University of Zagreb
Faculty of Education and
Rehabilitation Sciences

Croatian written and spoken corpora of speech with communication disorders

Gordana Hržica

University of Zagreb

Department of Speech and Language Pathology

CLARIN-DELAD Workshop, 27-28 January 2021

OVERVIEW

CroDA: a Croatian discourse corpus of speakers with aphasia

- Elicited speech
- Persons with aphasia
- 20 speakers
- 4 tasks per speaker
- Collected and transcribed according to rules of AphasiaBank (part of TalkBank) (MacWhinney at al., 2011)

Croatian Corpus of Non-Professional Written Language

- Elicited written language
- Persons with typical language status
- Persons with language disorders
- More that 400 speakers
- 8 taks per speaker

CroDA: a Croatian discourse corpus of speakers with aphasia

INTRODICTION

- Aphasia - acquired neurological disorder caused by stroke in 80% of cases, and in other cases by tumors, injuries and infections

- Corpus-based studies of such speakers are relatively recent

- Specialised corpora of speakers with aphasia in some languages, including Dutch (CoDAS; Westerhout, Monachesi, 2006), Greek (GREECAD, Varlokosta et al., 2016), and Russian (CliPS - Clinical Pear Stories; Khudyakova et al., 2016)
- → Annotated using ELAN software or TalkBank system (Codes for Human Analysis
- of Transcripts (CHAT) and annotation and analysis program Computerised Language Analysis (CLAN))

- AphasiaBank (http://aphasia.talkbank.org/) - a subsection of TalkBank that contains multimedia interactions for communication studies in aphasia

CroDA: a Croatian discourse corpus of speakers with aphasia

PROCEDURE

- AphasiaBank is based on structured interviews of speakers with aphasia conducted by clinicians (MacWhinney et al., 2011).

→ To ensure consistent experimental conditions and comparability across participants and languages (including the control group).
→ To ensure consistent experimental conditions and comparability across participants and languages

- 1) Personal narratives
- 2) Picture descriptions and short narratives
- 3) Storytelling
- 4) Prcedural discourse

CroDA: a Croatian discourse corpus of speakers with aphasia

PARTICIPANTS

Persons with aphasia

- 20 patients with aphasia
- All monolingual and right-handed prior to the event that caused their aphasia.
- Patients were additionally classified as fluent (N=10) or nonfluent (N=10)
- (Goodglass and Kaplan, 1983).

Control group:

- 20 speakers with typical language status
- All monolinguals and right-handed
- Age-matched and gender matched individually with the clinical group

CroDA: a Croatian discourse corpus of speakers with aphasia

PARTICIPANTS

- All participants signed an informed consent form
- Information for consent was delivered in writing and orally, with the assistance when needed
- Data were anonymised
- Coropora in the AphasiaBank are pasword-protected – users need to register

CroDA: a Croatian discourse corpus of speakers with aphasia

ANNOTATION

- CroDA is not morpholgically annotated (in progress)
- Transcription and coding acording to TalkBank system:
  - Codes for Human Analysis of Transcripts (CHAT)
  - Computerised Language Analysis (CLAN)

- Annotation of errors according to three groups:
  - Phonological errors
  - Morphsyntactic errors
  - Semantic errors

CroDA: a Croatian discourse corpus of speakers with aphasia

DATA SHARING AND REACH

CroDA was published in 2017

Access:
https://aphasia.talkbank.org/access/Croatian/Aphasia/Zagreb.html

Reference:
Kuvac Kraljevic, J., Hrzica, G., Lice, K. (2017). CroDA: a Croatian discourse corpus of speakers with aphasia. Croatian Review of Rehabilitation Research, 53, 2.
https://hrcak.srce.hr/191747

Projects:
CroDA was developed within the framework of the project Adult Language Processing (HRZZ-2421-UIP-11-2013)

4 citations (excluding self-citations)


Croatian Corpus of Non-Professional Written Language

INTRODICTION

- The aim was to design the first specialized written corpus in Croatian that consists of diverse texts produced by ordinary, non-professional typical speakers and speakers with different types of language disorders.

