# Croatian Corpus of Non-Professional Written Language – Typical Speakers and Speakers with Language Disorders

**Kuvač Kraljević, Jelena; Hržica, Gordana; Kologranić Belić, Lana**

*Source / Izvornik:* **Govor, 2020, 37, 125 - 147**

**Journal article, Published version**
**Rad u časopisu, Objavljena verzija rada (izdavačev PDF)**

https://doi.org/10.22210/govor.2020.37.07

**Izvorni znanstveni rad**
**Rukopis primljen 6. 11. 2019.**
**Prihvaćen za tisak 3. 3. 2021.**
**https://doi.org/10.22210/govor.2020.37.07**

**Jelena Kuvač Kraljević, Gordana Hržica**
*jkuvac@erf.hr, gordana.hrzica@erf.unizg.hr*
Faculty of Education and Rehabilitation Sciences, University of Zagreb
Croatia

**Lana Kologranić Belić**
*lana.kologranic@gmail.com*
Polyclinic for the Rehabilitation of Listening and Speech SUVAG, Zagreb
Croatia

# Croatian Corpus of Non-Professional Written Language – Typical speakers and speakers with language disorders

## Summary

Corpora, as annotated archives of human communication, are objective, reliable resources for language analysis. Here we present the corpus of non-professional written Croatian, based on 1-year sampling of writings by typical speakers and speakers with language disorders. This corpus provides a unique resource because it samples language used by non-professionals, in contrast to corpora based on texts by professional writers (such as journalists, scholars or novelists) sampled over more than a century. In addition, our corpus contains written language from typical and impaired speakers sampled under identical conditions, allowing detailed analyses of language use. This paper describes the language tasks (essay, story generation, non-formal and formal letter and dictation) used to elicit text production, and procedures for sampling and annotation used to generate the corpus. Its usefulness is illustrated through language productivity analyses of transcripts of different genres produced by writers of different age and language status. This corpus may prove useful for the analysis of writing skills in typical and language-impaired speakers of Croatian.

**Keywords:** Croatian Corpus of Non-Professional Written Language, written language, genres, language disorders

## 1. INTRODUCTION

A corpus is a body of written text or transcribed speech that can serve as an objective, reliable basis for linguistic analysis and description (Kennedy, 2014). The history of text analysis can be traced back to the 13th century, when the Christian Bible was manually indexed, and particularly impressive growth in the development of language corpora has occurred in the past 50 years. During this time, various types of corpora have been developed in different languages. They have been used in the range of areas, such as language teaching and learning, forensic linguistics, translation studies, sociolinguistics, and pragmatics (see McCarthy & O'Keeffe, 2010).

If a corpus is to serve as a source of evidence for linguistic descriptions and analyses of human communicative ability, it should linguistically describe a speaker's language performance (Leech, 1992, p. 107). Linguistic competence and performance are too complex to be described adequately by introspection and elicitation alone (Svartvik, 1992). Therefore, corpus analysis should be seen as complementary to the other methods of language analyses, including experiments. Indeed, a corpus is an empirical basis for testing principles of linguistic theories (Kennedy, 2014).

Corpora can be compiled for many different purposes, and the purpose helps determine corpus size, style and content. *General* or *core* corpora consist of a body of texts that enable linguists to address questions related to vocabulary, grammar or discourse structure. Examples are the British National Corpus (www.natcorp.ox.ac.uk /corpus/index.xml) or Croatian National Corpus (Tadić, 2009). *Specialized* corpora, in contrast, are designed with specific purposes in mind. Croatian examples are the Croatian Child Language Corpus (Kovačević, 2002), which provides information about the specificity of child language development; the Croatian Adult Spoken Language Corpus (HrAL; Kuvač Kraljević & Hržica, 2016), which provides information about spoken grammar and lexicon in adulthood; and the Croatian Discourse Corpus of Speakers with Aphasia (CroDA; Kuvač Kraljević, Hržica, & Lice, 2017), which supports analyses of spoken discourse skills and error production of adult speakers with aphasia. All three corpora are available within TalkBank (https://talkbank.org), a large database of spoken-language corpora covering different languages (MacWhinney, 2002; MacWhinney, Fromm, Forbes, & Holland, 2011; MacWhinney & Wagner, 2010).

Most corpora of written language are based on carefully selected texts produced mostly by professional writers. Corpora of professional writing provide much useful information but cannot be representative of everyday written language use, such as in emails, letters, notes, essays, and business correspondence. Spoken corpora are much more prone to include non-professional speakers, but there is a great discrepancy in size of written and spoken corpora. Raso and Mello (2014) warn that moving towards big data in corpus linguistics does not necessary fill a gap in linguistic resources i.e., does not provide linguists with the means to study spoken language. Similar can be said for non-professional writing. Such resources are rare, an often restricted to small number of words and to limited number of genres. For example, Schler, Koppel, Argamon, and Pennebaker (2006) have collected relatively large corpus of 140 million words, but it is restricted to blogs. Same stands for Enron Email Data Corpus (Federal Energy Regulatory Commission, 2012).

The aim of this paper is to present what appears to be the 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 biggest advantage of this corpus, in comparison to corpora based on the blog, twitter, web and other online language sources which also contain a large amounts of non-professional written texts, is the control of participants and the prescribed procedure of sampling a written language. The novelty of this corpus is that it provides insights into the writing skills on the productive level of Croatians who have at least four years of education, i.e., who have been exposed to the formal learning of writing. We describe in detail the principles guiding the sampling of written texts as to facilitate creation of similar corpora in other languages. We also provide examples of analysis of essays and narratives produced by typical individuals and those with language disorders illustrating some of the questions that can be addressed with this unique type of corpus.

