transcripts of the Saeima's sittings ≠ transcribed speech**

Concept of Balanced Spoken Latvian Corpus (1)

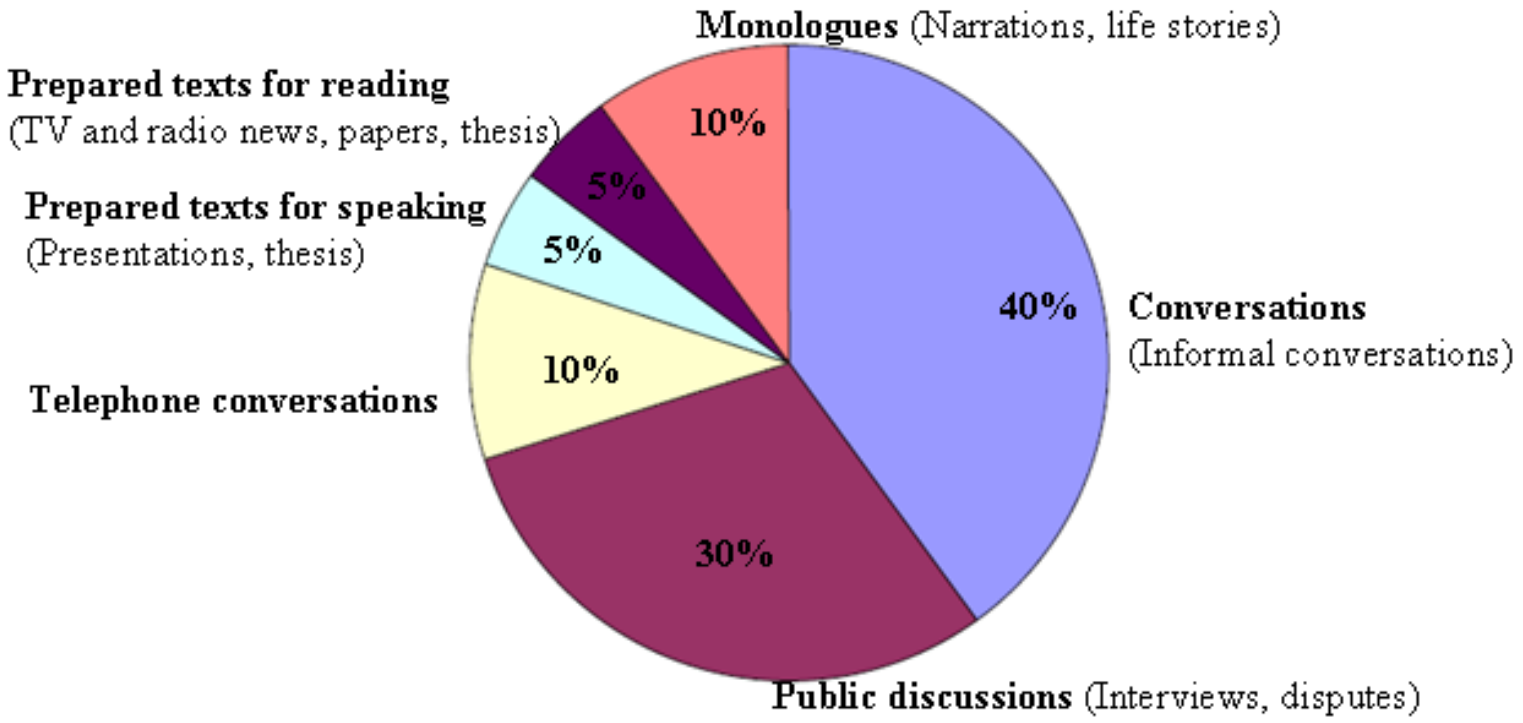

Figure 1. Prospective ratio of speech data (Latvian language corpus conception, 2005)

Concept of Balanced Spoken Latvian Corpus (2)

Planned division:

- Spontaneous speech (~80%) → dialogues and polilogues (phone calls; public discussions, interviews; private conversations etc); monologues (narrations, life stories)

- Planned speech (~20%) → monologues (TV and radio news; academic speeches, papers).

