DirKorp: A Croatian Corpus of Directive Speech Acts

Petra Bago*, Virna Karlić†

* Department of Information and Communication Sciences
† Department of South Slavic Languages and Literatures
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{pbago, vkarlic}@ffzg.hr

Abstract

In this paper we present recent developments on a new version (v2.0) of DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*), a Croatian corpus of directive speech acts developed for the purposes of pragmatic research. The corpus contains 800 elicited speech acts collected via an online questionnaire with role-playing tasks. Respondents were 100 Croatian speakers, all undergraduate or graduate students of the Faculty of Humanities and Social Sciences University of Zagreb. The corpus has been manually annotated on the speech act level, each speech act containing up to 12 features. It contains 12,676 tokens and 1,692 types. The corpus is encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, developed and maintained by the *Text Encoding Initiative Consortium* (TEI). We describe applied pragmatic annotation as well as the structure of the corpus.

1. Introduction

Corpus pragmatics is an interdisciplinary field of study that incorporates linguistic pragmatics and computer science, focusing on the development of natural language corpora in machine-readable form and their application for the purposes of studying pragmatics phenomena in written and spoken language. For a long time have linguists regarded a corpus approach to language incompatible with pragmatics (Romero-Trillo, 2008: 2). While the corpus approach to studying language implies processing authentic language material implementing quantitative research methods, pragmatic research is still predominantly of qualitative nature – based on the researcher's introspection, data obtained by elicitation methods or an analysis of authentic linguistic material of small size. The application of corpus analysis in the research of pragmatics phenomena represents a major turnaround in the development of pragmatics, primarily because it allows a systematic analysis of authentic language material of large size, and thus the detection of patterns of language use that "go below radar" through qualitative analyses (ibid.). In addition, it should be pointed out that the application of new technologies in linguistics, including pragmatics, did not only ensure, facilitate or accelerate numerous research processes, but opened the door to a new, different way of thinking about language (Leech, 1992).

The application of corpus methods on large pragmatic corpora allows one to systematically carry out empirically based pragmatic research (Bunt, 2017: 327). While the implementation of corpus research can result in minor adjustments to existing theories on the one hand, it can lead to a rethinking of pragmatics concepts and theoretical frameworks on the other hand, for example the development of the theory of dialogue acts (ibid.).

According to Rühlemann and Aijmer (2015), one of the major methodological problems that corpus pragmatics researchers encounter is the disproportionate relationship between pragmatic functions and language forms by which these functions are expressed. One form can perform multiple pragmatic functions in discourse, while one function can be expressed by different forms, which makes the process of querying a corpus according to the pragmatic function criterion considerably difficult. It is for this reason that corpus pragmatics researchers most often investigate conventional speech acts or functions performed by a limited number of language forms (Jucker, Scheier, and Hundt, 2009: 4). The aim of this paper is to present the first Croatian corpus of directive speech acts DirKorp, manually annotated for corpus pragmatic research.

The paper is structured as follows: Section 2 describes selected work related to pragmatic corpora, while the subsequent three section present the DirKorp corpus. Section 3 gives a description of the developed corpus, Section 4 describes 12 annotation features, and Section 5 presents the structure of the corpus encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2021). Finally, Section 6 contains conclusion and future work.

2. Related Work

The number of large corpora with systematically implemented pragmatic annotation is small so far. Due to a disproportionate relationship between pragmatic functions and language forms by which these functions are expressed, automatic corpus annotation does not produce satisfactory results. For this reason, only a small number of researchers have engaged in the creation of larger corpora of this sort. Generally, for the purposes of corpus pragmatic research, specialized corpora of smaller size are produced for individual research purposes. In addition, pragmatic research is sometimes carried out on corpora without pragmatic annotation.