## 1.1. A specialized corpus of non-professional written language

Corpus linguistics has long been biased in favour of professional writing. Typically, large national corpora claiming to be representative of a language sample professional writing from books, newspapers and academic sources, although there are some exceptions (McEnery & Wilson, 2001). The current trend in the building of web-based corpora has allowed increasing inclusion of non-professional texts, but web-based corpora require additional skills to access the non-professional writing therein.

To ensure representativeness, a general corpus may strive to sample a broad range of genres. For example, the spoken part of the Cambridge English Corpus (https://www.cambridge.es/en/about-us/cambridge-english-corpus) contains samples of everyday conversation, radio broadcasts and TV programs, presentations, speeches, meetings and lectures. A specialized corpus, in contrast, may strive to sample a demographically diverse range of speakers or writers, including various ages, socioeconomic statuses and geographic locations (e.g., Carter, 1997, 1998, 1999; Kamandulytė-Merfeldienė, 2017; Kuvač Kraljević & Hržica, 2016). This was, in fact, our concept in construction of the present corpus of non-professional written Croatian. Participants covered a broad age range, from 10 years until old age, and came from different Croatian counties. The rationale behind the choice of lower age is the fact that in this period writing becomes automatized, text generation includes more mature discourse structures in a variety of literary genres, and finally posttranslation reviewing/revising and advanced preplanning emerge (Berninger, Fuller, & Whitaker, 1996).

During the creation of this corpus, we developed a protocol that defined discourse elicitation tasks and the methods to be used for the data analysis which is similar to protocols for other corpora (e.g., MacWhinney et al., 2011). Our protocol stipulated six groups of tasks representing different writing styles (descriptive, expository, narrative, and letter) and different levels of formality. The content of the six groups of tasks differed slightly across participant age groups, but style and formality were constant.

## 2. WRITERS WITH LANGUAGE DISORDERS

Persons with language disorders (e.g., Developmental Language Disorder (DLD), dyslexia, aphasia) are a specific group of non-professional speakers and writers. All language disorders can be classified into two basic groups according to the time of their occurrence and aetiology. Some disorders emerge in early or middle childhood and some are acquired in adulthood, i.e., in the period when spoken language is already automatized (Trauner & Nass, 2017). Some disorders, such as aphasia or traumatic brain injury, have a clear aetiology, while others are vaguer considering the cause of their occurrence. Children with DLD (formerly known as Specific Language

Impairment, SLI) show a late onset of language in childhood and have difficulties comprehending and producing verbal information. Approximately 7.6% of children show difficulties in acquisition of their mother tongue even when their cognitive functioning is typical, hearing is intact and language environment is adequate (Reed, 2005; Tomblin et al., 1997). The prevalence of DLD drives interest in understanding affected individuals' language performance, which can strongly influence skills mastery and overall academic achievement. Poor spoken language skills can be a trigger for poor academic achievement, poor reading and writing. Children whose reading achievement falls significantly below the expected level with respect to their chronological age, measured intelligence, and age-appropriate education, will be recognized as children with dyslexia (WHO, 2012). The disturbance in reading and writing significantly interferes with academic achievement or with any activity of daily living that requires those skills. Aphasia, in contrast, is one of the most prevalent acquired language disorders, occurring as a result of stroke or brain injury. Aphasia destroys communication skills, so it can have a devastating effect on psychological well-being and participation in life. According to the American Speech-Language-Hearing Association (ASHA; https://www.asha.org/practice-portal/clinical-topics/aphasia/#collapse_1), 1 in 250 people live with aphasia. According to the National Institute of Neurological Disorders and Stroke (NINDS; https://www.ninds.nih.gov/About), traumatic brain injury (hereafter referred to as TBI), a form of acquired brain injury, occurs when a sudden trauma causes damage to the brain. TBI can result when the head suddenly and violently hits an object, or when an object pierces the skull and enters brain tissue.

The increasing incidence of language disorders in society has led to an increase in the number of clinical corpora. TalkBank (https://talkbank.org), a large database of spoken-language corpora in different languages, includes several databases of clinical corpora. For example, the largest database of spoken language samples produced by persons with acquired language disorder is AphasiaBank (MacWhinney et al., 2011), based primarily on individuals whose aphasia resulted from a stroke that was verified through neuroimaging or definitive medical diagnosis. Established in 2007, AphasiaBank contains narrative, procedural, personal, and descriptive discourse from 290 persons with aphasia, as well as 190 control participants (MacWhinney & Fromm, 2016). This and other specialized clinical corpora can contribute to planning therapy as well as developing language tests and software solutions for augmentative and alternative communication.

All types of language disorders have a particularly negative influence on writing skills: affected individuals produce text much more slowly and with lower phonological accuracy than those with typical language skills, and the texts tend to be shorter and to feature simpler sentences with less diverse vocabulary (Bishop & Clarkson, 2003; Dockrell, Lindsay, & Connelly, 2009). The language processing problems of affected individuals mean that they may make different types of errors than writers with typical language skills (e.g., Ramus, 2014; Salmelin, Service, Kiesilä, Uutela, & Salonen, 1996).

Despite the increasing availability of clinical corpora, the literature on writing skills of people with language disorders is not so extensive (Zourou, Ecalle, Magnan, & Sanchez, 2010). Detailed insights are lacking for most languages, including Croatian. This likely reflects, in part, the relatively small number of specialized corpora and their small size. For example, there are two corpora of texts produced by speakers with dyslexia; one in Spanish contains only approximately 1,000 tokens (Rello, Baeza-Yates, Saggion, & Pedler, 2012) and the second in English 12,000 tokens (Pedler, 2007).

