DirKorp: A Croatian Corpus of Directive Speech Acts (v3.0)

Petra BAGO
Faculty of Humanities and Social Sciences, University of Zagreb

Virna KARLIĆ
Faculty of Humanities and Social Sciences, University of Zagreb

In this paper, we present recent developments on a new version (v3.0) of DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*), the first 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, a 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 such acts of equal propositional content and illocutionary purpose used in the same controlled 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. The respondents were 100 Croatian speakers, all undergraduate (63%) or graduate students (37%) 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 14 features: (1) respondent ID, (2) familiarity/unfamiliarity, (3) utterance type, (4) directive performative verb in 1st person, (5) illocutionary force, (6) propositional content, (7) T/V form, (8) exhortative, (9) lexical marker of request, (10) lexical marker of apology, (11) lexical marker of gratitude, (12)

Bago, P., Karlić, V.: DirKorp: A Croatian Corpus of Directive Speech Acts (v3.0). Slovenščina 2.0, 11(1): 189–217.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.189-217

honorific title, (13) grammatical mood, and (14) modal verb in 2nd person. 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). DirKorp is available for download under the CC BY-SA 4.0 license from GitHub in TEI format. We describe applied pragmatic annotation as well as the structure of the corpus.

Keywords: corpus pragmatics, directive speech acts, DirKorp, Croatian language

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, linguists have regarded a corpus approach to language as incompatible with pragmatics (Romero-Trillo, 2008, p. 2). While the corpus approach to studying language implies processing authentic language material by implementing quantitative research methods, pragmatic research is still predominantly of a 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 language material of large size, and thus the detection of patterns of language use that "fly below the 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 to large pragmatic corpora allows one to systematically carry out empirically based pragmatic research (Bunt, 2017, p. 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 pragmatic concepts and theoretical frameworks on the other, such as the development of the theory of dialogue acts (ibid.).

According to Rühlemann and Aijmer (2015), one of the major methodological problems that corpus pragmatic 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 rather difficult. It is for this reason that corpus pragmatic researchers most often investigate conventional speech acts or functions performed by a limited number of language forms (Jucker et al., 2009, p. 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 corpus pragmatic research, Section 3 explores the definition, classification, and research methods of directive speech acts, while the subsequent three sections present the DirKorp corpus. Section 4 gives a description of the developed corpus, Section 5 describes 14 annotation features, and Section 6 presents the structure of the corpus encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2021). Finally, Section 7 contains the conclusions and some directions for future work.

This is a follow-up paper from the conference "Language Technologies and Digital Humanities" held in Ljubljana, Slovenia on 15th–16th September 2022, where we presented DirKorp v2.0. Here we present a new published version of the DirKorp (v3.0) with two additional annotation layers, as well as a new section clarifying the definition, classification, and research methods of directive speech acts (Section 3).

2 Related work

The number of large corpora with systematically implemented pragmatic annotation remains relatively small. Due to a disproportionate relationship between pragmatic functions and the language forms by which these functions are expressed, automatic corpus annotation does not produce satisfactory results. For this reason, only a few researchers have engaged in creating 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 was used for pragmatic research is the Birmingham Blog Corpus[1] (Kehoe and Gee, 2007; 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, and includes some 500 million words in English that were collected between 2000 and 2010. Automatic POS annotation was performed using the Stanford Core NLP tools[2] and included lemma annotations and part-of-speech categories[3] based on the Universal Dependencies framework,[4] while the documents contain metadata of the publication date. Pragmatic research on speech acts has been conducted on this corpus. For example, Lutzky and Kehoe (2017a; 2017b) used it to analyse apologies as speech acts that contain formulaic expressions, which facilitate their querying in a corpus when using the 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 the Croatian and Serbian languages 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 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.

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

[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

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 the Penn Discourse Treebank or PDTB[5] (Prasad et al., 2018) which 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* with a total of around 1 million 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 size. Additionally, for each corpus 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, namely HCRC Map Task (Anderson et al., 1991), Switchboard (Godfrey et al., 1992), TRAINS (Allen et al., 1995) and DBOX (Petukhova et al., 2014), and four of Dutch – DIAMOND (Geertzen et al., 2004), OVIS[7], Dutch Map Task (Caspers, 2000) and Schiphol (Prüst et al., 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; 2014) which contains 76 transcripts based on hour-long video recordings of engineering lectures held in English at three universities. It is manually annotated for three pragmatic features: humour, 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 on the board, etc.