An example of a corpus that does not contain pragmatic annotation, but which was used for pragmatic research is the Birmingham Blog Corpus[1] (Kehoe and Gee, 2007; Kehoe and Gee, 2012). In fact, this is a subcorpus of a larger set of corpora being developed at the department *Research and Development Unit for English Studies* at the Birmingham City University. It consists of blog posts and reader comments, sizing 500M words in English that were collected between 2000 and 2010. Automatic POS annotation was performed using the

[1] https://www.webcorp.org.uk/wcx/lse/corpora

Stanford Core NLP tools[2] and include lemma annotations and part-of-speech categories[3] based on the Universal Dependencies framework[4], while documents contain metadata of the publication date. Pragmatic research on speech acts was conducted on this corpus: For example, Lutzky and Kehoe (2017a; 2017b) used it to analyze apologies as speech acts that contain formulaic expressions, which facilitate its querying in a corpus when using available tools.

Similarly, we (Karlić and Bago, 2021) conducted research on the pragmatic functions and properties of imperatives using corpora without pragmatic annotation. We used hrWaC and srWaC (Ljubešić and Klubička, 2014), two large web corpora of Croatian and Serbian language with morphosyntactic annotation. For the purposes of the analysis, an additional pragmatic annotation of a representative sample of verbs in an imperative form was carried out manually. Other corpora of the Croatian spoken and written language with no pragmatic annotation have also been used as a resource for a corpus pragmatic research. For example, Hržica, Košutar, and Posavec (2021) used the Croatian Corpus of the Spoken Language of Adults (HrAL) (Kuvač Kraljević and Hržica, 2016) and the Croatian National Corpus of the written language (HNK) (Tadić, 1996) for the search and analysis of connectors and discourse markers.

According to Bunt (2017) the majority of corpora with pragmatic annotation contain labels on discourse relationships in written texts and on spoken dialogue acts. An example of such a larger corpus is Penn Discourse Treebank or PDTB[5] (Prasad, Webber, and Lee, 2018) that contains labels on discourse relations, i.e. discourse structure and its semantics. Discourse annotations were added to a subcorpus consisting of texts published in the newspaper *Wall Street Journal* sizing 1M tokens, included in a bigger corpus *Penn Treebank* (PTB). Bunt (2017) states that there are corpora of other languages developed for the purposes of studying the co-occurrence of discourse labels, such as Chinese, Czech, Dutch, German, Hindi and Turkish – emphasizing that these corpora are manually annotated and of modest sizes. Additionally, for each corpora a new schema was developed based on various theoretical starting points.

DialogBank[6] (Bunt et al., 2019) is one of a rare dialogue corpus annotated with an ISO 24617-2 standard. It contains already existing dialogue corpora annotated with various schemas. Four corpora are of English: HCRC Map Task (Anderson et al., 1991), Switchboard (Godfrey, Holliman, and McDaniel, 1992), TRAINS (Allen et al., 1995) and DBOX (Petukhova et al., 2014); and four of Dutch language: DIAMOND (Geertzen et al., 2004), OVIS[7], Dutch Map Task (Caspers, 2000) and Schiphol (Prüst, Minnen, and Beun, 1984). Dialogue act annotation involves segmenting a dialogue into defined grammatical units and augmenting each unit with one or more communicative function labels.

Another example of a corpus with a pragmatic annotation is the *Engineering Lecture Corpus*[8] (Alsop and Nesi, 2013; Alsop and Nesi, 2014) that contains 76 transcripts based on an hour-long video recordings of engineering lectures held in English on three universities. It is manually annotated for three pragmatic features: humor, storytelling and summary[9]. Each feature can be augmented with one of the attributes containing additional information that describes the feature in more detail. Further, the corpus contains labels regarding significant breaks, laughter, writing or drawing in the board, etc.

Finally, we present SPICE-Ireland corpus (*Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland*) (Kallena and Kirka, 2012), a part of a larger set of corpora ICE-Ireland (*International Corpus of English: Ireland Component*) containing pragmatic, discourse and prosodic features. The corpus contains various types of private and public, formal and informal dialogues and monologues of a length of about 2,000 words, sizing 625K words. It consists of spoken English. The pragmatic annotation of speech acts is based on Searle's classification (Searle, 1969; Searle, 1976): representatives, directives, commissives, expressives and declaratives.