The present Croatian Corpus of Non-Professional Written Language includes data from people diagnosed with various types of language disorders, including Developmental Language Disorder, dyslexia, aphasia and TBI. Producing texts to answer to specific language tasks requires integration of a number of language skills on different language levels. By analysing such text, we can better understand language deficits of people with language disorders. Corpus also includes individuals with typical language status, allowing detailed comparisons of the two populations sampled under comparable conditions. We expect that people with language disorders would show lower productivity across the various writing genres.

## 3. CREATING A CROATIAN CORPUS OF NON-PROFESSIONAL WRITTEN LANGUAGE

### 3.1. Participants

The corpus comprises written texts from 395 participants (Table 1), all of whom were native speakers of Croatian and 267 of whom were recruited from the following institutions where they were receiving therapy for language disorders: Polyclinic for the Rehabilitation of Listening and Speech SUVAG in Zagreb (N = 140) and Osijek (N = 14), the Clinical Hospital in Split (N = 12), the Clinical Hospital in Osijek (N = 22), the Clinical Hospital Sveti Duh in Zagreb (N = 7), the General Hospital in Požega (N = 7), the Dr. Josip Benčević General Hospital in Slavonski Brod (N = 4), the Special Hospital for Medical Rehabilitation in Krapinske Toplice (N = 36), the Polyclinic for Rehabilitation of People with Developmental Disorders in Split (N = 14) and the Specialized Hospital for Medical Rehabilitation in Lipik (N = 5). These individuals were recruited in 2015 and 2016. Participants with language disorders who have already been assigned one of the following diagnosis codes F80.1, F80.2, F80.9, F81.0, F81.1, F81.3, R47.0, S00.0, S01.0 and S06 (WHO, 2012) and who have already been involved in the speech and language therapy, were included in the study. Clinical decision on the presence of a language disorder was based on the results of at least two formal tests (e.g., Peabody Picture Vocabulary Test, Dunn et al., 2009 and Test for Reception of Grammar, Bishop, Kuvač Kraljević, Hržica, Kovačević, & Kologranić Belić, 2014), and authentic assessment measures. In parallel, individuals with typical language skills were recruited through public calls in Split (N = 23), Zagreb (N = 81), Krapina County (N = 16) and Slavonia County (Požega, Slavonski Brod, Osijek and Lipik, N = 14). Participants were selected randomly, but they had to meet several inclusion criteria, reported by their parents, teachers, or themselves: (a) they had no hearing impairments; (b) spoke Croatian as their primary language; (c) did not report any developmental problems such as cognitive and language or difficulties with attention and (d) had no history of special education services. Participants, whose age ranged from 10 to 80 years (mean age 52), were recruited from various locations all over Croatia in order to ensure the representativeness of the sample. Indeed, both groups of participants included speakers of all three dialects (kajkavian, chakavian and shtokavian).

Developmental Language Disorder (DLD) and dyslexia were more prevalent among the children and adolescents in the sample than among the adults, probably reflecting that clinical care and support for such disorders tends to occur earlier in life, despite the fact that the disorders are lifelong. Acquired language disorders such as aphasia and TBI, conversely, were more prevalent among adults in the sample, reflecting the fact that such disorders usually occur later in life.

Table 1. Data about participants
Tablica 1. Podatci o ispitanicima

| Age group / Dobna skupina | | | |
|---|---|---|---|
| | Male / Muški | Female / Ženski | Total / Ukupno |
| Children / Djeca | 111 | 59 | 170 |
| Adolescents / Adolescenti | 18 | 16 | 34 |
| Adults / Odrasli | 78 | 113 | 191 |
| Total / Ukupno | 207 | 188 | 395 |
| Language status / Jezični status | | | |
| | Male / Muški | Female / Ženski | Total / Ukupno |
| Typical / Tipičan | 51 | 83 | 134 |
| DLD / Razvojni jezični poremećaj (RJP) | 32 | 21 | 53 |
| Other LD known aetiology / Jezični poremećaji ostalih etiologija | 5 | 3 | 8 |
| Dyslexia / Disleksija | 76 | 34 | 110 |
| TBI / Traumatska ozljeda mozga (TOM) | 19 | 7 | 26 |
| Aphasia / Afazija | 34 | 30 | 64 |
| Total / Ukupno | 217 | 178 | 395 |

## 3.2. Discourse elicitation tasks

Language samples were collected over a period of eight months (from September 2015 to April 2016) by speech and language therapists. This short period of data collection makes the corpus synchronous. In the case of participants with language disorders, the therapists collected the data in the clinical setting. In the case of participants with typical language status, equivalent data were collected in a non-clinical setting (e.g., at home, at work). Tasks were designed to elicit two modes of discourse: description and narration (both storytelling and recounts – see Heath, 1986). Additionally, the formality of the tasks was varied, as it has been shown that it affects text properties (Biber, 1988, 2006). During one or more sessions, participants completed 10 tasks

(Table 2): two essays, two sets of questions, two stories (picture prompts), two non-formal letters, two formal letters and two dictations. Participants received precise instructions for each task, and writing had no time limit. Examiners prompted participants when they failed to complete tasks. These prompts, as well as the original instructions, were carefully scripted beforehand.

Since the goal was to capture non-professional written language, the tasks were designed to resemble typical language use (e.g., description of a familiar person, personal narratives). Both descriptions and narratives are regularly used by speakers of different age (e.g., Dipper & Pritchard, 2017). Most of the used tasks were the same for all age groups (e.g., "describe your home"), but two of them were slightly adjusted for different age groups. First, the formal letter-writing task involved different age-adjusted instructions, such as writing an invitation to a playdate (if the subject was a child), inviting a friend over (for adolescents), or scheduling a business meeting (if the subject was an adult). Second, a task of writing a non-formal letter included writing a postcard either to grandparents (if the subject was a child), or to family or friends (if the subject was an adolescent or an adult).