Concept of Balanced Spoken Latvian Corpus (3)

Levels of annotation:

- metadata
- orthographic annotation
- morphosyntactical annotation
- phonetic annotation
- prosodic annotation

  A small part of the corpus should be annotated phonetically and prosodically

What do we already have?

- **Collected speech data**
  - **Institutions**: (IMCS UL; Institute of Philosophy and Sociology (UL); Rezekne Higher Education Institution etc.)
- Common metadata and annotation standards are developed and used.
- **Some corpora are being developed, for example:**
  - *The Corpus of Public Discussion* (being developed; IMCS, UL)
  - *The Latvian Learner Corpus (developed at Latvian Associations of Language Teacher)*
  - *The Colloquial Speech Corpus* (being developed; Language Embassies & IMCS UL)

The Corpus of Public Discussions (1)

- Recordings of a radio discussion program called “Puškins pret Dantesu”, radio SWH, 2006.
- The corpus contains 11 recordings, average length of each record is 40 minutes (total record length is ~ 8 hours).
- Number of speakers is 21 (3 females and 18 male)
- The orthographic transcription and the annotation of non-linguistic acoustic events were chosen.
- The metadata are added.

The Corpus of Public Discussions (2)

Metadata

- **specification of speakers**: the information of speakers age, sex, education, accent etc.;
- **specification of recording**: the recording software, the specification of recording equipment, and acoustic environment;
- **specification of data**: the format and index of the data;
- **specification of annotation**

The Corpus of Public Discussions (3)

Corpus annotation:

- Orthographic transcription
- Morphosyntactic annotation: POS and chunking
- Phonetic annotation
- Prosodic annotation

The Corpus of Public Discussions (4)

The principal features of the **orthographic transcription** scheme are:

- Generally **orthographical standards** for the Latvian language are used; incorrect forms are annotated.
- **Capitalization**: initial words of sentences are capitalized only if they would be capitalized in the middle of the sentence.
- **Numbers** are spelled out following the standards of the Latvian language, using correct ending.
- The transcription includes only some **punctuation marks**: full stop, comma, question mark and exclamation mark.

The Corpus of Public Discussions (5)

During the process of transcription some **problems** already arose, for example:

- Non-standard spelling and pronunciation:
  - *lasam* (incorrect spelling), *lasām* (correct spelling) present 1st pl. *read*)
- In continues speech often it is not easy to decide where one utterance ends and other starts due to fast speech, mispronunciation, overlapping etc.

The Corpus of Public Discussions (6)

Annotation of non-linguistic events:

- Main non-linguistic acoustic events marked in orthographical are pause fillers, hesitations.
- Human noises, such us laughing, cough, expiration, inspiration etc.
- Mispronunciations, unintelligible words, unfinished words.
- Pauses: both micro pauses and pauses (silences longer than 1 sec.) are marked with full stop enclosed in brackets.

  etc.

The Corpus of Public Discussions (7)

Next stages:

- Morphosyntactic annotation: POS and chunking
  - the text morphosyntactic annotating tool will be adjusted and used to speech data processing
- Phonetic annotation
  - A part of the data in the corpus will enriched with authomatically obtained and a manually verified broad phonetic transcription
- Prosodic annotation

Discussion and conclusions (1)

- The development of a speech corpus is much more time consuming and much more expensive than development of a text corpus.
- This is because speech data has to be transcribed at first and only then it can be structurally and morphosyntactically anotated, by adding relevant meta information to speech data.
- Currently only some special speech corpora are being created
  - *The Colloquial Speech Corpus (planned size - 1 million running words)*
  - *The Corpus of Public Discussions*

Thank you for your attention!

ilze.auzina@lumii.lv