To the best of our knowledge, there exist no publicly available corpora of spoken or written Croatian language with pragmatic annotation. So far, Croatian linguists mostly dealt with speech acts from a theoretical perspective, referring primarily to the Austin's and Searle's theory (cf. Pupovac, 1991; Ivanetić, 1995; Miščević, 2018; Palašić, 2020). However, in recent times, the number of research based on qualitative and quantitative analysis of small-sized authentic linguistic materials (from literary texts and advertisements to email messages and political discourse in Croatian and other languages) has been increasing (cf. e.g., Pišković, 2007; Matić, 2011; Franović and Šnajder, 2012; Šegić, 2019).

In the following sections we present a new version (v2.0) of DirKorp, the first Croatian corpus of directive speech acts.

3. Corpus Description

DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*) (Karlić and Bago, 2021) is a Croatian corpus of directive speech acts developed for the purposes of pragmatic research. The corpus contains 800 elicited speech acts collected via an online questionnaire with role-playing tasks applying the method of simulated communication that is implemented under pre-set conditions. This method is suitable for researching speech acts due to the ability to collect a great number of examples of speech acts of the equal propositional content and illocutionary purpose used in the same controlled situations. The questionnaire included eight closed-type role-playing tasks. These types of tasks imply recording the speaker's reactions (in this case in writing) to the stimulus without feedback. In each task, the participants are presented with one textually described hypothetical situation asking them to refer a directive speech act to

[2] https://stanfordnlp.github.io/CoreNLP/

[3] See more about the POS tagset used for the Birmingham Blog Corpus: https://www.webcorp.org.uk/wcx/lse/guide.

[4] https://universaldependencies.org/u/pos/index.html

[5] https://doi.org/10.35111/qebf-gk47

[6] https://dialogbank.uvt.nl/

[7] http://www.let.rug.nl/vannoord/Ovis/

[8] www.coventry.ac.uk/elc

[9] https://www.coventry.ac.uk/research/research-directories/current-projects/2015/engineering-lecture-corpus-elc/annotations-and-mark-ups/

their interlocutor. Their assignment was to imagine they were in the presented situation and to give a written statement they would use in the described situations. The presented situations are classified into two categories with regard to the relationship between the participants of the communication act: (1) situations involving interlocutors who are not in a familiar relationship; (2) situations involving interlocutors in a familiar relationship. Assignments of the two categories are organized into four pairs, asking respondents to share a speech act of similar propositional content: "I want you to return something that belongs to me" (for text of role-playing tasks see Example 1 when interlocutors have (a) an unfamiliar relationship and (b) a familiar relationship); "I want you to answer my inquiry"; "I want you to change something that bothers me"; "I want you to stop behaving inappropriately"[10].

Example 1

(a) Upravo si pojeo/la ručak u restoranu. Posluživao te stariji konobar koji se odnosio prema tebi ljubazno i profesionalno. Prilikom plaćanja računa konobar ti vraća 100 kuna manje nego što je trebao. Želiš da ti konobar vrati novac. Zamisli da se konobar nalazi pred tobom i napiši što bi mu točno rekao/la u danoj situaciji (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).

(Eng. *You just ate lunch at a restaurant. You were served by an elderly waiter who treated you kindly and professionally. When paying the bill, the waiter refunds you 100 kunas less than he should have. You want the waiter to give you your money back. Imagine the waiter was in front of you and write what exactly you would say to him in the given situation (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

(b) Posudio/la si knjigu najboljem prijatelju (ili prijateljici). Rekao ti je da će ti je uskoro vratiti, no nije održao riječ. Sjedite zajedno u kafiću, situacija je opuštena, razgovarate o svakodnevnim stvarima. Želiš mu dati do znanja da ti treba čim prije vratiti knjigu. Zamisli da se tvoj prijatelj nalazi pred tobom i napiši što bi mu točno rekao/la u danoj situaciji (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).