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

[6] https://dialogbank.lsv.uni-saarland.de/

[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/

Finally, we present the SPICE-Ireland corpus (*Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland*) (Kallen and Kirk, 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, with a size of some 625,000 words. It consists of spoken English. The pragmatic annotation of speech acts is based on Searle's classification (Searle, 1969; 1976): representatives, directives, commissives, expressives, and declaratives.

When it comes to corpus research of speech acts, researchers have two options: (1) to analyse examples from existing corpora of authentic linguistic material, or (2) to analyse examples of elicited linguistic material. In the second case, different types of data completion tests are usually applied, and based on the obtained results smaller custom-made corpora are created for the needs of individual research (and therefore not publicly available). This method is most often used in cross-linguistic, contrastive research, but it is also used in the study of individual languages (e.g., Barron, 2008; Trosborg, 1995). For an overview of pragmatic research of speech acts (including directives) on elicited linguistic material, see, for example, Wojtaszek (2008).

To the best of our knowledge, there exist no publicly available corpora of spoken or written Croatian with pragmatic annotation. So far, Croatian linguists have mostly dealt with speech acts from a theoretical perspective, referring primarily to the Austin's and Searle's theory (cf. Pupovac, 1990; Ivanetić, 1995; Miščević, 2018; Palašić, 2020). However, in recent years the number of research projects based on the 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).

3 Directive speech acts: definition, classification, and research methods

During verbal communication, speakers express their thoughts in the form of utterances, through which they convey information, express their emotions and attitudes, or try to modify the addressee's behaviour (Capone, 2009, p. 1015).

Speech acts are utterances with specific properties and communicative functions (ibid.):

> A speech act (...) is not merely the expression of a thought. It is the vocalization of a certain representation of the world (external or internal) aimed at making official the display of an intention to change a state of things and at changing things by the public display of that intention.

Therefore, speech acts can be briefly defined as "actions performed via utterances" (Yule, 2002, p. 47). According to Searle (1975), there are five types of speech acts: (1) representatives – statements that can be evaluated as true or false; (2) directives, which speakers use to influence the addressee's wishes and actions; (3) commissives, through which speakers commit to perform some action in the future; (4) expressives, which speakers use to express their feelings or attitudes; and (5) declaratives or institutionalized declarations that formally change the state of affairs in extralinguistic reality (cf. Karlić and Bago, 2021).

Directive speech acts, or directives, are a type of speech act by which speakers express their "(...) desire/wish for the addressee to do something. (...) In using a directive, the speaker intends to elicit some future course of action on the part of the addressee, thus making the world match the words via the addressee" (Huang, 2009, p. 1004).

Directives differ with respect to their illocutionary force. The illocutionary force of a directive depends on how binding it is for the addressee. If the speaker insists on its realization, the illocutionary force of the directive is strong – and vice versa. According to this criterion, directive speech acts are classified into orders, commands, requests, pleads, incentives, advice, etc. (cf. Piper et al., 2005, p. 1021; Karlić and Bago, 2021, p. 37).

Directive speech acts can be direct or indirect. Direct directives contain an explicit directiveness marker – an imperative (*Close the window*) or a performative verb in the first person of the present tense (*I ask you to close the window*). Directiveness can be expressed implicitly, through assertions without a performative verb (*You should close the window*), interrogative utterances (*Can you close the window?*), or elliptical utterances (*Um... the window...*). Just like illocutionary force, the propositional content of directive speech acts can also be expressed explicitly and implicitly (*It is cold here* [implicature: *Close the window*]) (cf. Huang, 2009, p. 1005; Karlić and Bago, 2021, p. 39).

According to Brown and Levinson (1987, p. 65–66), directives represent a typical example of face-threatening acts. For this reason, when using them, speakers often apply various politeness strategies that mitigate their illocutionary force (e.g., implicatures and lexical or grammatical modifiers of illocutionary force).

The foundations of speech act theory were laid by the philosophers John Austin and John Searle in works published in the 1960s and 1970s. Since then, numerous studies of speech acts have been conducted. In the beginning, they were non-empirical, based on the researcher's intuition. In recent years, however, the number of empirical studies of speech acts has grown significantly. Jucker (2009) distinguishes three types of data collection methods for the needs of empirical research of speech acts – field, laboratory, and armchair (Flöck and Geluykens, 2015, p. 10):

> While armchair approaches investigate participants' intuitions and attitudes about language use, field and laboratory approaches aim at studying actual language use. They differ, however, in the way language data are produced. While in laboratory approaches, language use is elicited by researchers (by employing role-plays or administering discourse completion tasks), field data are defined by the absence of such elicitation techniques. Field methods are therefore observational in nature, i.e., they require an authentic communicative intent by participants to produce language.

