*Jelena Kuvač Kraljević, Gordana Hržica*

# CROATIAN ADULT SPOKEN LANGUAGE CORPUS (HrAL)*

*dr. sc. Jelena Kuvač Kraljević, Edukacijsko-rehabilitacijski fakultet, jkuvac@erf.hr, Zagreb*
*dr. sc. Gordana Hržica, Edukacijsko-rehabilitacijski fakultet, gordana.hrzica@erf.hr, Zagreb*

*pregledni članak*
UDK 811.163.42(038)
rukopis primljen: 13. 9. 2016.; prihvaćen za tisak: 15. 12. 2016.

*Interest in spoken-language corpora has increased over the past two decades leading to the development of new corpora and the discovery of new facets of spoken language. These types of corpora represent the most comprehensive data source about the language of ordinary speakers. Such corpora are based on spontaneous, unscripted speech defined by a variety of styles, registers and dialects.*
*The aim of this paper is to present the* Croatian Adult Spoken Language Corpus *(HrAL), its structure and its possible applications in different linguistic subfields. HrAL was built by sampling spontaneous conversations among 617 speakers from all Croatian counties, and it comprises more than 250,000 tokens and more than 100,000 types. Data were collected during three time slots: from 2010 to 2012, from 2014 to 2015 and during 2016.*
*HrAL is today available within* TalkBank*, a large database of spoken-language corpora covering different languages (https://talkbank.org), in the* Conversational Analyses corpora *within the subsection titled* Conversational Banks*. Data were transcribed, coded and segmented using the transcription format* Codes for Human Analysis of Transcripts *(CHAT) and the* Computerised Language Analysis *(CLAN) suite of programmes within the* TalkBank *toolkit. Speech streams were segmented into communication units (C-units) based on syntactic criteria. Most transcripts were linked to their source audios. The* TalkBank *is public free, i.e. all data stored in it can be shared by the wider community in accordance with the basic rules of the* TalkBank*.*
*HrAL provides information about spoken grammar and lexicon, discourse skills, error production and productivity in general. It may be useful for sociolinguistic research and studies of synchronic language changes in Croatian.*

**Key words:** Croatian Adult Spoken Language Corpus *(HrAL); language sampling; spontaneous speech corpora*

* Acknowledgment: The work on this paper was supported by the Croatian National Foundation, grant HRZZ-2421 for the project Adult language processing.

## 1. Introduction

The term *spoken language corpus* refers to a collection of language and communication data that serves as a tool for understanding the nature of human language and communication (Wichmann, 2008). Such a corpus is based on spontaneous, unscripted speech, i.e. speech that occurs in real time, so it contains hesitations, repetitions, false starts and unintentional errors. Spoken language is a multidimensional phenomenon defined by a variety of styles, registers and dialects. All these possible variations are impossible to capture in a single corpus, so corpora are designed depending on the researcher's objectives.

The first spoken-language corpora were developed in different linguistic subfields for different purposes. For example, Carpenter Fries (1952) developed a 250 000-word corpus based on transcripts of conversations between speakers of American English residing in the north-central United States, with the aim of building a grammar based on everyday spoken language. *Oral Vocabulary of the Australian Workers* (Schonell et al., 1956) was developed to allow analyses of idiomatic words and phrases, while the *London-Lund Corpus* (Svartvik & Quirk, 1980; Svartvik, 1990) was created to assess the grammatical accuracy of adult language. Brown (1973) collected child spoken-language samples to study language acquisition. Interest in spoken-language corpora has increased over the past two decades, leading to research that has revealed new facets of spoken language (see Leech, 2000).

With respect to the following, three criteria are important when designing a corpus: (1) the authenticity of language in the corpus, (2) the representativeness of language in the corpus and (3) the criteria used to sample language (Tognini Bonelli, 2001). Fulfilling these criteria can be challenging when assembling a corpus of spoken language because this type of language involves less clear orthographic boundaries, syntactic chunking, and spontaneous, simultaneous conversation.

A corpus of spoken language is inevitably smaller than a corpus of written (text-based) language. Some authors, such as Leech (2000) and McCarthy & O'Keeffe (2008), argue that the composition of a corpus in terms of genres and other design features is more important than its size. O'Keeffe & Farr (2003) suggest that while written corpora with fewer than 5 million words should be considered quite small, spoken corpora with more than 1 million words should be considered large.