(Eng. *You lent a book to your best friend. (S)he told you (s)he'd give it back to you soon, but (s)he didn't keep her/his word. You are sitting together in a café, the situation is relaxed, you talk about everyday things. You want to let her/him know you need to get your book back as soon as possible. Imagine if your friend was in front of you and wrote what exactly you would say to her/him in the given situation (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

Respondents were 100 Croatian speakers, all undergraduate (63 %) or graduate students (37 %) of the Faculty of Humanities and Social Sciences University of Zagreb, ages between 18 to 33. Croatian is the mother tongue for the majority of the respondents (96 %). The questionnaire was carried out during December 2020 and January 2021. All respondents voluntarily participated in the study. The questionnaire was conducted anonymously, and the collected language material was used exclusively for scientific purposes.

The elicitation of language production by the role-playing method has its advantages and disadvantages. On the one hand, it enables the collection of a large number of speech acts with the same propositional content and illocutionary purpose. On the other hand, users of the corpus should keep in mind that the language material collected by this method does not reflect the features of actual language use. It rather shows what speakers think they would say and/or do in hypothetical situations.

DirKorp contains 12,676 tokens and 1,692 types[11]. Since it consists of 800 speech acts, it is a relatively small corpus. However, as the first Croatian corpus with detailed pragmatic annotation, DirKorp can serve as a useful resource for researching speech acts, politeness strategies and other related pragmatic phenomena in the Croatian language. In addition, we hope that it will contribute to the development of larger corpora of the Croatian language with pragmatic annotation, and that it will encourage a wider application of the corpus-pragmatic research method.

We have conducted corpus pragmatic analyses of the collected speech acts in order to investigate ways and means of expressing directives, and their pragmatic characteristics and functions. For example, we confirmed that indirect directives are more frequent than direct, especially among interlocutors who are not in a familiar relationship. Regarding (un)familiar relationship between interlocutors, we detected that explicit illocutionary force is more frequent in communication between interlocutors with a familiar relationship, while implicit illocutionary force is more frequent in communication between interlocutors with an unfamiliar relationship. Additionally, we have identified that imperative utterances are a more frequent type of direct directives than utterances with a directive performative verb in 1st person. For more such corpus pragmatic analyses see Karlić and Bago (2021).

4. Corpus Annotation

Collected language material has been manually annotated on the speech act level by two independent annotators with university graduate degrees in the field of philology. Annotators received oral and written instructions, including illustrative examples for all the features they had to annotate.

The categorization of speech acts and their formal and pragmatic properties was carried out according to the theory of speech acts by Austin (1962), Searle (1969; 1976) and their successors; the politeness theory of Brown and Levinson (1978), and the grammars of contemporary Croatian and Serbian languages (Silić and Pranjković, 2007; Piper et al., 2005). For more on individual

[10] Full texts of role-playing tasks are available in the corpus header.

[11] Respondents' answers contain utterances, but also text about what they would do in the given situation. At this moment, we have not analyzed average length of a response. Generally, we can only state that some speech acts contain only one utterance, while some contain more than one.

categories, see Karlić and Bago (2021). In the new version of DirKorp (v2.0), each speech act can contain up to 12 features. The first 8 features were part of the corpus version v1.0, while features 9-12 are newly added. For frequency distribution of all features see Karlić and Bago (2021).

(1) **Respondent ID** – This mandatory feature contains information on identification of the respondent uttering the speech act.

(2) **Familiarity / unfamiliarity** – This mandatory feature contains information on the category of the proposed situation in which the speech act was uttered. Four situations are labelled 'unfamiliar' (involving interlocutors who are not in a familiar relationship), while the other four situations are labelled 'familiar' (involving interlocutors who are in a familiar relationship).

(3) **Utterance type** – This mandatory feature contains information on the utterance type regarding its structural organization. It contains five labels: (a) an imperative utterance, (b) an assertive utterance (a statement), (c) an utterance in the form of a question, (d) an utterance in the form of an ellipsis, (e) a nonverbal signal, (f) a case of avoidance of executing a speech act (see Example 2).