Each of the mentioned methods has its advantages and disadvantages. For the purposes of creating the DirKorp corpus, we applied the laboratory method of eliciting language production by role-playing. The main advantage of this method is that it gives "full variable control to the researcher (...), and can generate large amounts of data; however, participants use language without their own intrinsic communicative intent in fictional scenarios" (Flöck and Geluykens, 2015, p. 11). This method allowed us to collect a large amount of mutually comparable directives with the same propositional content and produced in the same controlled circumstances.

In the following sections, we present a new version (v3.0) of DirKorp, the first Croatian corpus of directive speech acts.

4 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 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 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 (i.e., interlocutors who are not close and are not equal in terms of power relations, and communicate in more or less in/formal situations); (2) situations involving interlocutors in a familiar relationship (i.e., interlocutors in a close and equal relationship who communicate in more or less in/formal situations). 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 this role-playing task pair see Example 1 when interlocutors have (a) an unfamiliar relationship (label "NEFAM1") and (b) a familiar relationship (label "FAM1")); "I want you to answer my inquiry" (for text of this role-playing task pair see Example 2 when interlocutors have (a) an unfamiliar relationship (label "NEFAM2") and (b) a familiar relationship (label "FAM2")); "I want you to change something that bothers me" (for text of this role-playing task pair see Example 3 when interlocutors have (a) an unfamiliar relationship (label "NEFAM3") and (b) a familiar relationship (label "FAM3")); "I want you to stop behaving inappropriately" (for text of this role-playing task pair see Example 4 when interlocutors have (a) an unfamiliar relationship (label "NEFAM4") and (b) a familiar relationship (label "FAM4"))[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 your friend was in front of you and write 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).*)

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

Example 2

(a) Poslao/la si e-mail profesoru s upitom možeš li pohađati njegov izborni kolegij i hitno trebaš njegov odgovor i potvrdu u mailu. Međutim, profesor ne odgovara već tjedan dana, a rok za upis završava sutradan. Želiš ponovno zatražiti njegovu povratnu informaciju. Napiši kratak e-mail profesoru kakav bi mu uputio/la u navedenoj situaciji.
(Eng. *You sent an email to the professor asking if you can attend his elective course and urgently need his response and confirmation in the email. However, the professor has not responded for a week, and the admission deadline ends the next day. You want to ask for his feedback again. Write a short email to the professor as you would in this situation.*)

(b) Poslao/la si poruku (WhatsApp, Viber, Messenger) najboljem prijatelju (ili prijateljici) s pozivom na druženje sljedeće večeri. On je vidio poruku, ali nije odgovorio do sutradan. Želiš da ti odgovori čim prije kako bi mogao/la isplanirati ostatak dana. Napiši kratku poruku prijatelju kakvu bi mu uputio/la u navedenoj situaciji.
(Eng. *You sent a message (WhatsApp, Viber, Messenger) to your best friend with an invitation to hang out the next night. (S)he saw the message but did not respond until the next day. You want her/him to reply as soon as possible so you can plan the rest of the day. Write a short message to your friend as you would in this situation.*)

Example 3

(a) Voziš se u taksiju. Prozori su otvoreni i želiš da ih taksist zatvori jer ti je hladno. Zamisli da se nalaziš u navedenoj situaciji i napiši što bi točno rekao/la taksistu (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You're riding in a cab. The windows are open, and you want the taxi driver to close them because you're cold. Imagine that you are in this situation and write down what exactly you would say to the taxi driver (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

(b) Voziš se u autu na suvozačkom mjestu. Vozač je tvoj najbolji prijatelj (ili prijateljica). Budući da vozi prebrzo i gleda u mobitel, ne osjećaš se ugodno i želiš da uspori. Zamisli da se nalaziš u danoj situaciji i napiši što bi mu točno rekao/la (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).

(Eng. *You are riding in the car in the passenger seat. The driver is your best friend. Because (s)he's driving too fast and looking at her/his cell phone, you don't feel comfortable and want her/him to slow down. Imagine that you are in a given situation and write what exactly you would say to her/him (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

Example 4

(a) Nalaziš se u dućanu i čekaš u redu pred blagajnom. Velika je gužva. Ispred tebe se u red ugura gospođa srednje dobi. Ljudi u redu iza tebe negoduju jednako kao i ti. Želiš da gospođa stane na kraj reda. Zamisli da se nalaziš u danoj situaciji i napiši što bi točno rekao/la gospođi (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You're in the store waiting in line at the cash register. It's crowded. A middle-aged lady squeezes in front of you. The people in line behind you are just as resentful as you are. You want the lady to stand at the end of the line. Imagine that you are in a given situation and write down what exactly you would say to the lady (do not recount but formulate the statement as if you were directly addressing the interlocutor).*)