While most spoken-language corpora are based on free conversation (e.g., McEnery & Wilson, 2001), a trend in corpus linguistics is to include the spoken language of professional speakers such as TV hosts and lecturers. This approach raises important methodological questions. How much spontaneity is present in this type of conversation if many broadcasts and lectures are scripted and participants strive to use formal language varieties? How well does this type of conversation represent spoken varieties in a given language?

In larger corpora, such as those at the national level, the financial and time costs of transcription mean that spoken language, usually from free conversation, makes up

only a small part of the overall corpus. For example, the proportion of spoken language in the *British National Corpus* is only 10% (Burnard, 2000). This raises questions about the authenticity, balance and representativeness of national corpora.

The aim of the present work was to compile the first *Croatian Adult Spoken Language Corpus* (HrAL). HrAL was developed with the intent of being representative of non-professional spoken language, reflecting the horizontal and vertical variability in language.

## 2. Building HrAL

### 2.1. Procedure

Data for the corpus were collected from 2010 to 2012, from 2014 to 2015 and during 2016. Speakers of the Croatian language from different parts of Croatia with access to groups of speakers (friends and families) were recruited and trained to collect samples of spoken language. These trained investigators were responsible for recruiting participants, i.e. groups of 3-8 adult speakers who regularly engage in informal spoken interaction. Different groups of investigators (and speakers) participated in each time period during which the corpus was collected. The majority of investigators were students, and the groups of speakers were their families or friends. All speakers were informed about the speech sampling procedure and the study objectives, and they were told that they could withdraw from the study at any time with no questions asked. All speakers gave a written informed consent.

Sampling was performed in informal situations, predominantly spontaneous conversations among friends, relatives or acquaintances during family meals, informal gatherings or socialising. To mitigate the effect of the Observer's paradox (Labov, 1972), two procedures were followed. First, the participants were informed, when they gave their consent, that they would be recorded at some point in the coming four weeks, but that they would not know exactly when. Thus, speakers were unaware of when they were being recorded. Second, the investigators were instructed to distance themselves from the situation, participating as little as possible in the conversation or performing other activities during the recording. Occasionally, the investigators were even absent while the conversations were recorded. The investigators' utterances were coded with "INV" during transcription to allow easy removal during subsequent analyses as required.

Most recording sessions lasted approximately 15 min (Figure 1). Most samples shorter than 15 min were fragments of longer conversations that were recorded over multiple samples. Samples belonging to the same conversation were marked accordingly so that they could be analysed as a single conversation in subsequent analyses (see section 3.1).