Example 2
(a) E vrati mi onu knjigu koju sam ti posudio.
(Eng. *Hey, give me back that book I lent you.*)
(b) Oprostite, ali mislim da ste mi krivo vratili novce.
(Eng. *Excuse me, but I think you gave me my money back wrong.*)
(c) Možete li molim vas zatvoriti prozore?
(Eng. *Could you please close the windows?*)
(d) E, moja knjiga??
(Eng. *Hey, my book??*)
(e) [Samo bih zavrtjela očima da vide moje neodobravanje, ali ne bih ništa rekla.]
(Eng. *[I'd just roll my eyes so that they see my disapproval, but I wouldn't say anything.]*
(f) [Ne bih ništa rekao.]
(Eng. *[I wouldn't say anything.]*)

(4) **Directive performative verb in 1st person** – This optional feature contains information on the representation of a directive performative verb in 1st person as part of the speech act, only for assertive utterances and utterances in the form of a question. It contains two labels: (a) yes and (b) no (see Example 3).

Example 3
(a) Oprostite, molim da odete na kraj reda.
(Eng. Excuse me*, I am imploring you to go to the end of the line.*)
(b) Gospođo, morate na kraj reda stati.
(Eng. *Madam, you must move to the end of the line.*)

(5) **Illocutionary force** – The optional feature contains information on explicitness or implicitness of the illocutionary force of a speech act. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question and in the form of an ellipsis). It contains two labels: (a) explicit and (b) implicit (see Example 4).

Example 4
(a) Daj mi donesi više onu knjigu, treba mi!
(Eng. *Bring me that book already, I need it!*)
(b) Kaj je s onom knjigom koju sam ti posudio?
(Eng. *What happened to that book I lent you?*)

(6) **Propositional content** – This optional feature contains information on explicitness or implicitness of the propositional content of a speech act. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question and in the form of an ellipsis). It contains two labels: (a) explicit and (b) implicit (see Example 5).

Example 5
(a) Gledaj na cestu, pusti mobitel.
(Eng. *Look at the road, leave the cell phone.*)
(b) Ti hoćeš da poginemo?
(Eng. *You want us to die?*)

(7) **T/V form** – This optional feature contains information on how the respondent addressed the interlocutor, using an informal (T-form) or a formal *you* (V-form). It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question and in the form of an ellipsis). It contains three labels: (a) T-form, (b) V-form and (c) impossible to determine (see Example 6).

Example 6
(a) Oprosti, dao si mi manje novca
(Eng. *Sorry, you$_{T\text{-}form}$ gave me less change.*)
(b) Oprostite, mislim da ste mi ipak još dužni100 kuna.
(Eng. *Excuse me, I think you$_{V\text{-}form}$ still owe me 100 kunas.*)
(c) Hmm... još 100 kuna, zar ne?
(Eng. *Hmm… another 100 kunas, right?*)

(8) **Exhortative** – This optional feature contains information on the representation of an exhortative as part of the speech act. It contains two labels: (a) yes and (b) no (see Example 7).

Example 7
(a) Daj mi više vrati knjigu, treba mi za knjižnicu.
(Eng. *Bring me back my book already, I need it for the library.*)
(b) Jel se sjećaš one knjige koju sam ti posudila? Potrebna mi je. Možeš li mi ju donijeti sutra na faks?
(Eng. *Do you remember that book I lent you? I need it. Could you bring it tomorrow to uni?*)

(9) **Request** – This optional feature contains information on whether the speech act includes a lexical marker of request. It contains two labels: (a) yes and (b) no (see Example 8).

Example 8
(a) E da, jel bi mi mogao/la vratiti knjigu, molim te?
(Eng. *Oh yeah, could you bring the book back, please?*)
(b) Zaboravio si mi vratiti knjigu, jel se možeš idući put sjetiti?

(Eng. *You forgot to bring me back the book, can you remember next time?*)

(10) **Apology** – This optional feature contains information on whether the speech act includes a lexical marker of apology. It contains two labels: (a) yes and (b) no (see Example 9).

