



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

Non-professional and specialised corpora

Gordana Hržica

JOTA
February 28, 2017

|
Laboratory for
Psycholinguistic Research



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

Polin

Child language

Spoken

Written

Aphasia

|
Laboratory for
Psycholinguistic Research



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



|
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




**Language
acquisition**

- **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 (MZOŠ 0130131484-1488) and Language Acquisition in Cross-Linguistic Context: Psycho- and Neurolinguistic Approach (MZOŠ 0013002).

|
Croatian Corpus of Child
Language



**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

|
Croatian Corpus of Child
Language



**Language
acquisition**


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

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)

|
Croatian Corpus of Child
Language




**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

|
Croatian Corpus of Child
Language




Language acquisition

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

- basic data of language acquisition in different languages – examples:
 - Brown, R. (1973). A first language: The early years. Cambridge, MA: Harvard University Press.
 - Xanthos, A. et. al. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language* 31 (4), 461-479
 - Gagarina, N. et al. (2012). Multilingual Assessment Instrument for Narratives

|
Croatian Corpus of Child
Language



Language acquisition

- 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
 - ...

|
Croatian Adult Spoken
Language Corpus (HrAL)



**Adult
spoken
language**

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

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

|
Croatian Adult Spoken
Language Corpus (HrAL)



**Adult
spoken
language**

- **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

|
Croatian Adult Spoken
Language Corpus (HrAL)



**Adult
spoken
language**

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

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

|
Croatian Adult Spoken
Language Corpus (HrAL)



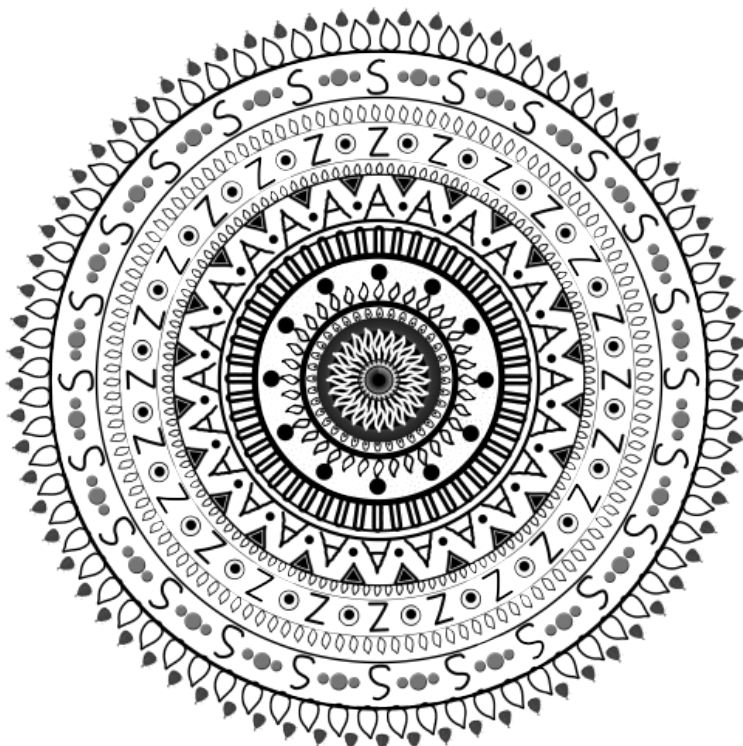
**Adult
spoken
language**

- Background
 - The corpus
 - **Methodological issues**
 - Research
-
- Representativeness of the corpus
 - Diversity of speech
 - Language varieties
 - Sample of participants

↓
Croatian Adult Spoken
Language Corpus (HrAL)