Overall, the appropriateness of tasks for a particular age group was established by: (1) taking into consideration the speaker's experience (e.g., everyday situations a speaker encounters), (2) using prompts already established as relevant and appropriate for eliciting discourse (such as sets of pictures from *Expression, reception and recall of narrative instrument* (Bishop, 2004) used in a number of studies of child and adult language).

Dictation consisted of a paragraph which was read aloud from a book. For children and adolescents, book paragraphs from the age-appropriate obligatory school reading lists were used (see Table 2). For adults, the selected texts included classic works of Croatian literature.

The tasks also differed in format depending on the age group. For adolescents and adults, each written sample consisted of 10 handwritten tasks and two computer-based writing tasks (one set of answers to questions and one dictation). For children, none of the tasks were computer-based. Children wrote one dictation, while adults wrote two.

If a participant became tired, he or she could continue writing in the next session. Most participants completed all 10 writing tasks within a single session lasting 40 minutes. Some participants, primarily older adults with language disorders, required two sessions in order to complete all tasks.

**Table 2.** Writing tasks
**Tablica 2.** Pisani zadatci

| | Writing task / Pisani zadatak | Theme / Tema | Writing style / Stil pisanja |
|---|---|---|---|
| 1 | Essays (two) / Eseji (dva) | a) My best friend<br>b) My home | Descriptive |
| 2 | Responses to questions (two sets) / Odgovori na pitanja (dva seta) | a) Questions 1:<br>1) What do you do during the weekend?<br>2) How do you celebrate your birthday?<br>3) What is your favourite school subject and why? / What do you like about your job (current or previous)?<br>4) What do you see through the window?<br>5) What are you wearing today?<br><br>b) Questions 2:<br>1) What are your hobbies?<br>2) What do you do during recess? / What do you do during the Christmas holidays?<br>3) What is your favourite TV show and why?<br>4) Describe your favourite professor. / Describe your favourite actor.<br>5) What do you see through the window of your room? | Expository / Descriptive |
| 3 | Stories (two) / Priče (dvije) | a) Beach Story*<br>b) Fish Story* | Narrative |
| 4 | Non-formal letters (two) / Neformalno pismo (dva) | a) Letter to friend<br>b) Postcard to grandparents / Postcard to family or friends | Non-formal discourse |
| 5 | Formal messages (two) / Formalna poruka (dvije) | a) Invitation to playdate / meeting<br>b) Cancelling training / lecture / meeting | Formal discourse |
| 6 | Dictations (one for children, two for adolescents and adults) / Diktati (jedan za djecu, dva za adolescente i odrasle) | Full list of books used for dictations:<br>Hrvoje Hitrec: *Eko Eko* (10 years), Sanja Pilić: *Mrvice iz dnevnog boravka* (11), Melita Rundek: *Psima ulaz zabranjen* (12), Damir Miloš: *Bijeli klaun* (13), Višnja Stahuljak: *Don od Tromeđe* (14), Vjenceslav Novak: *U glib* (adolescents), Pavao Pavličić: *Dobri duh Zagreba* (adolescents), Miroslav Krleža; *Povratak Filipa Latinovicza* (adults), Dinko Šimunović: *Muljika* (adults) | Narrative default discourse |

* Participant wrote stories based on sets of pictures from *Expression, reception and recall of narrative instrument* (Bishop, 2004).

In all, 395 participants produced more than half a million tokens in more than 41,000 utterances (Table 3).

**Table 3.** Basic corpus information
**Tablica 3.** Osnovne informacije o korpusu

| | Children / Djeca | | Adolescents / Adolescenti | | Adults / Odrasli | | Total / Ukupno | |
|---|---|---|---|---|---|---|---|---|
| | N | No. of tokens / Broj pojavnica | N | No. of tokens / Broj pojavnica | N | No. of tokens / Broj pojavnica | N | No. of tokens / Broj pojavnica |
| Typical language status / Tipičan jezični status | 17 | 24,538 | 15 | 26,745 | 101 | 205,100 | 133 | 356,383 |
| Dyslexia / Disleksija | 95 | 98,528 | 8 | 12,656 | 7 | 12,771 | 110 | 123,955 |
| Developmental Language Disorder / Razvojni jezični poremećaj | 48 | 52,010 | 3 | 3,422 | 2 | 2,997 | 53 | 123,955 |
| Other language disorders known aetiology / Jezični poremećaji ostalih etiologija | 8 | 9,161 | 0 | 0 | 0 | 0 | 8 | 9,161 |
| Broca's aphasia / Brokina afazija | 1 | 316 | 0 | 0 | 42 | 56,891 | 43 | 57,207 |
| Wernicke's aphasia / Wernickeova afazija | 0 | 0 | 0 | 0 | 1 | 949 | 1 | 949 |
| Anomic aphasia / Anomija | 0 | 0 | 0 | 0 | 7 | 12,194 | 7 | 12,194 |
| Other types of aphasia / Ostale vrste afazija | 0 | 0 | 0 | 0 | 14 | 20,945 | 14 | 20,945 |
| Traumatic brain injury / Traumatska ozljeda mozga | 0 | 0 | 8 | 14,441 | 18 | 24,549 | 26 | 38,990 |
| Total / Ukupno | 169 | 184,553 | 34 | 57,264 | 192 | 336,396 | 395 | 743,739 |

### 3.3. Annotation

The corpus was annotated using the morphosyntactic annotations from version 4 of the MULTEXT-East Morphosyntactic Specifications for Croatian (Ljubešić, 2013). Due to its specificities (non-professional corpus, clinical corpus), the corpus was in a large part annotated manually. However, inflectional lexicon for Croatian (hrLex – Ljubešić, 2019; Ljubešić, Klubička, Agić, & Jazbec, 2016) was used to facilitate manual annotation. Annotators were provided with possible options for annotations retrieved from hrLex, but they could also add their own. Manual annotation consisted of three phases. First, surface forms were corrected. This included correction of tokens: tokens could be divided into two or merged if word boundaries were displaced. Second, manual morphosyntactic annotation was performed. Third, errors were marked according to one of 12 types (Štefanec, Ljubešić, & Kuvač Kraljević, 2016). Annotators were experienced in text normalization and lemmatization. In order to retain the characteristics of the language used by participants, annotators were instructed to keep non-standard language features such as regionalisms or slang. Only unintentional or orthographic errors were corrected.