Example 9
(a) Oprostite, ovdje fali još 100 kuna
(Eng. *Excuse me, 100 kunas is missing here.*)
(b) Možete li molim vas pritvoriti prozore, hladno mi je?
(Eng. *Could you please close the windows, I'm cold?*)

(11) **Gratitude** – This optional feature contains information on whether the speech act includes a lexical marker of gratitude. It contains two labels: (a) yes and (b) no (see Example 10).

Example 10
(a) Molim te mi samo javi da znam zbog organizacije hoćeš li doći. Hvala ti!
(Eng. *Please just let me know whether you're coming so that I know because of the organization. Thank you!*)
(b) Heej, jel dolaziš večeras na druženje? Moram znati zbog organizacije. xoxo
(Eng. *Heeey, are you coming tonight to hang out? I need to know because of the organization. xoxo*)

(12) **Honorific title** – This optional feature contains information on whether the speech act includes an honorific title. It contains two labels: (a) yes and (b) no (see Example 11).

Example 11
(a) Gospođo, kraj reda je dolje
(Eng. *Madam, the end of the line is back there.*)
(b) Oprostite, tamo je kraj reda!
(Eng. *Excuse me, the end of the line is there!*)

5. Corpus Format

DirKorp is encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, developed and maintained by the Text Encoding Initiative Consortium (TEI) (TEI Consortium, 2021). The TEI document is comprised of a header and the body of the corpus. The content of the elements and attributes are in Croatian. Metadata of the corpus is given in the header including: bibliographic information; the editorial practice; a structured taxonomy describing categories used for each of the 12 pragmatic features in the annotation process (see Figure 1 for an example), including full text of the eight situations on the questionnaire; a list of questionnaire participants with information on their age, gender, undergraduate or graduate level of study, enrollment in a philological/non-philological/combined study program and mother tongue (see Figure 2 for an example); and a list of revisions of the DirKorp versions. The body of the corpus is composed of one division containing utterances with pragmatic features (see Figure 3 for an example).

DirKorp is available for download under the CC BY-SA 4.0 license from GitHub in TEI format (https://github.com/pbago/DirKorp).

```
<taxonomy xml:id="tiVi">
  <category xml:id="ti">
    <catDesc>Govorni čin sadržava
obraćanje na ti (atribut se odnosi
na tipove iskaza koji uključuju
verbalna sredstva [imperativni,
tvrdnja, upitni,
eliptični]).</catDesc>
  </category>
  <category xml:id="vi">
    <catDesc>Govorni čin sadržava
obraćanje na Vi (atribut se odnosi
na tipove iskaza koji uključuju
verbalna sredstva [imperativni,
tvrdnja, upitni,
eliptični]).</catDesc>
  </category>
  <category xml:id="persNeodredivo">
    <catDesc>Nije moguće odrediti
sadržava li govorni čin obraćanje na
ti ili Vi (atribut se odnosi na
tipove iskaza koji uključuju
verbalna sredstva [imperativni,
tvrdnja, upitni,
eliptični]).</catDesc>
  </category>
</taxonomy>
```

Figure 1: An example of a pragmatic feature description – how the respondent addressed the interlocutor (V-form, T-form or impossible to determine).

```
<person xml:id="I001" sex="F">
  <p>ispitanik/ispitanica, 20
godina, spol Ž, preddiplomski studij
Filozofskog fakulteta, nefilološko
usmjerenje, materinji jezik
hrvatski</p>
</person>
```

Figure 2: An example of participant information.

```
<u who="#I001" ana="#NEFAM1 #tvdrnja
#dpg1N #isI #psI #vi #adhorativN
#molbaN #isprikaY #zahvalaN
#honorifikN">Ispričavam se, pardon,
fali još sto kuna. Oprostite.</u>
```

Figure 3: An example of an utterance containing all 12 pragmatic features.

6. Conclusion and Future Work

We have presented DirKorp, the first Croatian corpus of directive speech acts, containing 800 elicited speech acts collected via an online questionnaire with role-playing tasks, specifically developed for pragmatic research studies. Respondents were 100 Croatian speakers, all students of the Faculty of Humanities and

Social Sciences University of Zagreb. The corpus has been manually annotated on the level of a speech act, each speech act containing up to 12 features. It contains 12,676 tokens and 1,692 types. The corpus is available for download under the CC BY-SA 4.0 license from GitHub in TEI format.