(b) Slušaš predavanje na fakultetu. Sjediš pored dvoje kolega s kojima si inače vrlo blizak/bliska. U jednom trenutku oni počinju glasno razgovarati i smijati se. Njihov razgovor ti smeta jer ne možeš pratiti predavanje, a i nastavnik pogledava u vašem smjeru. Želiš da prestanu. Zamisli da se nalaziš u danoj situaciji i napiši što bi im točno rekao/la (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You're listening to a lecture in college. You're sitting next to two colleagues with whom you are otherwise very close. At some point, they start talking loudly and laughing. Their conversation bothers you because you can't follow the lecture, and the lecturer looks in your direction. You want them to stop. Imagine that you are in a given situation and write down exactly what you would say to them (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

The respondents were 100 Croatian speakers, all undergraduate (63%) or graduate students (37%) of the Faculty of Humanities and

Social Sciences (University of Zagreb), aged between 18 to 33, with Croatian being the native language for the majority (96%). The questionnaire was administered in December 2020 and January 2021. Before completing the questionnaire, all the respondents were informed of the purposes of the study as well as what data would be collected. All the respondents voluntarily participated in the study and were made aware that they could withdraw from it at any time. By choosing to participate in the study, respondents gave informed consent for their data to be processed for the stated research purposes. The questionnaire was administered anonymously via an online survey, and the language material collected was used exclusively for research 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, but instead 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 compared to some of the corpora with pragmatic annotation presented in Section 3. However, as the first Croatian corpus with detailed pragmatic annotation, DirKorp can serve as a useful resource for researching the characteristics of speech acts on a formal and content level, the application of politeness strategies in communication in different situations, and the properties of other grammatical-pragmatic and lexical-pragmatic phenomena in the Croatian language that are annotated in the corpus. In addition, we believe that DirKorp can serve as a complement to research on speech acts that are conducted on authentic language materials and as a starting point for conducting contrastive research on the characteristics and use of speech acts in other languages. In addition, we hope that it will contribute to the development of larger corpora of

[11] Respondents' answers contain utterances but also text about what they would do in the given situation. At this moment, the corpus contains no annotation of utterances of speech acts, and therefore we cannot analyse the average length of a response. Generally, we can only state that some speech acts contain only one utterance, while some contain more than one.

the Croatian language with pragmatic annotation, and that such work will encourage a wider application of the corpus-pragmatic research method.

In Karlić and Bago (2021), we have conducted corpus pragmatic analyses of the collected speech acts 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 ones, especially among interlocutors who are not in a familiar relationship. Regarding a(n) (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).

5 Corpus annotation

The collected language material has been manually annotated on the speech act level by two independent annotators[12] 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.

Basic categorization of speech acts (directive; direct and indirect; explicit and implicit) and their formal and pragmatic properties (i.e., performative verbs) was carried out according to the theory of speech acts by Austin (1962), Searle (1969; 1976), and their successors. The features and components of speech acts related to the phenomenon

[12] When comparing the annotations of two annotators, in all categories, disagreements were found in at most 1.2% of examples and were mostly the result of accidental mistakes by one of the annotators. Once such mistakes were corrected, a consensus was reached among the annotators. The only category in which the disagreement was higher (2.5%) was the category “Illocutionary force”. In most cases, these were examples with generalized conversational implicature (one annotator marked speech acts with this type of implicature as explicit, and another as implicit). Based on the instruction to label speech acts with all types of implicature as implicit, a consensus was reached among the annotators for this category as well.

of politeness (familiarity, use of T/V forms, and certain lexical modifiers of the illocutionary force of speech acts) are taken according to the politeness theory of Brown and Levinson (1987), while the grammatical characteristics (utterance type, grammatical mood, modal verbs) of speech acts are categorized according to the grammatical descriptions of contemporary Croatian and Serbian languages (Silić and Pranjković, 2007; Piper et al., 2005). For more on individual categories, see Karlić and Bago (2021). In the new version of DirKorp (v3.0), each speech act can contain up to 14 features. The first eight features were part of the corpus version v1.0, features nine to 12 were part of v2.0, while features 13 and 14 were newly added. Appendix A contains the frequency distribution of features two to 14. For a more detailed frequency distribution of all features see Karlić and Bago (2021).

(1) **Respondent ID** – This mandatory feature contains information on the 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 six 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 a predicate ellipsis[13], (e) a nonverbal signal, (f) a case of avoidance of executing a speech act (see Example 5).