## 4. ILLUSTRATIVE ANALYSES

To illustrate how transcripts from the Croatian Corpus of Non-Professional Written Language can be applied, we performed one analysis from a developmental perspective and one from a clinical perspective. It is important here to emphasize that analyses will not be comprehensive, since the idea of this paper is only to provide an example of how and for what purpose the corpus can be used.

Using exploratory and confirmatory factor analyses, Puranik, Lombardino, and Altmann (2008) and Wagner et al. (2011) demonstrated that four factors can be conceptualized from written texts: productivity (e.g., total number of words, number of different words, total number of sentences), complexity (e.g., mean length of T- or C-unit or clausal density), accuracy (the proportion of grammatical and spelling errors to the total number of sentences), and mechanics (number of capitalization and punctuation errors). These four factors were part of the translation component of the writing process, i.e., the phase of production of written text and should be considered when evaluating writing (Koutsoftas & Gray, 2012). In the two analyses that follow, we applied productivity measures, and in the second we applied additional basic analysis of accuracy and mechanical errors.

These analyses will be conducted on two different genres: on essays and narratives. Research shows that the differences among participants in writing ability are ensured by different genres, as well as by familiarity with the topic being written about (Hržica, Košutar, & Kramarić, 2019). Therefore, the essay topic was defined by the writers or participants themselves while in narration it was defined by researchers. The analyses presented here are just a few made possible by the corpus.

## 4.1. Language productivity from a developmental perspective

Both analyses – the average numbers of tokens (words) and average numbers of utterances generated during the task – were based on language productivity, which can reliably assess language proficiency (Tilstra & McMaster, 2007). The number of tokens and utterances in essays were calculated for a randomly selected subset of participants with typical language development from three i.e., four age groups (Table 4). The number of tokens increased with age up to 60 years of age, after which the number decreased. This decline is biologically determined: elderly, especially starting around 70 years old, have difficulties in producing spoken and written language (Kemper, 1994; Rao, 2015). Children, adolescents and young adults produced similar numbers of utterances, and, as Table 4 shows, elderly produced the fewest utterances.

**Table 4.** Average numbers of tokens and utterances in different age groups
**Tablica 4.** Prosječan broj pojavnica i iskaza za različite dobne skupine

| Main age ranges in corpus / Raspon dobi u korpusu | Specific analysed age groups / Analizirane dobne skupine | N | Average number of tokens / Prosječan broj pojavnica | Average number of utterances / Prosječan broj iskaza |
|---|---|---|---|---|
| < 15 | 10–12 | 10 | 124.4 | 10.8 |
| 16–20 | 17–19 | 10 | 179.8 | 11.0 |
| > 21 | 30–32 | 10 | 177.7 | 10.8 |
| | 60–70 | 10 | 116.7 | 8.8 |

This developmental analysis illustrates how the corpus data can support studies into, for example, normative values for language measures in different age ranges and age-related differences such as in types of sentences, conjunctions, as well as temporal and referential anaphora. The data allow analysis of linguistic improvements in style, content, and grammar of written text during elementary school; such studies would need to control for how demanding the school curriculum is.

## 4.2. Language productivity, accuracy and errors of mechanics from a clinical perspective

The two written narratives produced by young writers with Developmental Language Disorder (DLD) and older writers with Broca's aphasia will be contrasted here. DLD is diagnosed when the language skills of a child without any known biomedical condition, such as autism spectrum disorder or cognitive deficit, are persistently below the level expected for the child's age. This disorder interferes with the child's ability to communicate effectively with other people and strongly affects academic achievement (Bishop, 1999). Aphasia, in contrast, occurs later in life as a result of a clear neurological disturbance, usually brain stroke, and it affects reading, writing, speaking and language comprehension (Hegde, 2006). These two types of disorders manifest with similar language symptomatology, yet they usually occur at quite different ages and have quite different causes.

Table 5. Average numbers of tokens and utterances in participants with DLD and aphasia
Tablica 5. Prosječan broj pojavnica i iskaza kod ispitanika s RJP-om i afazijom

| Group / Skupina | N | Age (yrs.) range / Raspon dobi (god.) | Average number of tokens / Prosječan broj pojavnica | Average number of utterances / Prosječan broj iskaza |
|---|---|---|---|---|
| Children with Developmental Language Disorder (DLD) / Djeca s razvojnim jezičnim poremećajem (RJP) | 41 | 11–15 | 122.9 | 12.3 |
| Adults with Broca's aphasia / Odrasli s Brokinom afazijom | 45 | > 21 | 118.8 | 10.9 |

Language productivity analysis revealed that both groups produced similar number of tokens and utterances (Table 5). Based on the analysis of narrative text from the same corpus, Kuvač Kraljević, Matić, and Olujić Tomazin (in press) determined that adolescents and adults with typical language skills produce approximately 180 words in their narratives which is significantly more than their peers with language disorders. This is in line with other studies that claim that adolescents with DLD and adults with aphasia produce significantly shorter texts than TD adolescents and adults regardless the language orthography (e.g., Mackie & Dockrell, 2004; Marini, Andreetta, del Tin, & Carlomagno, 2011).