Further work is planned on the corpus, which includes an evaluation of the developed scheme for annotating directive speech acts, annotation at the levels smaller than a speech act, as well as augmentation with additional features such as information on grammatical mood used in a speech act, information on representation of modal verb in 2nd person as part of a speech act, and information on various politeness strategies applied in a speech act.

7. Acknowledgements

This paper is generously co-financed by the institutional project of the Faculty of Humanities and Social Sciences "South Slavic languages in use: pragmatic analyses" (principle researcher Virna Karlić). We wish to thank all our annotators.

8. References

James F. Allen, Lenhart K. Schubert, Geoge Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesio, and David R. Traum. 1995. The TRAINS Project: A Case Study in Building a Conversational Planning Agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Sian Alsop and Hilary Nesi. 2013. Annotating a Corpus of Spoken English: The Engineering Lecture Corpus (ELC). In: *Proceedings of GSCP 2012: Speech and Corpora*, pages 58–62. Firenze University Press, Florence.

Sian Alsop and Hilary Nesi. 2014. The Pragmatic Annotation of a Corpus of Academic Lectures. In: *The International Conference on Language Resources andEvaluation 2014 Proceedings*, pages 1560–1563. European Language Resources Association, Reykjavik.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus, *Language and Speech*, 34(4):351–366.

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Harry Bunt. 2017. Computational Pragmatics. In: *Oxford Handbook of Pragmatics*, pages 326-345. Oxford University Press, New York.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. 2019. The DialogBank: Dialogues with Interoperable Annotations. In: *Language Resources and Evaluation*, 53(2):213–249.

Johanneke Caspers. 2000. Melodic Characteristics of Backchannelsin Dutch Map Task Dialogues. In: *Proceedings, 6th International Conference on SpokenLanguage Processing*, pages 611–614. China Military Friendship Publish , Beijing, https://www.isca-speech.org/archive/icslp_2000/.

Tin Franović and Jan Šnajder. 2012. Speech Act Based Classification of Email Messages in Croatian Language. In: *Proceedings of the Eighth Language Technologies Conference*, pages 69-72. Information Society, Ljubljana.

Jeroen Geertzen, Yann Girard, Roser Morante, Ielka Van der Sluis, Hans Van Dam, Barbara Suijkerbuijk, Rintse Van der Werf, Harry Bunt. 2004. The DIAMOND Project. In: *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004),* Barcelona.

John Godfrey, Edward Holliman, and Jande McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1*, pages 517–520. IEEE Computer Society, San Francisco.

Gordana Hržica, Sara Košutar, and Kristina Posavec. 2021. Konektori i druge diskursne oznake u pisanome i spontanome govorenom jeziku. *Fluminensia: časopis za filološka istraživanja*, 33(1):25–52.

Nada Ivanetić. 1995. *Govorni činovi*. Zagreb: FF-press, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. (eds.). 2009. *Corpora: Pragmatics and Discourse*. Rodopi, Amsterdam.

Jeffrey L. Kallen and John M. Kirk. 2012. *SPICE-Ireland: A User's Guide*. https://pure.qub.ac.uk/en/publications/spice-ireland-a-users-guide.

Virna Karlić and Petra Bago. *(Računalna) pragmatika: temeljni pojmovi i korpusnopragmatičke analize*. FF Press , Zagreb, 2021. https://openbooks.ffzg.unizg.hr/index.php/Ffpress/catalog/book/125.

Andrew Kehoe and Matt Gee. 2007. New Corpora from the Web: Making Web Text More 'Text-Like'. In: *Studies in Variation, Contacts and Change in English 2*. https://varieng.helsinki.fi/series/volumes/02/kehoe_gee/.

Andrew Kehoe and Matt Gee. 2012. Reader Comments as an Aboutness Indicator in Online Texts: Introducing the Birmingham Blog Corpus. In: *Studies in Variation, Contacts and Change in English 12*. https://varieng.helsinki.fi/series/volumes/12/kehoe_gee/.