[13] Utterances in the form of a predicate ellipsis were singled out as a separate category due to: (1) the absence of a verb (and potentially other components of the sentence structure) and therefore the default indirectness and implicitness of the speech act, which makes them incomparable to other utterances in the corpus; (2) impossibility to determine the type of utterance for all examples due to their elliptical structure.

Example 5

(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.][14]
(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 6).

Example 6

(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 the 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 7).

[14] Descriptions of non-verbal situations can be found in Example 5 (e) and (f). All other examples contain actual utterances. DirKorp v3.0 does not contain annotations of utterances. Therefore, it is currently not possible to filter the speech acts with regard to actual utterances or descriptions of non-verbal situations.

Example 7

(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 the 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 8).

Example 8

(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 9).

Example 9

(a) Oprosti, dao si mi manje novca
(Eng. *Sorry*$_{\textit{T-form}}$*, you*$_{\textit{T-form}}$ *gave me less change.*)

(b) Oprostite, mislim da ste mi ipak još dužni 100 kuna.
(Eng. *Excuse*$_{\textit{V-form}}$ *me, I think you*$_{\textit{V-form}}$ *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 (a lexical

mean used to express encouragement, i.e., incentive particles). 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) yes and (b) no (see Example 10).

Example 10

(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 (e.g., "please"). 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) yes and (b) no (see Example 11).

Example 11

(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 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) yes and (b) no (see Example 12).

Example 12

(a) Oprostite, ovdje fali još 100 kuna
(Eng. *Excuse me, 100 kunas are 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 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) yes and (b) no (see Example 13).

Example 13

(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 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) yes and (b) no (see Example 14).

Example 14

(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!*)

(13) **Grammatical mood** – This optional feature contains information on grammatical mood used in a speech act. It is only applied to indirect speech acts (assertive utterances and utterances in the form of a question) since it is understood that direct imperative

speech acts contain verbs in the imperative mood. Accordingly, this feature contains two labels: (a) indicative mood and (b) conditional mood (see Example 15).

Example 15

(a) Oprostite, ali ovo nije kraj reda.
(Eng. *Excuse me, but this is not the end of the line*.)

(b) Oprostite jel bi mogli zatvorit prozore? Malo mi je hladno.
(Eng. *Excuse me, could you close the windows? I'm a little cold.*)

(14) **Modal verb in 2nd person** – This optional feature contains information on the representation of modal verb in 2nd person as part of a speech act. It is only applied to indirect speech acts (an assertive utterance and an utterance in the form of a question). It contains two labels: (a) yes and (b) no (see Example 16).

Example 16

(a) Oprostite, mislim da je došlo do pogreške, trebate mi vratiti još 100 kuna.
(Eng. *Sorry, I think there was a mistake, you have to return another 100 kunas*.)

(b) Malo je hladno ovdje, možemo možda zatvoriti prozor?
(Eng. *It's a little cold in here, can we possibly close the window?*)

6 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. The 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 14 pragmatic features in the annotation process (see Figure 1 for an example), including the 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, enrolment in a philological/non-philological/combined study program and native language (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, annotation feature 7 from Section 5).

```
<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 #tvrdnja #dpg1N #isI #psI #vi
#adhorativN #molbaN #isprikaY #zahvalaN #honorifikN #gnI
#mg2N”>Ispričavam se, pardon, fali još sto kuna. Oprostite.</u>
```

Figure 3: An example of a speech act containing all 14 pragmatic features.

7 Conclusion and future work

In this article we have presented DirKorp v3.0, 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. The respondents were 100 Croatian speakers, all students of the Faculty of Humanities and Social Science (University of Zagreb). The corpus has been manually annotated on the level of a speech act, with each speech act containing up to 14 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 schema for annotating directive speech acts (e.g., test-retest reliability on a sample of data to evaluate stability and consistency of the schema, domain experts reviewing the schema to determine if it adequately captures the relevant aspects of the data, reviewing the adequacy of encoding choices regarding attributes and its values), annotation at the levels smaller than a speech act, as well as augmentation with additional features such as information on various politeness strategies applied in a speech act.

Acknowledgments

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

References

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., ..., & Traum, D. R. (1995). The TRAINS Project: A Case Study in Building a Conversational Planning Agent. *Journal of Experimental & Theoretical Artificial Intelligence, 7*(1),7–48.

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

Alsop, S., & Nesi, H. (2014). The Pragmatic Annotation of a Corpus of Academic Lectures. In *The International Conference on Language Resources and Evaluation 2014 Proceedings* (pp. 1560–1563). Reykjavik: European Language Resources Association.