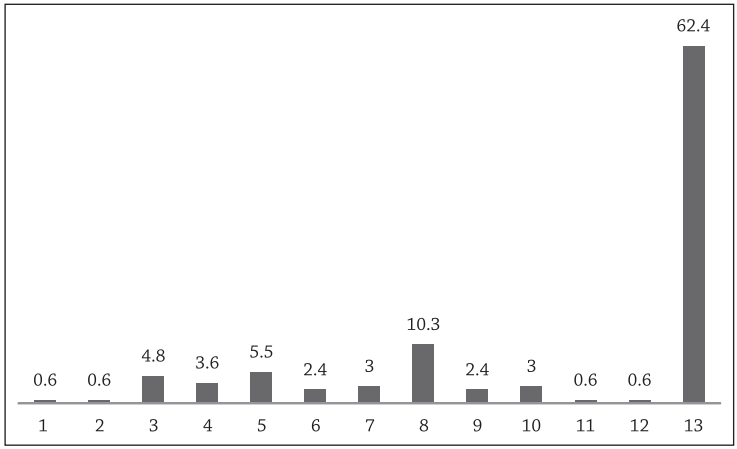


***Figure 1:*** *Percentage of language samples of the indicated duration (min).*

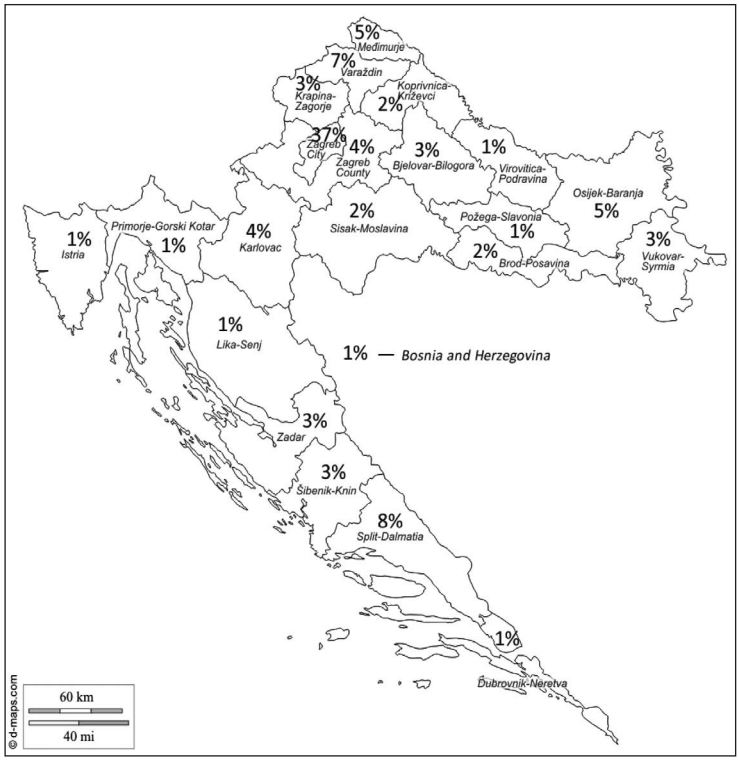


***Figure 2:*** *Percentage of language samples recorded in Croatia and Bosnia and Herzegovina. (Map from d-maps.com/carte.php?num_car=58521&lang=en)*

Sampling was performed in every Croatian county, mostly in the capital city, Zagreb, where more than a quarter of the country's population resides (Figure 2). Speakers from Zagreb come from all over the country and their spoken language is likely to reflect a mixture of varieties covering the entire country. Some sampling was also performed in parts of Bosnia and Herzegovina where the Croatian language is officially the dominant language and therefore the first language that the speakers acquire.

Sound files were transcribed and communicational features and events were annotated using special codes, such as hesitations, repetitions, breaks, and overlapping. Since transcripts cannot capture all the details of the original spoken language, as many transcripts as possible in HrAL are linked to the corresponding source audio files.

### 2.2. Participants

The participants were adults who speak Croatian as their mother tongue and first language. All the investigators were instructed to collect samples from speakers with no documented difficulties in language or in cognitive development or status. The original group consisted of 636 speakers, but 19 speakers withdrew during sampling or analysis, thus leaving 617 participants and 165 language transcripts. Some participants were investigators who participated in the conversations. The results of corpus analysis are presented below in two forms: one that includes the contributions of investigators (marked with '+'), and one that excludes them (marked with '-').

Transcripts were annotated to include the age and gender of the speakers, as well as the location of the conversation. A separate spreadsheet lists the speakers' origin, where they have spent most of their lives and their level of education. While age and gender data were available for all speakers (Table 1), the information about their origin was available for only 80% and the information about education for only 60%. These data are more complete for samples collected from 2014 onwards.

***Table 1:*** *Completeness of the data on the characteristics of speakers represented in HrAL**

| Sampling period | Speakers, n (+/-) | No. of speakers (+/-) for which data are available | | | |
|---|---|---|---|---|---|
| | | Age | Sex | Origin | Education |
| 2010-2012 | 69/53 | 69/53 | 69/53 | 47/36 | 18/2 |
| 2014-2015 | 287/214 | 287/214 | 287/214 | 236/180 | 146/113 |
| 2016 | 251/206 | 251/206 | 251/206 | 203/168 | 206/117 |
| All | 607/473 | 607/473 | 607/473 | 486/384 | 370/286 |

* Data are reported for all speakers including investigators (+) or excluding them (-).

*2.2.1. Participant gender*

Speakers represented in HrAL are predominantly female (+, 74%; -, 67%), and these proportions are higher than those of females in the most recent Croatian Census (2011; Figure 3), indicating a feminine bias in gender representation in HrAL.


***Figure 3:*** *Percentages of speakers represented in HrAL by gender.*

*2.2.2. Participant age*

A substantial percentage of speakers represented in HrAL were 30 years old or younger. This reflects the fact that the majority of investigators were in this age group, so many of the family members and friends that they recruited as speakers were also in this age group. The distribution of speakers roughly corresponds to that found in the most recent Croatian Census, though younger individuals are overrepresented, and some older age groups are underrepresented (Figure 4).

