Non-professional and specialised corpora

Gordana Hržica
Laboratory for Psycholinguistic Research

University of Zagreb
Faculty of Education and Rehabilitation Research
Department of Speech and Language Pathology
University of Zagreb
Faculty of Education and Rehabilitation Research
Department of Speech and Language Pathology

Polin
Laboratory for Psycholinguistic Research

Language acquisition
Language knowledge
Language processing
Laboratory for Psycholinguistic Research

University of Zagreb
Faculty of Education and Rehabilitation Research
Department of Speech and Language Pathology

Typical language development/status

Atypical language development/status
Croatian Corpus of Child Language

- Background
- The corpus
- Methodological issues
- Research

The research leading to these corpus has received funding from the Ministry of Science, Education and Sports of the Republic of Croatia for the projects Higher Cortical Functions and Language: Developmental and Acquired Disorders (MZOS 0130131484-1488) and Language Acquisition in Cross-Linguistic Context: Psycho- and Neurolinguistic Approach (MZOS 0013002).
Language acquisition

- **Background**
- The corpus
- Methodological issues
- Research

CLAN

(Computerized Language Analysis)

CHAT

(Codes for Human Analyses of Transcripts)

CHILDES

(Child Language Data Exchange System – MacWhinney 1987, 2000)
- Data sharing
- Community building
- Development of technologies
- Unification of the transcription

Catherine Snow & Brian MacWhinney (1983)
Croatian Corpus of Child Language
(Kovačević, 2002)
- Three children
- Longitudinal corpus
- Home environment
- Child directed speech: around 210 000 tokens
- Child language: around 110 000 tokens
- Linked with audio files
- Morphologically marked (partially)
Language acquisition

- Background
- The corpus
- Methodological issues
- Research

- Amount of data
- SES of participants
- Sampling method
- Frequency effect

Croatian Corpus of Child Language
Language acquisition

- Background
- The corpus
- Methodological issues
- Research

→ Dense sampling
  Max Planck corpus
  (e.g. Behrenes 2006.)
  MIT Human Speechome Project
→ Cross-sectional sampling
→ Elicitation methods
→ Development of advanced measures
  lexical richness
  syntactic complexity
→ Cross-linguistic research
- basic data of language acquisition in different languages – examples:

- Background
- The corpus
- Methodological issues
- Research

- basic data of language acquisition in different languages – examples (Croatian):

- Acquisition of nouns
- Acquisition of adjectives
- Acquisition of verbs
- The role of diminutives in child language
- ...
The research leading to these corpus has received funding from the Croatian Science Foundation, grant HRZZ-2421 for the project "Adult language processing". The Croatian Adult Spoken Language Corpus (HrAL) - Background - Methodological Issues - The Corpus - Research
Croatian Adult Spoken Language Corpus (HrAL)

- **Background**
- The corpus
- Methodological issues
- Research

**CLAN**
(Computerized Language Analysis)

**CHAT**
(Codes for Human Analyses of Transcripts)

**TalkBank**
(MacWhinney 2007)
- Data sharing
- Community building
- Development of technologies
- Unification of the transcription

Catherine Snow and Brian MacWhinney (1983)
Croatian Adult Spoken Language Corpus (HrAL) (Kuvač Kraljević & Hržica 2016)

- Informal conversation
- More than 600 speakers from all Croatian counties
- More than 160 transcripts
- Around 280 000 tokens
- Information about participants’ age, gender, education and origin
- Linked with audio files
- Available in the TalkBank
- Adult spoken language
  - Background
  - The corpus
  - Methodological issues
  - Research

- Representativeness of the corpus
  - Diversity of speech
  - Language varieties
  - Sample of participants
## Adult Spoken Language

<table>
<thead>
<tr>
<th>Spoken</th>
<th>Written</th>
<th>Aphasia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult Spoken Language Corpus (HrAL)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Background
- The corpus
- **Methodological issues**
- Research
Croatian Adult Spoken Language Corpus (HrAL)

- Background
- The corpus
- Methodological issues
- Research

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Spoken</th>
<th>Written</th>
<th>Aphasia</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-20</td>
<td>14</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>21-30</td>
<td>27</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>31-40</td>
<td>15</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>41-50</td>
<td>25</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>51-60</td>
<td>29</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>61-70</td>
<td>15</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>71-80</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>81-90</td>
<td>2</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Adult Spoken Language

- Background
- The corpus
- Methodological issues
- Research

- Language technologies
- Description of adult language
- Adult language as a new baseline for studying other language groups
The research leading to these corpus has received funding from the European Regional Developmental Fund 2007-2013 under grant agreement No. RC.2.2.08-0050 (project RAPUT – Računalni asistent za podršku pri unosu teksta osobama s jezičnim poremećajima)
Croatian Corpus of Non-professional Written Language

- **Background**
- **The corpus**
- Methodological issues
- Research

- Almost 390 speakers (Age: 10 to 80)
- Around 500,000 tokens
- More than 50% of participants are persons with language disorders
- Elicited production
- Morphologically coded (Version 4 of the MULTEXT-East Morphosyntactic Specifications for Croatian (Ljubešić, 2013))
- Still not publicly available (in preparation)
Non-professional written language

- Background
- The corpus
- **Methodological issues**
- Research

- Elicitation materials
  - Age-adjusted
  - Differ in structure and type: essay, answers to questions, narratives, letter, invitation
  - Large differences in text size (within the same age groups)
Non-professional written language

- Background
- The corpus
- Methodological issues
- Research

- Non-professional writing vs. professional writing
- Writing of persons with language disorders


The research leading to these corpus has received funding from the Croatian Science Foundation, grant HRZZ-2421 for the project “Adult language processing”.

- **Background**
- The corpus
- Methodological issues
- Research

Croatian Corpus of Aphasic Speech
- **Aphasic Speech**
  - Background
  - The corpus
  - Methodological issues
  - Research

- AphasiaBank (part of the TalkBank)
  - Interviews between aphasic participants and clinicians
  - Consistent protocol for interviews (personal narratives, picture description, story telling, procedural discourse)
  - Data-sharing
  - Control groups
  - Using familiar transcription protocols

Croatian Corpus of Aphasic Speech
Aphasic Speech

- Background
- The corpus
- Methodological issues
- Research

- Croatian Corpus of Aphasic Speech
  - 15 persons with aphasia
  - Anamnestic data
  - Control group (in preparation)
- Background
- The corpus
- Methodological issues
- Research

- Diverse target group:
  - Type of aphasia
  - Time of the stroke
  - Individual differences
  - Effects of the therapy
Aphasic Speech

- Background
- The corpus
- Methodological issues
- Research

- More than 300 studies in AphasiaBank bibliography
To be used in:

Research
- comparing groups of speakers
- Comparing genres
- Collocation studies
- Language varieties
- Language measures
- ...

Language assessment
- Frequency
- Syntax
- Data collected on specific tasks
- Language measures
- ...

...
Future work:

• Prepare the data according to current standards in corpus linguistics
• Make all corpora publically available and searchable
Thank you!

Questions?