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., ..., & Weinert, R. (1991). The HCRC Map Task Corpus, *Language and Speech*, *34*(4), 351–366.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford: Clarendon Press.

Barron, A. (2008). The structure of requests in Irish English and English English. In K. P. Schneider & A. Barron (Eds.), *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages* (pp. 35–68). John Benjamins Publishing Company.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Bunt, H. (2017). Computational Pragmatics. In *Oxford Handbook of Pragmatics* (pp. 326–345). Oxford University Press, New York.

Bunt, H., Petukhova, V., Malchanau, A., Fang, A. & Wijnhoven, K. (2019). The DialogBank: Dialogues with Interoperable Annotations. In *Language Resources and Evaluation*, *53*(2), 213–249.

Capone, A. (2009). Speech Acts, Classification and Definition. In *Concise Encyclopedia of Pragmatics* (pp. 1015–1017). Oxford: Elsevier.

Caspers, J. (2000). Melodic Characteristics of Backchannels in Dutch Map Task Dialogues. In *Proceedings, 6th International Conference on Spoken Language Processing* (pp. 611–614). Beijing: China Military Friendship Publish,. Retrieved from https://www.isca-speech.org/archive/icslp_2000/

Flöck, I., & Geluykens, R. (2015). Speech Acts in Corpus Pragmatics: A Quantitative Contrastive Study of Directives in Spontaneous and Elicited Discourse. In *Yearbook of Corpus Linguistics and Pragmatics* (pp. 7–37). Springer International Publishing.

Franović, T., & Šnajder, J. (2012). Speech Act Based Classification of Email Messages in Croatian Language. In *Proceedings of the Eighth Language Technologies Conference* (pp. 69–72). Ljubljana: Information Society.

Geertzen, J., Girard, Y., Morante, R., Van der Sluis, J., Van Dam, H., Suijkerbuijk, B., Van der Werf, R., & Bunt, H. (2004). The DIAMOND Project. In: *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004),* Barcelona.

Godfrey, J., Holliman, E. & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *IEEE International*

Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 517–520). San Francisco: IEEE Computer Society.

Hržica, G., Košutar, S., & Posavec, K. (2021). Konektori i druge diskursne oznake u pisanome i spontanome govorenom jeziku. *Fluminensia: časopis za filološka istraživanja*, *33*(1), 25–52.

Huang, Y. (2009). Speech Acts. In *Concise Encyclopedia of Pragmatics* (pp. 1000–1009). Oxford: Elsevier.

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

Jucker, A. H. (2009). Speech Act Research between Armchair, Field and Laboratory: The Case of Compliments. *Journal of Pragmatics*, *41*, 1611–1635.

Jucker, A. H., Schreier, D., & Hundt, M. (Eds.). (2009). *Corpora: Pragmatics and Discourse*. Rodopi, Amsterdam.

Kallen, J. L., & Kirk, J. M. (2012). *SPICE-Ireland: A User's Guide*. Retrieved from https://pure.qub.ac.uk/en/publications/spice-ireland-a-users-guide

Karlić, V., & Bago, P. (2021). *(Računalna) pragmatika: temeljni pojmovi i korpusnopragmatičke analize*. Zagreb: FF Press. Retrieved from https://openbooks.ffzg.unizg.hr/index.php/Ffpress/catalog/book/125.

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

Kehoe, A., & Gee, M. (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*. Retrieved from https://varieng.helsinki.fi/series/volumes/12/kehoe_gee/

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

Leech, G. N. (1992). Corpora and Theories of Linguistic Performance. In *Directions in Corpus Linguistics* (pp. 105–122). De Gruyter, Berlin.

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

Lutzky, U., & Kehoe, A. (2017a). I Apologize for My Poor Blogging: Searching for Apologies in the Birmingham Blog Corpus. *Corpus Pragmatics*, *1*(1), 37–56.

Lutzky, U., & Kehoe, A. (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.

Matić, D. (2011). *Govorni činovi u političkome diskursu*. PhD thesis. Zagreb: Faculty of Humanities and Social Sciences.

Miščević, N. (2018). *Rođenje pragmatike*. Orion Art, Beograd.

Palašić, N. (2020). *Pragmalingvistika – lingvistički pravac ili petlja?* Zagreb: Hrvatska sveučilišna naklada.

Petukhova, V., Gropp, M., Klakow, D., Eigner, G., Topf, M., Srb, S., Motlicek, P., ... Potard, ..., & Schmidt, A. (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)* (pp. 252–258). European Language Resources Association, Reykjavik.