The relationship between productivity and many other measures of written language output is confirmed. For example, it is well known that productivity correlates with accuracy – the more children or adults with typical language status write, they make less spelling and grammatical errors (Dockrell, Lindsay, Connelly, & Mackie, 2007), and produce more informative stories (Koutsoftas & Gray, 2012). Preliminary analysis of texts produced by adolescents with DLD and adults with Broca's aphasia showed that both groups produce most often simple errors that differ from intended word by only one single grapheme, such as omission and addition of grapheme, multiple errors that differ in more than one grapheme, such as omission of syllables (Rello et al., 2012). In our sample adolescents with DLD produced 48% and persons with aphasia 40% of these errors. Further, they have problem with the rules of capitalizations and punctuation, such as substitution of upper- and lower-case letters, diacritical marks omission, commas, and dots (26% in DLD group and 14% in aphasia). The latter types of errors are those that Koutsoftas and Gray (2012) call mechanic errors. Further, accuracy with special focus on grammatical errors such as inappropriate inflection, copula omission or duplication, stringing sentences, and inappropriate word order revealed that adults with aphasia have more problems with stringing sentences than children with DLD. Problems with noun inflection, verb changes and omission of copula are common in both groups. Problem with retrieving the correct lexical item is a more prominent feature of texts written by adults with aphasia. Consequently, persons with aphasia more often produce neologisms. Deeper analyses of different types of errors based on this corpus can be found in Luketin (2015), Kuvač Kraljević et al. (in press), and Štefanec et al. (2016). Besides the productivity and accuracy, the corpus data allows comparisons of macro-organization, informativeness, text quality, morphological complexity and other language characteristics between different types of disorders.

## 5. CONCLUSION

The aim of this study was to present the first Croatian Corpus of Non-Professional Written Language, and it is only Croatian corpus that includes language samples from persons with language disorders. Such clinical corpora are less common, despite the increasing amount of data reported in the clinical literature. The corpus described here offers a unique language resource, annotated on multiple levels, for diverse lines of research, for which we provide here only a sampling.

While many believe that written corpora are easier to develop than spoken ones, creating a reliable, representative written corpus means taking into account several factors, which we have tried to describe in our case. It is important to note that non-professional writers differ regarding their age, gender, socioeconomic status and other characteristics. Therefore, it would be interesting to examine similarities and differences in writing skill with respect to these demographic features in further analyses. Also, written texts differ according to genre and level of formality. Such questions are important when planning language sampling for written corpora. We believe that this corpus has an adequate size to pose numerous research questions, since we were careful to include diverse written genres and a broad age range of participants who were speakers of all three major dialects of the language. Specialized corpora are much smaller than core corpora because they tend to focus on specific areas of language to respond to narrow clinical or pedagogical needs (Nelson, 2010).

This corpus has already been used in several studies aimed to illuminate the skill of writing in the Croatian language (Hržica et al., 2019; Štefanec et al., 2016), as well as for the development of automatic tools for morphosyntactic description, namely lemmatizers (Ljubešić & Štefanec, 2000a, 2000b) and models for morphosyntactic annotation (Ljubešić & Štefanec, 2000c, 2000d).

In the future, this corpus will be expanded with transcripts based on new genres produced by new groups of writers. Also, the plan is to network it with the European Research Infrastructure for Language Resources and Technology (www.clarin.eu). In this way, the corpus will be publicly available and it will fulfil the criteria of a "FAIR" resource that is findable, accessible, interoperable and reusable for fundamental and clinical research in writing skills and language processing.

## ACKNOWLEDGMENTS

This study was carried out within the project Computer Assistant for Text Input for Persons with Language Impairment (RAPUT; RC.2.2.08 – 0050), funded by the European Structural and Investment Funds. We are grateful to the speech and language pathologists and all participants who helped develop the corpus. We sincerely appreciate Vanja Štefanec for assistance with corpus annotation.

## REFERENCES

American Speech-Language-Hearing Association. (n.d.). *Aphasia* (Practice Portal). Retrieved October 1st, 2019, from www.asha.org/Practice-Portal/Clinical-Topics/Aphasia/

**Berninger, V. W., Fuller, F., & Whitaker, D.** (1996). A process model of writing development across the life span. *Educational Psychology Review*, *8*(3), 193–218.

**Biber, D.** (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

**Biber, D.** (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam; Philadelphia: J. Benjamins.

**Bishop, D. V. M.** (1999). *Uncommon understanding: Development and disorders of language comprehension in children*. Hove: Psychology Press.

**Bishop, D. V. M.** (2004). *Expression, reception and recall of narrative instrument (ERRNI)*. London: Pearson Assessment.

**Bishop, D. V. M., & Clarkson, B.** (2003). Written language as a window in to residual language deficits: A study of children with persistent and residual speech and language impairments. *Cortex*, *39*(2), 215–237. https://doi.org/10.1016/S0010-9452(08)70106-0

**Bishop, D. V. M., Kuvač Kraljević, J., Hržica, G., Kovačević, M., & Kologranić Belić, L.** (2014). *Test razumijevanja gramatike (TROG-2:HR)* [*Test of receptive grammar (TROG-2:HR)*]. Zagreb/Jastrebarsko: Naklada Slap.

British National Corpus. Retrieved October 5th, 2019, from http://www.natcorp.ox.ac.uk/corpus/index.xml

Cambridge English Corpus. Retrieved November 25th, 2020, from https://www.cambridge.es/en/about-us/cambridge-english-corpus

**Carter, R. A.** (1997). Speaking Englishes, speaking cultures, using CANCODE. *Prospect*, *12*(2), 4–11.

**Carter, R. A.** (1998). Orders of reality: CANCODE, communication and culture. *ELT Journal*, *52*(1), 43–56.

**Carter, R. A.** (1999). Common language: Corpus, creativity and cognition. *Language and Literature*, *8*(3), 1–21.

**Dipper, L. T., & Pritchard, M.** (2017). Discourse: Assessment and therapy. In F. Fernandes (Ed.), *Advances in speech-language pathology*. IntechOpen. doi: 10.5772/66241

**Dockrell, J. E., Lindsay, G., & Connelly, V.** (2009). The impact of specific language impairment on adolescents' written text. *Exceptional Children*, *75*(4), 427–446. doi.org/10.1177/001440290907500403

**Dockrell, J. E., Lindsay, G., Connelly, V., & Mackie, C.** (2007). Constraints in the production of written text in children with specific language impairments. *Exceptional Children*, *73*(2), 147–164.