Jelena Kuvač Kraljević and Gordana Hržica. 2016. Croatian Adult Spoken Language Corpus (HrAL). *Fluminensia: časopis za filološka istraživanja*, 28(2):87–102.

Geoffrey N. Leech. 1992. Corpora and Theories of Linguistic Performance. In: *Directions in Corpus Linguistics*, pages 105–122. De Gruyter, Berlin.

Ursula Lutzky and Andrew Kehoe. 2016. Your Blog is (the) Shit: A Corpus Linguistic Approach to the Identification of Swearing in Computer Mediated Communication. *International Journal of Corpus Linguistics*, 21(2): 165–191.

Ursula Lutzky and Andrew Kehoe. 2017a. 'I Apologize for My Poor Blogging': Searching for Apologies in the

Birmingham Blog Corpus. *Corpus Pragmatics*, 1(1):37–56.

Ursula Lutzky and Andrew Kehoe. 2017b. 'Oops, I Didn't Mean to Be so Flippant'. A Corpus Pragmatic Analysis of Apologies in Blog Data. *Journal of Pragmatics*, 116:27–36.

Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr}WaC-Web Corpora of Bosnian, Croatian and Serbian. In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Association for Computational Linguistics , Gothenburg, https://aclanthology.org/W14-0405.pdf.

Daniela Matić. 2011. *Govorni činovi u političkome diskursu*. PhD thesis. Faculty of Humanities and Social Sciences, Zagreb.

Nenad Miščević. 2018. *Rođenje pragmatike*. Orion Art, Beograd.

Nikolina Palašić. 2020. *Pragmalingvistika – lingvistički pravac ili petlja?* Hrvatska sveučilišna naklada, Zagreb.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz, Steffen Liersch, and Anna Schmidt. 2014. The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 252–258. European Language Resources Association, Reykjavik.

Predrag Piper et al. 2005 = Предраг Пипер, Ивана Антонић, Бранислава Ружић, Срето Танасић, Људмила Поповић, Бранко Тошовић. 2005. *Синтакса савременог српског језика*. Проста реченица, Београд: Институт за српски језик САНУ, Београдска књига, Матица српска.

Tatjana Pišković. 2007. Dramski diskurs između pragmalingvistike i feminističke lingvistike. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje,* 33(1):325–341.

Olumide Popoola. 2017. A Dictionary, a Survey and a Corpus Walked into a Courtroom...: An Evaluation of Resources for Adjudicating Meaning in Trademark Disputes. In: *The 9th International Corpus Linguistics Conference*. Birmingham: Birmingham University.https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2017/general/paper134.pdf.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse Annotation in the PDTB: The NextGeneration. In: *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97. Santa Fe: Association for Computational Linguistics. https://aclanthology.org/W18-4710.pdf.

Hub Prüst, Guido Minnen, and Robbert-Jan Beun. 1984. Transcriptie dialooogesperiment juni/juli 1984, *IPORapport 481*. Institute for Perception Research, Eindhoven University of Technology, Eindhoven.

Milorad Pupovac. 1990. *Jezik i djelovanje*. Biblioteka časopisa Pitanja, Zagreb.

Jesús Romero-Trillo (ed.). 2008. *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. De Gruyter, Berlin.

Christoph Rühlemann and Karin Aijmer. 2015. Introduction. Corpus pragmatics: laying the foundations. In: *Corpus pragmatics*, pages 1-28.

John R. Searle. 1969. *Speech Acts*. Cambridge University Press, Cambridge.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5:1–23.

Josip Silić and Ivo Pranjković. 2007. *Gramatika hrvatskoga jezika za gimnazije i visoka učilista*. Školska knjiga , Zagreb.

Tea Šegić. 2019. Tata kupi mi auto und Nivea Milk weil es nichts Besseres für die Hautpflege gibt. *Filologija*, 73:103–116.

Marko Tadić. 1996. Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika,* 41-42:603–611.

TEI Consortium (ed.). 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.