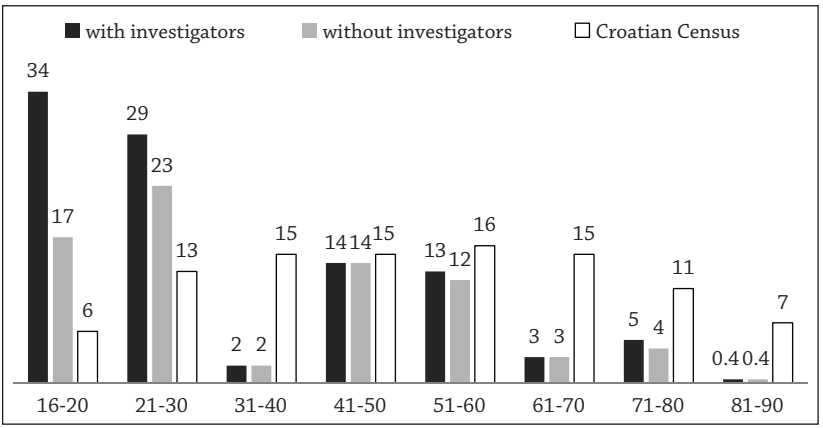


***Figure 4:*** *Percentages of speakers represented in HrAL by age group.*

However, the large number of speakers allows for selecting specific groups of participants, thus building subcorpora that are more representative, or that are more appropriate for specific research (e.g., young speakers, older population etc.).

Sampling was performed in various age groups of Croatian speakers. Future cycles of language sampling should aim to balance the proportions of each group so as to be more representative of the general population.

*2.2.3. Participant origin*

The county distribution of speakers represented in HrAL roughly corresponds to the distribution of inhabitants recorded in the most recent *Croatian Census* (Figure 5). However, the proportion of HrAL participants from the northern counties of Varaždin, Međimurje and Krapina-Zagorje is twice as high as their proportion in the census. Conversely, some counties such as Primorje-Gorski Kotar are underrepresented in the HrAL.