**Dunn, L. M., Dunn, L. M., Kovačević, K., Padovan, N., Hržica, G., Kuvač Kraljević, J., Mustapić, M., Dobravac, G., Palmović, M.** (2009). *Peabody Picture Vocabulary Test (PPVT-III-HR)*. Zagreb/Jastrebarsko: Naklada Slap.

Federal Energy Regulatory Commission. (2012). *Enron email data corpus*. Retrieved October 1st, 2019, from https://www.cs.cmu.edu/~enron/

**Heath, S.** (1986). Taking a cross-cultural look at narratives. *Topics in Language Disorders*, *7*(1), 84–94.

**Hegde, M. N.** (2006). *A coursebook on aphasia and other neurogenic language disorders*. NY: Delmar.

**Hržica, G., Košutar, S., & Kramarić, M.** (2019). Rječnička raznolikost pisanih tekstova osoba s razvojnim jezičnim poremećajem [Lexical diversity in written texts of persons with developmental language disorders]. *Hrvatska revija za rehabilitacijska istraživanja* [*Croatian Review of Rehabilitation Research*], *55*(2), 14–30.

**Kamandulytė-Merfeldienė, L.** (2017). Grammatically coded corpus of spoken Lithuanian: Methodology and development. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, *11*(4), 821–825.

**Kemper, S.** (1994). Elderspeak: Speech accommodations to older adults. *Aging, Neuropsychology, and Cognition*, *1*(1), 17–28.

**Kennedy, G. D.** (2014). *Introduction to the corpus linguistics*. NY: Routledge.

**Koutsoftas, A. D., & Gray, S.** (2012). Comparison of narrative and expository writing in students with and without language-learning disabilities. *Language, Speech, and Hearing Services in Schools*, *43*(4), 395–409.

**Kovačević, M.** (2002). *Croatian child language corpus*. CHILDES. http://childes.psy.cmu.edu/data/Slavic

**Kuvač Kraljević, J., & Hržica, G.** (2016). Croatian adult spoken language corpus (HrAL). *Fluminensia*, *28*(2), 87–102.

**Kuvač Kraljević, J., Hržica, G., & Lice, K.** (2017). CroDA: A Croatian discourse corpus of speakers with aphasia. *Croatian Review of Rehabilitation Research*, *53*(2), 61–71. https://doi.org/10.31299/hrri.53.2.5

**Kuvač Kraljević, J., Matić, A., & Olujić Tomazin, M.** (in press). Written narratives of adolescents with developmental language disorder, typically developing adolescents and adults. *Written Language & Literacy.*

**Leech, G.** (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Trends in Linguistics: Directions in Corpus Linguistics*. *Studies and Monographs 65* (pp. 105–122). Berlin/NY: Mouton de Gruyter.

**Luketin, J.** (2015). *Analiza pisanog jezika odraslih osoba s jezičnim poremećajima* [*Written language analysis of adults with language disorders*] (Master thesis). Zagreb: Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu.

**Ljubešić, N.** (2013). MULTEXT-East Croatian Morphosyntactic Specifications [Computer software]. Revised Version 4. Retrieved September 5th, 2019, from http://nlp.ffzg.hr/data/tagging/msd-hr.html

**Ljubešić, N.** (2019). *Inflectional lexicon hrLex 1.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1232

**Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I.-P.** (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4264–4270). Portorož, Slovenija: European Language Resources Association (ELRA).

**Ljubešić, N., & Štefanec, V.** (2020a). *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Croatian 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1333

**Ljubešić, N., & Štefanec, V.** (2020b). *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Serbian 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1333

**Ljubešić, N., & Štefanec, V.** (2020c). *The CLASSLA-StanfordNLP model for morphosyntactic annotation of non-standard Croatian 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1331

**Ljubešić, N., & Štefanec, V.** (2020d). *The CLASSLA-StanfordNLP of non-standard Serbian 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1332

**Mackie, C., & Dockrell, J. E.** (2004). The nature of written language deficits in children with SLI. *Journal of Speech, Language, and Hearing Research*, *47*(6), 1469–1483. doi: 10.1044/1092-4388(2004/109)

**MacWhinney, B.** (2002). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

**MacWhinney, B., & Fromm, D.** (2016). AphasiaBank as BigData. *Seminars in Speech and Language*, *37*(1), 10–22. doi: 10.1055/s-0036-1571357

**MacWhinney, B., Fromm, D., Forbes, M., & Holland, A.** (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286–1307. doi:10.1080/02687038.2011.589893

**MacWhinney, B., & Wagner, J.** (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung*, *11*, 154–173.

**Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S.** (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology*, *25*(11), 1372–1392. doi:10.1080/02687038.2011.584690(11)

**McCarthy, M., & O'Keeffe, A.** (2010). Historical perspectives: What are corpora and how they evolved? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook Of Corpus Linguistics* (pp. 3–13). Oxford: Routledge.

**McEnery, T., & Wilson, A.** (2001). *Corpus linguistics: An introduction.* Edinburgh: Edinburgh University Press.

National Institute of Neurological Disorders and Stroke. Retrieved November 25th, 2020, from https://www.ninds.nih.gov/About

**Nelson, M.** (2010). Building a written corpus: What are the basic? In A. O`Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook Of Corpus Linguistics* (pp. 53–65). Oxford: Routledge.