Adult
spoken
language

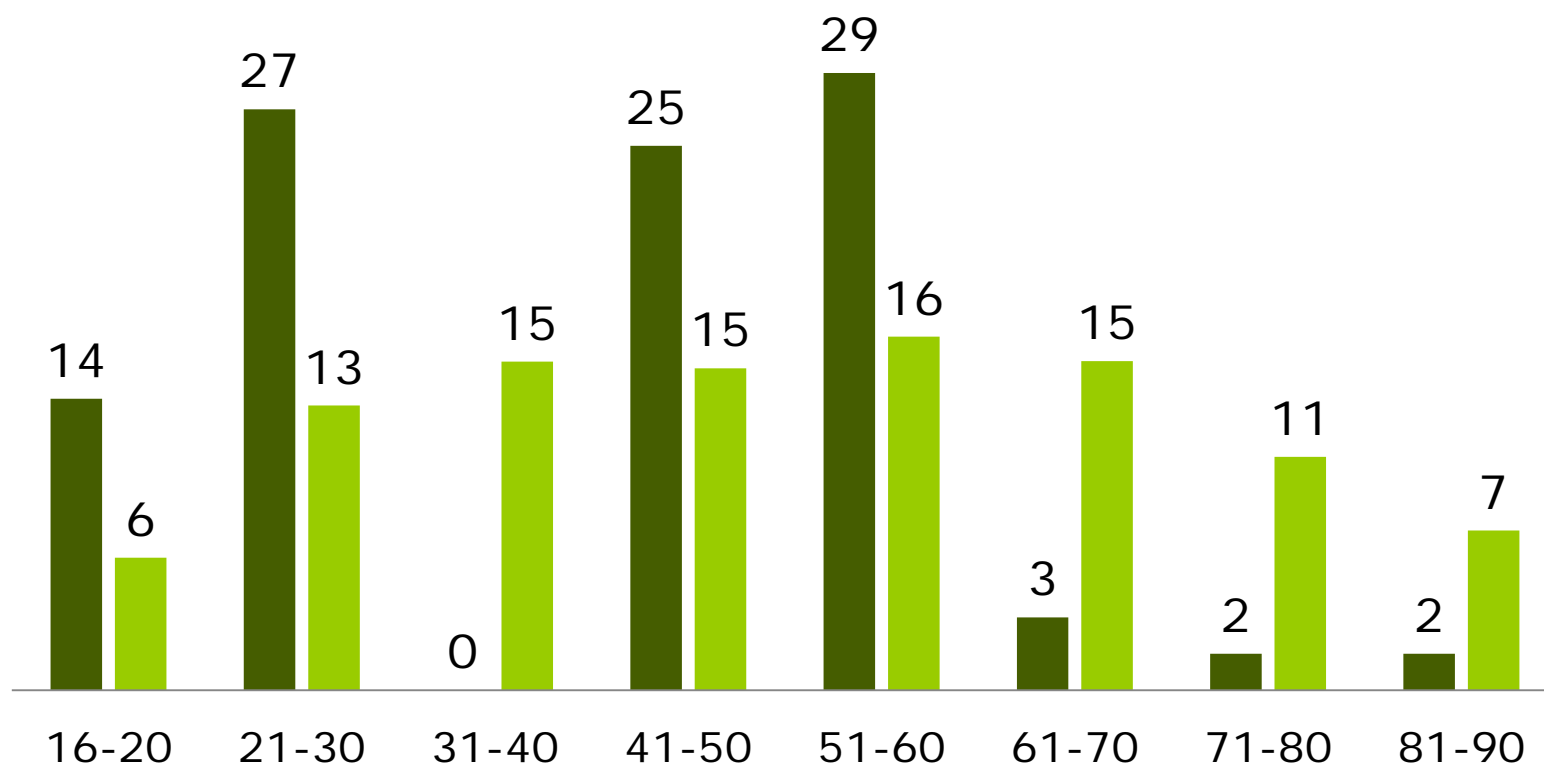
- Background
- The corpus
- **Methodological issues**
- Research



↓
Croatian Adult Spoken
Language Corpus (HrAL)

Adult spoken language

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




|
Croatian Adult Spoken
Language Corpus (HrAL)



**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

|
Croatian Corpus of Non-
professional Written
Language



**Non-
professional
written
language**


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



R A P U T

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




**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)


|
Croatian Corpus of Non-
professional Written
Language



**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)

|
Croatian Corpus of Non-
professional Written
Language




**Non-
professional
written
language**

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

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

Štefanec, V., Ljubešić, N., Kuvač Kraljević, J. (2016). Croatian Error-Annotated Corpus of Non-Professional Written Language. LREC 2016.


Kuvac Kraljević, J., Matić, A., Kologranić Belić, L., and Olujić, M. (under review). Written narratives of adolescents with specific language impairment: discourse analysis.



Aphasic Speech


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

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



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



Aphasic Speech

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


- *Croatian Corpus of Aphasic Speech*
 - 15 persons with aphasia
 - Anamnestic data
 - Control group (in preparation)



Aphasic Speech

- 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**
-
- MacWhinney et al. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology*. 2011; 25(11): 1286–1307.
 - 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?