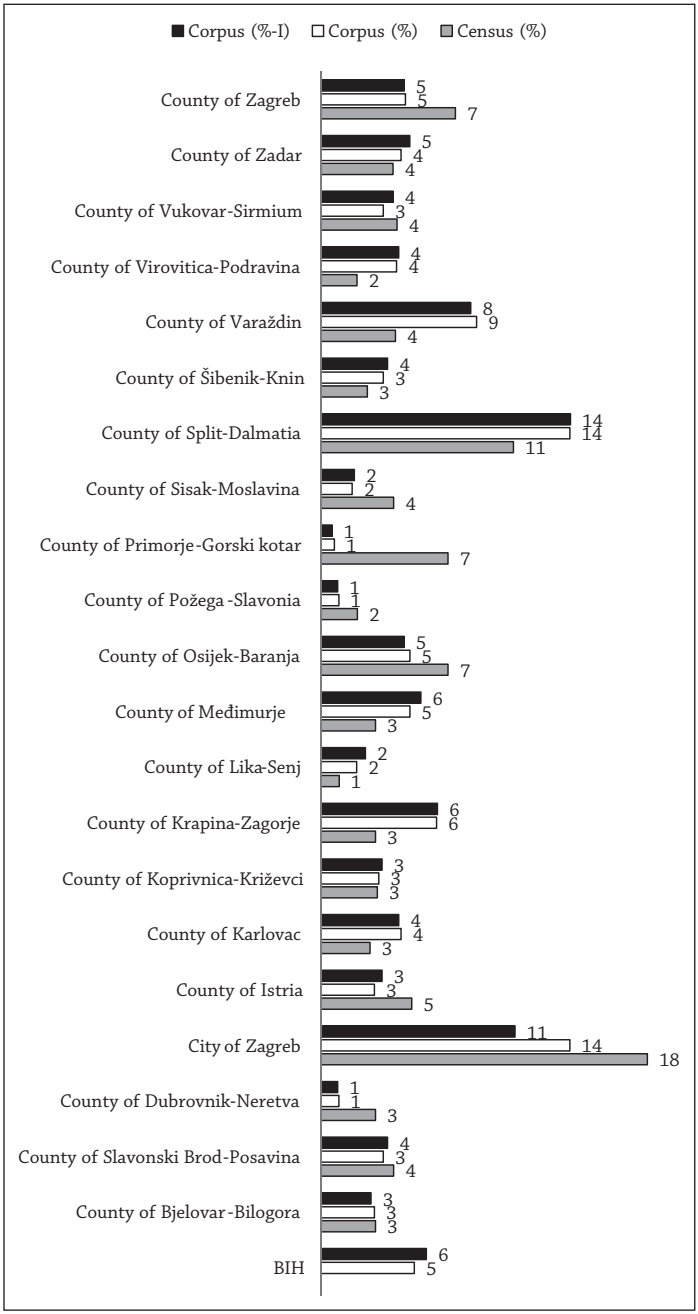


***Figure 5:*** *Distribution of speakers represented in HrAL by Croatian counties.*

### 2.3. Transcription, coding and media linkage

HrAL is available within *TalkBank*, a large database of spoken-language corpora covering different languages (https://talkbank.org, MacWhinney, 2007). *TalkBank* contains corpora from several subfields of communication, including first language acquisition, second language acquisition, conversation analysis and clinical group communication. HrAL is published in the subsection *Conversation Banks* under the heading *Conversation Analysis Bank* (direct link: https://talkbank.org/access/CABank/Croatian.html). The *TalkBank* allows download of individual corpora as well as access to browsable transcripts.

Transcripts were prepared in accordance with the standard *TalkBank* rules of contribution, using *TalkBank* support for transcription, editing and analysis. Data were transcribed, coded and segmented using the transcription format *Codes for Human Analysis of Transcripts* (CHAT) and the *Computerised Language Analysis* (CLAN) suite of programmes within the *TalkBank* toolkit (MacWhinney, 2007). CHAT codes capture a range of conversational structures as well as morphosyntactic and discourse features. First, samples were transcribed by the investigators who had previously received approximately 10 hours of basic training and who were supervised by experienced transcribers. Next, the speech streams were segmented into communication units (C-units, Loban, 1966) based on syntactic criteria. A C-unit is based on a T-unit, which is defined as the 'one main clause plus whatever subordinate clauses happen to be attached or embedded within it' (Hunt, 1966: 735). Thus, C-units include all the T-units, in addition to isolated phrases not accompanied by a verb. Such phrases typically appear in answers to questions (Crookes, 1990).

Each transcript was re-checked twice by experienced researchers, once for C-unit segmentation and again for coding. Each transcript was also verified by the CHECK programme of the CLAN suite.

Most transcripts were linked to their source audio file using the Transcriber function in the CLAN suite. This allows for both continuous media playback and utterance-level playback. Media files are available in the downloaded version of HrAL. The transcripts of samples from 2010-2012 are not linked to source audio files, while the transcripts from later samples are.

## 3. HrAL

The corpus consists of three folders, one for each sampling period. An accompanying spreadsheet contains information about participants. The following three types of files – transcripts, sound files and participant spreadsheet – are available.

### 3.1. Transcript files

Transcript files are in the .cha format. Files are named according to the year of sampling and the investigator who collected the sample. For example, the file named

2015_1.cha in Figure 6 is a sample collected in 2015 by investigator 1. In those cases where the same investigator collected several samples, the corresponding filenames contain an additional number: in Figure 6, the files 2015_4_1.cha and 2015_4_2.cha were recorded by investigator 4 at different times in 2015.

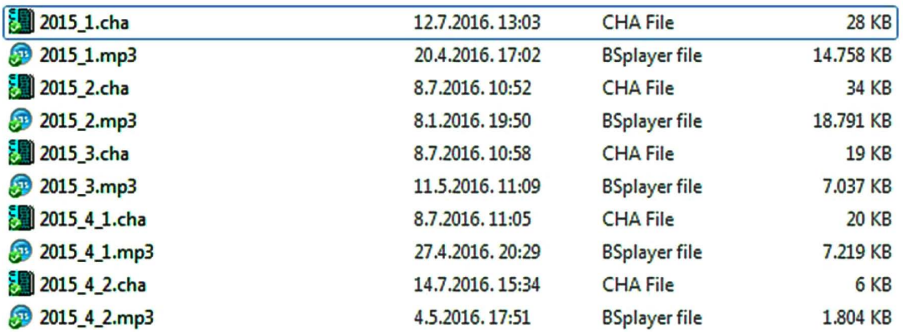


***Figure 6:*** *File structure in HrAL.*

Each transcript comprises a header, with information about the transcript and the speakers, and the transcription itself (Figure 7). Each participant is defined using a unique three-letter code which was assigned during anonymisation and which bears no similarity to the participant's real name. The same speaker may appear in multiple samples if the samples were transcribed by the same individual. Investigators who participated in conversations were also assigned a three-letter code in addition to the invariant code INV.

The header indicates, when available, age, gender, origin (county and city), educational level and role of the participants.