**Pedler, J.** (2007). *Computer correction of real-word spelling errors in dyslexic text* (Doctoral thesis). London: London University.

**Puranik, C. S., Lombardino, L. J., & Altmann, L. J. P.** (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech Language Pathology*, *17*(2), 107–120.

**Ramus, F.** (2014). Neuroimaging sheds new light on the phonological deficit in dyslexia. *Trends in Cognitive Sciences*, *18*(6), 274–275.

**Rao, P. K. S.** (2015). Cognitive-communicative decline with aging: Do speech-language pathologists contribute to clinical decisions? *Indian Journal of Gerontology*, *29*(1), 1–22.

**Raso, T., & Mello, H.** (2014). *Spoken corpora and linguistic studies.* Amsterdam: John Benjamins Publishing.

**Reed, V. A.** (2005). *An introduction to children with language disorders*. NY: Pearson Education.

**Rello, L., Baeza-Yates, R., Saggion, H., & Pedler, J.** (2012). A first approach to the creation of a Spanish corpus of dyslexic texts. In N. Calzolari (Ed.), *Proceedings on LREC Workshop: Natural Language Processing for Improving Textual Accessibility* (pp. 22–27). European Language Resources Association (ELRA).

**Salmelin, R., Service, E., Kiesilä, P., Uutela, K., & Salonen, O.** (1996). Impaired visual word processing in dyslexia revealed with magnetoencephalography. *Annals of Neurology*, *40*(2), 157–162.

**Schler, J., Koppel, M., Argamon, S., & Pennebaker, J.** (2006). Effects of age and gender on blogging. *Proceedings of the 2006 AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 191–197).

**Svartvik, J.** (1992). Corpus linguistics comes of age. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 7–15). Berlin/NY: Mouton de Gruyter.

**Štefanec, V., Ljubešić, N., & Kuvač Kraljević, J.** (2016). Error-annotated corpus of non- professional written language. In N. Calzolari (Ed.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/index.html

**Tadić, M.** (2009). New version of the Croatian national corpus. U D. Hlaváčková, A. Horák, K. Osolsobě, & P. Rychlý (Eds.), *After half a century of Slavonic natural language processing* (pp. 199–205). Brno: Masaryk University.

TalkBank. Retrieved October 1st, 2019, from http://talkbank.org/

**Tilstra, J., & McMaster, K.** (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency?

*Communication Disorders Quarterly*, *29*(1), 43–53. https://doi.org/10.1177/1525740108314866

**Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M.** (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1245–1260.

**Trauner, D. A., & Nass, R. D.** (2017). Developmental language disorders. In K. Swaiman, S. Ashwal, & D. M. Ferriero et al. (Eds.), *Swaiman's pediatric neurology: Principles and practice* (6th ed.) (pp. 431–436). Philadelphia, PA: Elsevier.

**Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Gehron Wilson, L., Tschinkel, E., & Thatcher Kantor, P.** (2011). Modeling the development of written language. *Reading and Writing*, *24*, 203–220.

Word Health Organization. (2012). *ICD-10: International statistical classification of diseases and related health problems*, 10th revision (4th ed.). Geneva, Switzerland: Autor.

**Zourou, F., Ecalle, J., Magnan, A., & Sanchez, M.** (2010). The fragile nature of phonological awareness in children with specific language impairment: Evidence from literacy development. *Child Language Teaching and Therapy*, *26*(3), 347–358. doi: 10.1177/0265659010369288

**Jelena Kuvač Kraljević, Gordana Hržica**
*jkuvac@erf.hr, gordana.hrzica@erf.unizg.hr*
Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu
Hrvatska

**Lana Kologranić Belić**
*lana.kologranic@gmail.com*
Poliklinika za rehabilitaciju slušanja i govora SUVAG, Zagreb
Hrvatska

# Hrvatski korpus neprofesionalnoga pisanog jezika osoba s jezičnim poremećajima i osoba bez jezičnih poremećaja

## Sažetak

Korpusi, anotirani arhivi ljudske komunikacije, objektivan su i pouzdan izvor materijala za jezičnu analizu. U ovom se radu predstavlja hrvatski korpus neprofesionalnoga pisanog jezika nastao tijekom jednogodišnjega prikupljanja pisanih uzoraka osoba s jezičnim poremećajima i osoba bez jezičnih poremećaja. Ovaj korpus ima jedinstvenu vrijednost zbog jezičnih uzoraka neprofesionalaca, u usporedbi s korpusima temeljenima na tekstovima profesionalnih autora (npr. novinara, znanstvenika ili pisaca) koji obuhvaćaju uzorke stare više od stoljeća. K tome, ovaj korpus uključuje jezične uzorke osoba s jezičnim poremećajima i jezične uzorke osoba bez jezičnih poremećaja prikupljene u istim uvjetima, što otvara prostor za detaljnu analizu jezične uporabe. U radu se opisuju jezični zadatci (esej, pisanje priče, neformalno i formalno pismo te diktat) korišteni za proizvodnju teksta te procedure uzorkovanja i anotacije korištene za stvaranje korpusa. Korisnost je ilustrirana putem analiza jezične proizvodnje, tj. transkripata različitih žanrova koje su proizveli autori različite dobi, odnosno jezičnoga statusa. Opisani korpus može biti koristan za analizu jezičnih vještina govornika hrvatskoga jezika, bilo da se radi o osobama s jezičnim poremećajima ili osobama bez jezičnih poremećaja.

**Ključne riječi:** hrvatski korpus neprofesionalnoga pisanog jezika, pisani jezik, žanrovi, jezični poremećaji