Piper, P. 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.

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

Prasad, R., Webber, B., & Lee, A. (2018). Discourse Annotation in the PDTB: The NextGeneration. In: *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (pp. 87–97). Santa Fe: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-4710.pdf

Prüst, H., Minnen, G. & Beun, R. (1984). Transcriptie dialooogesperiment juni/juli 1984, *IPORapport 481*. Eindhoven: Institute for Perception Research, Eindhoven University of Technology.

Pupovac, M. (1990). *Jezik i djelovanje*. Zagreb: Biblioteka časopisa Pitanja.

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

Rühlemann, C., & Aijmer, K. (2015). Introduction. Corpus pragmatics: laying the foundations. In: *Corpus pragmatics* (pp. 1–28).

Searle, J. R. (1969). *Speech Acts*. Cambridge University Press, Cambridge.

Searle, J. R. (1975). A Taxonomy of Speech Acts. In: *Minnesota Studies in the Philosophy of Science* (Vol. 9, pp. 344–369). University of Minnesota Press.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society, 5*, 1–23.

Silić, S. & Pranjković, I. (2007). *Gramatika hrvatskoga jezika za gimnazije i visoka učilista*. Zagreb: Školska knjiga.

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

Tadić, M. (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.

Trosborg, A. (1995). *Interlanguage Pragmatics: Requests, Complaints, and Apologies*. Berlin; New York: Mouton de Gruyter.

Wojtaszek, A. (2016). Thirty years of Discourse Completion Test in Contrastive Pragmatics research, *Linguistica Silesiana, 37*, 161–173.

Yule, G. (2002). *Pragmatics*. Oxford, New York: Oxford University Press.

DirKorp: hrvaški korpus direktivnih govornih dejanj (v3.0)

V prispevku predstavljamo razvoj nove različice (v3.0) korpusa DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*), prvega hrvaškega korpusa direktivnih govornih dejanj, ki je bil izdelan za namene raziskav pragmatike. Korpus vsebuje 800 govornih dejanj, ki so bila zbrana s spletnim vprašalnikom z nalogami igranja vlog – gre za metodo stimulirane komunikacije, ki poteka pod vnaprej določenimi pogoji. Metoda je primerna za raziskovanje govornih dejanj, saj lahko na ta način zberemo veliko število primerov z enako propozicijsko vsebino in ilokucijskim namenom, ki so uporabljeni v enaki kontrolirani situaciji. Predstavljene situacije razdelimo v dve kategoriji glede na odnos med udeleženci komunikacijskega dejanja: (1) situacije, ki vključujejo sogovorce, ki niso v sorodstvenem razmerju; (2) situacije z govorci v sorodstvenem razmerju. Naloge v obeh kategorijah so razdeljene v štiri pare, od sodelujočih pa zahtevajo, da pripišejo govorno dejanje s podobno propozicijsko vsebino. V vprašalniku je sodelovalo 100 govorcev hrvaščine; vsi so bili dodiplomski (63 %) ali podiplomski študenti (37 %) Fakultete za humanistiko in družbene vede (Univerza v Zagrebu). Korpus je bil ročno označen na ravni govornih dejanj, vsako dejanje pa vsebuje do 14 značilnosti: (1) ID sodelujočega, (2) sorodstveno/nesorodstveno razmerje, (3) tip izjave, (4) direktivni performativni glagol v prvi osebi, (5) ilokucijska sila, (6) propozicijska vsebina, (7) tikanje/vikanje, (8) prepričevalnost, (9) leksikalni označevalec za prošnjo, (10) leksikalni označevalec za opravičilo, (12) naziv spoštovanja, (13) slovnični naklon, (14) modalni glagol v drugi osebi. Korpus vsebuje 12.676 pojavnic in 1.692 različnic, enkodiran pa je v skladu s smernicami *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, ki jih razvija in vzdržuje konzorcij *Text*

Encoding Initiative Consortium (TEI). DirKorp je v formatu TEI na voljo za prenos pod licenco CC BY-SA 4.0 na platformi GitHub. V prispevku opišemo označevanje in strukturo korpusa.