The roles were Target_Adult, Investigator and Target_Child (if the participant was younger than 16). The following codes were used for the educational level: OŠ, completed primary school (8 years); SSS, completed secondary school (3-4 years); STUD, current undergraduate students; VŠS, completed undergraduate or equivalent degree (2-4 years); VSS, completed master's or equivalent degree (undergraduate degree + 1 or 2 years); MDR, postgraduate education beyond 2 years (e.g., a PhD degree).

The header also includes the information about the length (@TimeDuration), location (@Location) and situation (@Situation) of the source recording. In many cases, additional comments (@Comments) may report difficulties encountered during transcription (e.g., some parts of the sound file were difficult to understand or the background noise was excessive) or comments on the conversation itself (e.g. excessive overlapping).

```
@Participants:    ALP Target_Adult, ALR Target_Adult, ALT Target_Adult, INV
                  ALS Investigator
@ID:    hrv|HrAL|ALP|48;|male|Međimurska_(Kotoriba)||Target_Adult|SSS||
@ID:    hrv|HrAL|ALR|71;|female|Međimurska_(Donji_Vidovec)||Target_Adult|OŠ||
@ID:    hrv|HrAL|ALT|21;|male|Međimurska_(Kotoriba)||Target_Adult|SSS||
@ID:    hrv|HrAL|INV|20;|female|Međimurska_(Kotoriba)||Investigator|STUD||
@Comment:         U pozadini se često čuje zvuk posuđa i micanja stolaca po podu.
                  Cijeli se razgovor odvija na dijalektu.
@Media:           2015_7_1, audio
@Time Duration:           00:09:00
@Date:            25-DEC-2015
@Location:        Kotoriba
@Situation:       Razgovor nakon obiteljskog ručka.
```

***Figure 7:*** *Example of a transcript header.*

The body of the transcript consists of rows, each of which represents one C-unit (Figure 8). Each row starts with the unique three-letter code of the participant. This coding makes it straightforward to exclude specific participants from analyses when necessary.

```
*AOE:   misliš?
*AOE:   pa plus dva je.
*AOE:   misliš da će smrznit?
*AOD:   +< pa ne znam.
*AOD:   po noći će bit minus ako je sad plus dva.
```

***Figure 8:*** *Example of a body of a transcript.*

### 3.2. Sound files

Sound files in mp3 format are located in the folders corresponding to the period when they were sampled. Each sound file carries the same name as its transcript. They can be played directly from the transcript using the options 'Mode' → 'Continuous skip play' (to play the whole sound file) or 'Mode' → 'Play bullet media' (to play individual utterances).

### 3.3. Participant spreadsheet

A Microsoft Excel spreadsheet contains information about each participant (Figure 9). Each row presents information for one participant in each transcript. For most participants, this information includes the location where he or she has spent most of his or her life. This can help in determining the individual's dialectal status. The

data in this spreadsheet can facilitate statistical analyses of the database, such as the average number of participants per transcript or the average number of transcripts per county. It can also facilitate the selection of subsets of transcripts or participants for specific research purposes.

| Transcript | County of recording | Place of recording | Participant | origin (County) | origin (place) | majority of life (County) | majority of life (Place) | Age | Gender | Education |
|---|---|---|---|---|---|---|---|---|---|---|
| 2014_20_3 | Grad Zagreb | Zagreb | ADL | Grad Zagreb | Zagreb | Grad Zagreb | Zagreb | 21 | m | STUD |
| 2014_20_3 | Grad Zagreb | Zagreb | ADR | Grad Zagreb | Zagreb | Grad Zagreb | Zagreb | 55 | f | VŠS |
| 2014_21_1 | Grad Zagreb | Zagreb | ADS | Splitsko-dalmatinska | Split | Splitsko-dalmatinska | Split | 20 | f | STUD |
| 2014_21_1 | Grad Zagreb | Zagreb | ADT | Splitsko-dalmatinska | Split | Splitsko-dalmatinska | Split | 55 | f | VSS |

***Figure 9:*** *An extract of participant information provided in the spreadsheet companion to HrAL.*

### 3.4. Corpus size

HrAL contains 279 885 tokens and 129 061 types when the investigators are included, or 228 273 tokens and 104 035 types when they are excluded. As intended, the investigators contributed relatively few words to the corpus (Table 2). Each speaker contributed an average of 500 tokens to the corpus (range, 1 to 2 101).