- The novelty of this corpus is that it provides insights into the writing skills on productive level of Croatians who have “typical” education and level of exposure to writing.

- Despite the increasing availability of clinical corpora, the literature on writing skills of people with language disorders is not so extensive (Zourou et al, 2010). Detailed insights are lacking for most languages, including Croatian.


Croatian Corpus of Non-Professional Written Language

PROCEDURE

- A protocol that defined discourse elicitation tasks of the participants and the methods to be used in analyzing the data
- Similar to protocols for other corpora
- Four groups of tasks representing different writing styles:
  - Descriptive (2 tasks)
  - Expository (2 tasks)
  - Narrative (2 tasks)
  - Letter (2 tasks)
- different levels of formality


Croatian Corpus of Non-Professional Written Language

PARTICIPANTS

Participants

- A broad age range: from the period when writing should become automatized until old age
- Geraphic diversity of participants: coming from different Croatian counties
- Persons with language disorder are a specific group of non-professional speakers and writers:
  - Developmental language disorder
  - Dyslexia
  - Aphasia

PARTICIPANTS

Male	213
Female	188
Age group	
Children (10-15 years old)	170
Adolescents (16-21 years old)	36
Adults (at least 22 years old)	195
Language status	N
Typical language status	134
Language disorders	
Developmental language disorder	61
Dyslexia	110
Dysgraphia	5
Aphasia	91


Croatian Corpus of Non-Professional Written Language

ETHICS

- All participants signed an informed consent form according to the Helsinki Ethical Principles for Medical Research
- Information for consent was delivered in writing and orally, with the assistance of a speech-language pathologist and family members when needed
- All relevant approvals of Ethics Committeees or similar bodies were obtained. All names of the participants were changes
- All names and other details that might point to the identity of the speaker were either changed or removed form the transcripts
- Corpus is still not publically available. Prior to publication, additional precoutions might be taken to preserve the identity of participants


Croatian Corpus of Non-Professional Written Language

ANNOTATIONS

- 401 participants produced more than half a million tokens in more than 41 000 utterances
- The corpus was morphologically annotated (followed version 4 of the MULTEXT-East Morphosyntactic Specifications for Croatian (Ljubešić 2013))
- Annotation was done manually and automatically using additional layers to take into account that the writers were not professionals, the text was not proof-red and many of them had language disorders.
    - The first layer: correction of tokens, and tokens could be divided into two or merged if word boundaries were displaced.
    - The second layer: errors were marked according to one of 12 types

Croatian Corpus of Non-Professional Written Language

DATA SHARING AND REACH

Still not published (but working on it)

Reference:
Kuvac Kraljevic, J., Hrzica, G., Kologranic Belic, L. (accepted). Croatian Corpus of Non-Professional Written Language. Journal of Speech.

Papers from corpus:
1. Štefanec, V, Ljubešić, N., Kuvač Kraljević, J. (2016) Error-Annotated Corpus of Non-Professional Written Language. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (ed. N. Calzolari), http://www.lrec-conf.org/proceedings/lrec2016/index.html
2. Hrzica, G., Kosutar, S., Kramaric, M. (2019). Lexical Diversity in Written Texts of Persons with Developmental Language Disorder. Croatian review of rehabilitation research, 55, 2. https://doi.org/10.31299/hrri.55.2.2

Projects:
“Computer assistant for text input for persons with language impairment” (RAPUT; RC.2.2.08 - 0050), funded by the European Structural and Investment Funds.

CONSLUSION

- We believe that we have developed two specialised corpora of an adequate size to pose numerous research questions.
- We were careful to include diverse written genres and a broad age range of participants from locations across the country, which even allowed us to capture the three major dialects of the language.
- We are trying to find an optimal path for sharing the data.

INTRODUCTION TO THE WORKSOP:

Corpora of speech of individuals with communication disorders (CSD) ***are hard to obtain****. They are* ***costly to collect*** *and* ***difficult to share*** *due to privacy issues. Moreover, they are often* ***small in size*** *and very* ***specific*** *in terms of communication impairments addressed. These factors make re-use a challenge on the one hand, and a necessity on the other.*