Ključne besede: korpusna pragmatika, direktivna govorna dejanja, DirKorp, hrvaški jezik

Appendix A: Frequency distribution of annotated features 2-14

		Utterance type	∑	Directive performative verb in 1st person		Illocutionary force		Propositional content		T/V form		
				Yes	No	Explicit	Implicit	Explicit	Implicit	T-form	V-form	to determine
A	NEFAM1	Imperative	2	N/A	N/A	2	0	0	2	0	2	0
		Assertive	88	3	85	3	85	11	77	2	86	0
		Question	10	0	10	0	10	0	10	0	9	1
		Ellipsis	0	N/A	N/A	0	0	0	0	0	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
B	FAM1	Imperative	22	N/A	N/A	22	0	21	1	22	0	0
		Assertive	15	1	14	1	14	7	8	15	0	0
		Question	60	0	60	0	60	33	27	60	0	0
		Ellipsis	3	N/A	N/A	0	3	2	1	2	0	1
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
C	NEFAM2	Imperative	0	N/A	N/A	0	0	0	0	0	0	0
		Assertive	87	39	48	39	48	56	31	0	87	0
		Question	13	0	13	0	13	12	1	0	87	0
		Ellipsis	0	N/A	N/A	0	0	0	0	0	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D	FAM2	Imperative	40	N/A	N/A	40	0	38	2	40	0	0
		Assertive	8	2	6	2	6	4	4	8	0	0
		Question	46	0	45	0	46	5	41	45	0	1
		Ellipsis	3	N/A	N/A	0	3	1	2	1	0	2
		Nonverbal signal	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
E	NEFAM3	Imperative	2	N/A	N/A	2	0	2	0	0	2	0
		Assertive	1	0	1	0	1	1	0	0	1	0
		Question	96	1	95	1	95	96	0	0	95	1
		Ellipsis	0	N/A	N/A	0	0	0	0	0	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
F	FAM3	Imperative	86	N/A	N/A	86	0	80	6	86	0	0
		Assertive	4	0	4	0	4	1	3	4	0	0
		Question	9	0	9	0	9	5	4	9	0	0
		Ellipsis	1	N/A	N/A	0	1	0	1	1	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
H	NEFAM4	Imperative	19	N/A	N/A	19	0	19	0	0	19	0
		Assertive	55	9	46	9	46	12	43	0	55	0
		Question	12	0	12	0	12	10	2	0	11	1
		Ellipsis	1	N/A	N/A	0	1	1	0	0	1	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	13	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
G	FAM4	Imperative	43	N/A	N/A	43	0	40	3	1	0	42
		Assertive	1	0	1	0	1	0	1	0	0	1
		Question	12	0	12	0	12	12	0	0	0	12
		Ellipsis	37	N/A	N/A	0	37	33	4	1	1	35
		Nonverbal signal	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Avoidance	5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Exhortative		Request		Apology		Gratitude		Honorific title		Grammatical mood		Modal verb in 2nd person	
Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Indicative	Conditional	Yes	No
1	1	2	0	1	1	0	2	0	2	N/A	N/A	N/A	N/A
0	88	3	85	88	0	0	88	4	84	83	5	11	77
0	10	2	8	8	2	0	10	1	9	8	2	9	1
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
17	5	7	15	0	22	0	22	0	22	N/A	N/A	N/A	N/A
1	14	1	14	0	15	0	15	0	15	8	7	2	13
2	58	7	53	3	57	0	60	0	60	55	5	28	32
0	3	0	3	0	3	0	3	0	3	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
0	87	44	43	10	77	40	47	66	21	54	33	1	86
0	13	6	7	2	11	6	7	10	3	12	1	11	2
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
14	26	13	27	0	40	2	38	0	40	N/A	N/A	N/A	N/A
1	7	2	6	0	8	1	8	0	8	6	2	1	7
0	46	2	44	0	46	0	46	0	46	45	1	13	33
0	3	0	3	1	2	1	2	0	3	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
0	2	0	2	1	1	1	1	0	1	N/A	N/A	N/A	N/A
0	1	0	1	0	1	1	0	0	1	1	0	0	1
0	96	42	54	63	33	4	92	3	93	71	25	84	12
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
59	27	16	70	0	86	0	86	0	86	N/A	N/A	N/A	N/A
1	3	0	4	0	4	0	4	0	4	2	2	0	4
0	9	2	7	0	9	0	9	0	9	8	1	7	2
1	0	1	0	0	1	0	1	0	1	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	16	11	8	4	15	0	19	14	5	N/A	N/A	N/A	N/A
0	55	10	45	27	28	0	55	35	20	52	3	4	51
0	12	3	9	7	5	0	12	5	7	11	1	9	3
0	1	1	0	0	1	0	1	1	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
27	16	11	32	1	42	0	43	0	43	N/A	N/A	N/A	N/A
0	1	0	1	0	1	0	1	0	1	1	0	0	1
1	11	6	6	1	11	1	11	0	12	12	0	12	0
22	15	2	35	0	37	0	37	0	37	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A