***Table 2:*** *Numbers of tokens and types contributed to HrAL per speaker.**

| | N | Min | Max | M | SD |
|---|---|---|---|---|---|
| Types (all) | 607 | 1 | 703 | 215 | 151 |
| Tokens (-) | 473 | 1 | 2101 | 495 | 429 |
| Types (-) | 473 | 1 | 703 | 226 | 156 |
| Tokens (+) | 134 | 1 | 2030 | 371 | 342 |
| Types (+) | 134 | 1 | 634 | 180 | 129 |
| Tokens (all) | 607 | 1 | 2101 | 466 | 414 |

* Data are reported for all speakers including investigators (+) or excluding them (-).

## 4. Data sharing

*TalkBank* provides eight levels of data sharing, ranging from full web access with no attempt at anonymisation to archiving-only functionality (http://talkbank.org/share/irb/options.html). To protect the participants in HrAL, which is especially important given the small communities in which many samples were recorded, the last names and exact addresses of participants have been withheld. The participants' first names and/or initials have been pseudonymised. Each participant is identified using a unique three-letter sequence.

The authors of HrAL are happy to share the corpus and the related data with the wider community in accordance with the basic rules of *TalkBank* (https://talkbank.org/share/rules.html). In particular, any published works derived from *TalkBank* corpora should cite the references listed in the manuals, as well as the present article in the case of HrAL. Any such works should also cite the original *TalkBank* publication (MacWhinney, 2007). HrAL is available under a *Creative Commons Attribution-ShareAlike 4.0 International Public License* (https://creativecommons.org/licenses/by-sa/4.0/legalcode).

## 5. Possible applications

Over the past two decades, spoken-language corpora have been used increasingly to gain more direct insights into human communication, since they reflect the actual use of language in everyday situations. For example, usage-based theory privileges spoken corpora because they can reveal how social interactions influence language (see Tomasello & Stahl, 2004; Lieven, 2010). On the contrary, spoken-language corpora are usually not of interest to formal linguists and are only sometimes of interest to functional linguists (see Wichmann, 2008). Although theoretical approaches differ when it comes to how valuable and objective they consider spoken-language corpora to be, these corpora remain the most comprehensive data source available for addressing the basic linguistic question of what characterises the language of ordinary speakers.

HrAL provides information about spoken grammar and lexicon, discourse skills, error production and productivity in general. Since HrAL contains key demographic data on the majority of participants, these linguistic variables can be analysed for different age, gender, or dialectal subgroups. Age-related subgroup analysis may support research into how the specificity of language use changes over time. Such work may provide evidence about language socialisation across the life span in collaborative and competitive settings (see Ochs, 1999). HrAL contains all regional varieties of Croatian, so it may be quite useful for sociolinguistic research, especially in the domain of dialectology.

HrAL can serve as a valuable sorce for spoken-language research. As such, it represents a complement to written language corpora based mainly on texts written by professional writers. Howewer, in order to obtain more detailed insight into all the aspects of the Croatian language, it is neccessary to establish sources of non-professional written

language. Such a corpus is currently being developed. Language samples have been collected, and are in the process of annotation (Kuvač Kraljević, Hržica & Kolografić Belić, in press). Spoken and written language are based on the same grammatical repertoire but they differ in how that grammar is implemented (Leech, 2000). Spoken language tends to employ simpler and less concrete syntactic constructions, since grammar plays a less prominent role in spoken than written communication. Comparing the written corpus with HrAL may shed light on simplification in Croatian, which in turn may be relevant to studies of simplification in other languages.

HrAL has clear potential for underpinning studies of synchronic language changes in Croatian. If the corpus can be developed longitudinally over a longer period, then it may serve as a powerful tool for studying diachronic changes as well.

**References**

Brown, R. (1973) *A First Language. The Early Stages,* Cambridge. MA: Harvard University Press.

Burnard, L. (2000) *Reference Guide for the British National Corpus (World Edition),* Available from http://www.natcorp.ox.ac.uk/docs/userManual/ [Accessed 2007-01-11].

Croatian Bureau of Statistics (2011) *Croatia – Population and Housing Census 2011*.

Crookes, G. (1990) The utterance and other basic units for second language discourse analysis, *Applied Linguistics*, 11, 183–199.

Fries Carpenter, C. (1952) *The Structure of English*. New York: Harcourt Brace.

Hunt, K. W. (1966) Recent measures in syntactic development. *Elementary English 43,* 732–739.

Kuvač Kraljević, J., Hržica, G. & Kolografić Belić, L. (in press) *Croatian Corpus of Non-professional Written Language*. University of Zagreb, Laboratory for Psycholinguistic Research, Zagreb.

Labov, W. (1972) *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Leech, G. (2000) Grammars of Spoken English: New Outcomes of Corpus-oriented Research. *Language Learning*, 50(4), 675–724.

Lieven, E. (2010) Input and first language acquisition: Evaluating the role of frequency. *Lingua*, 120(11), 2546–2556.

Loban, W. (1966) *Language Ability: Grades Seven, Eight, and Nine*. Washington, DC: Government Printing Office.

MacWhinney, B. (2007) The TalkBank Project. In: J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora: Synchronic Databases,* Volume 1. Houndmills: Palgrave-Macmillan, 163–180.

McCarthy, M. J. & O'Keeffe, A. (2008) *Corpora and the Study of Spoken Language*. In: A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook*. Volume 2. Berlin: Mouton de Gruyter, 1008–1024.

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

Ochs, E. (1999) Language socialization. In: A. Duranti & M. A. Mldan (Eds.) *Key terms in language and culture*. MA: Blackwell, 227–230.

O'Keeffe, A. & Farr, F. (2003) Using Language Corpora in Language Teacher Education: pedagogic, linguistic and cultural insights. *TESOL Quarterly*, 37(3), 389–418.

Schonell, F., Meddleton, I., Shaw, B., Routh, M., Popham, D., Gill, G., Mackrell, G. & Stephens, C. (1956) *A Study of the Oral Vocabulary of Adults*. Brisbane and London: University of Queensland Press/University of London Press.

Svartvik, J. (1990) *The London-Lund Corpus of Spoken English: Description and Research*. (Lund Studies in English 82.) Lund: Lund University Press.

Svartvik, J. & Quirk, R. (1980) *A Corpus of English Conversation*. Lund: Gleerup.

Tognini Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tomasello, M. & Stahl, D. (2004) Sampling children's spontaneous speech: how much is enough? *Journal of Child Language*, 31(1), 101–121.

Wichmann, A. (2008) Speech corpora and spoken corpora. In: A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook*. Volume 1. Berlin: Mouton de Gruyter, 187–207.

SUMMARY

Jelena Kuvač Kraljević, Gordana Hržica

HRVATSKI KORPUS GOVORNOG JEZIKA (HrAL)

Zanimanje za korpuse govornog jezika posljednja dva desetljeća raste, pri čemu nastaju i razvijaju se novi istovrsni korpusi koji omogućuju uvid u nove činjenice o govornom jeziku. Ova vrsta korpusa predstavlja najiscrpniji izvor podataka o jeziku prosječnoga govornika. Ti se korpusi temelje na spontanom i nestrukturiranom govorenju koje je određeno različitim stilovima, registrima i dijalektima.

Cilj je ovog rada predstaviti *Hrvatski korpus govornog jezika odraslih* (HrAL), njegovu strukturu i moguću primjenu u različitim lingvističkim granama. HrAL je oblikovan uzorkovanjem spontane konverzacije između 617 govornika iz svih hrvatskih županija i sadrži više od 250.000 pojavnica i više od 100.000 različnica. Podatci su prikupljani u tri vremenska razdoblja: od 2010. do 2011., od 2014. do 2015. te tijekom 2016. godine.

HrAL je danas dostupan u *TalkBank*-u, bazi korpusa govornih jezika prikupljenih u različitim jezicima (https://talkbank.org), i to u pododjeljku *Conversational analyses corpora* unutar *Conversational Bank*. Podatci su transkribirani, kodirani i segmentirani rabeći transkripcijske formate *Codes for Human Analysis of Transcripts* (CHAT) i *Computerised Language Analysis* (CLAN), iz niza programa *TalkBank*-a. Govorni nizovi segmentirani su na komunikacijske jedinice (C-jedinice) temeljene na sintaktičkom kriteriju. Većina je transkripata povezana sa svojim audiozapisom. *TalkBank* je javno dostupan, odnosno svi podatci pohranjeni u njemu mogu biti slobodno upotrijebljeni prema osnovnim pravilima *TalkBank*-a.

HrAL daje informacije o gramatici i leksikonu govornog jezika, diskursnim vještinama, proizvedenim pogreškama i produktivnosti općenito. Koristan je za sociolingvistička istraživanja kao i za istraživanja sinkronijskih jezičnih promjena u hrvatskom.

**Ključne riječi:** *Hrvatski korpus govornog jezika odraslih (HrAL); jezično uzorkovanje; korpus spontanoga govora*