A Computational Lexicon of Ukrainian Discourse Connectives

Tatjana Scheffler[1], Veronika Solopova[2], Olha Zolotarenko[3] and Mariia Razno[4]

[1]*Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany*
[2]*Freie Universität Berlin, Berlin, Germany*
[3]*Universität Potsdam, Potsdam, Germany*
[4]*Kharkiv Polytechnic University, Kharkiv, Ukraine*

Abstract

We introduce a new lexicon of discourse connectives for the Ukrainian language. Discourse connectives like 'because', 'therefore' are grammatical elements which link clauses and sentences semantically and play a crucial role in discourse structure. They have shown to be useful for many tasks in natural language processing from argumentation mining to authorship analysis.

We introduce a semi-automatic method for inventorizing discourse connectives in underresourced languages, by leveraging existing lexicons from other languages. As a result, we provide the first computer-readable lexicon of 129 Ukrainian discourse connectives. We provide syntactic as well as semantic information for these items. Finally, we carry out a small pilot study using the lexicon for discourse level corpus annotation, and report on the distribution of connectives in Ukrainian in two different types of media.

Keywords

discourse connectives, Ukrainian language, lexicon, coherence relations, speech vs. writing

1. Introduction

With applications ranging from automatic coherence analysis [1, 2], authorship analysis [3], and essay scoring [4], to argumentation mining [5], detection of discourse connectives and understanding of their nature is an important area of current research, and a valuable contribution linguistics continues to bring into natural language processing (NLP). Discourse connectives are words or phrases ('but', 'in contrast') that indicate semantic relations between larger text chunks such as clauses. Standing "at the dawn" of Ukrainian NLP, we argue that the apprehension of the discourse structure gives us a new and deeper look into Ukrainian language specifics, opening doors to these many fields currently being newly investigated

In this work, we synthesize the international experience on discourse structure analysis and discourse level annotation practices. We, then, introduce our computational lexicon of Ukrainian discourse connectives, and describe the process behind its creation. Our approach,

ICTERI-2021, Vol II: Workshops, September 28 – October 2, 2021, Kherson, Ukraine

tatjana.scheffler@rub.de (T. Scheffler); veronika.solopova@fu-berlin.de (V. Solopova);
olha.a.zolotarenko@gmail.com (O. Zolotarenko); mari.razno@gmail.com (M. Razno)
https://tscheffler.github.io/ (T. Scheffler)
0000-0001-7498-6202 (T. Scheffler)

using semi-automatic procedures and publicly available multilingual resources, can serve as example for quickly creating discourse lexicons for other languages which lack those resources. Finally, we perform a case-study annotating a part of the GRAK corpus [6] with connectives and their senses from our lexicon, as a pilot study analyzing statistical differences of spoken and written discourse structure of the Ukrainian language.

2. Related Work

2.1. NLP Resources for Ukrainian

We are not aware of any formalized resources of Ukrainian connectives. Indeed, computational resources of any kind for the Ukrainian language are still quite scarce. There is already some existing work on creating corpora of Ukrainian, the most notable of which is the Brown Corpus of Ukrainian Language [7], and more recently UA-GEC (Grammatical Error Correction and Fluency Corpus for the Ukrainian Language) [8]. In addition, the volunteer project Lang-uk [9] is working on various NLP taggers and corpora creation. In terms of NLP tools, several tokenizers have been adapted specifically for Ukrainian[1]. To our knowledge, none of the existing projects specifically address discourse level information.

2.2. Discourse Level Resources in Other Languages

Discourse level annotation and processing beyond the sentence is an important domain in natural language processing [10]. As applications become more sophisticated, it is increasingly important to incorporate information from the discourse context. One relevant dimension is discourse structure, i.e. the hierarchical structure of a text, created by semantic relations between different sentences or paragraphs. It is known that discourse relations are frequently marked overtly by certain words and phrases, called discourse connectives [11, 12]. Note that discourse connectives are a heterogeneous class of markers: words of many different parts of speech, as well as larger phrases, can fill the role of discourse connective in a given language. The detection of discourse connectives is therefore an important first step in discourse parsing [13].

To enable discourse parsing, as well as the linguistic analysis of discourse structure, human and computer readable discourse connective lexicons have been constructed for several languages, including German [14, 15], English [16], Italian [17], and others. Stede, Scheffler and Mendes [18] present the multilingual online database Connective-lex[2], which provides standardized access to all available, inter-operable discourse connective lexicons in 10 languages. Prior to the work reported in this paper, the only Slavic language available on Connective-lex was Czech [19]. Among other, more application-oriented, uses, discourse connective lexicons, especially when they are interfaced with each other across languages, allow for multilingual and cross-linguistic comparison of discourse structure (see e.g., [20, 21]).

[1] https://github.com/lang-uk/tokenize-uk, https://github.com/brown-uk/nlp_uk, https://stanfordnlp.github.io/stanza/available_models.html

[2] http://connective-lex.info/

In this paper, we report on the construction of the Ukrainian discourse connective lexicon UK-Dimlex, which has been already integrated into Connective-lex.

2.3. Discourse Connectives and Text Genres

Some previous work has investigated the distribution of discourse connectives in different types of text. For example, it has been demonstrated that there are significant differences in both discourse structure and its marking between the spoken and written language. There are a few previous annotation efforts for manually identifying discourse connectives in spoken language, in particular for Italian telephone help-desk conversations [22], English telephone conversations and broadcast interviews [23], and TED talks in several languages [20]. Some conceptual linguistic work has also looked at spoken genres in particular and compared the means of discourse marking for speech vs. other media.

Since the studies are not very systematic yet and often only analyze small corpora for individual languages, a mixed picture emerges. In synopsis, speech is said to contain fewer discourse relations (whether marked explicitly or unmarked) overall [22], but in speech, the discourse relations that occur are more often marked explicitly using connectives: About 2 out of 3 relations are marked in speech, while only half of discourse relations in writing contain connectives [22, 23]. In addition, discourse relation structures have been found to be frequently truncated (one or more argument clauses are missing) in spoken language [24]. Even just investigating explicitly marked discourse relations, there are consistent differences between genres. For example, speech uses more connectives with large scope and vague, multi-functional semantics [24], especially temporal and causal relations [22, 23]. In writing, relations between entities dominate. Some works also propose new types of discourse relations for conversations that are not often found in monological text, such as Repetition and the question-answer relation Hypophora [22, 20]. In our work, we for the first time investigate Ukrainian spoken and written discourse wrt. discourse connectives, and will test whether these general tendencies carry over to this new language and domain.

3. Ukrainian Connective Lexicon Construction

In this section, we describe our construction method for the first lexicon of Ukrainian discourse connectives. We employed an automatic method using a bilingual aligned corpus for selecting a set of connective candidates from the Ukrainian language, and then in a second step filtered these candidates and enriched the lexicon with syntactic and semantic information manually via a crowd-sourcing process.

We assume the definition of “discourse connective” given in [18]: A lexical item or phrase is a connective if (i) it is not inflectable, (ii) its meaning is a two-place relation, (iii) whose arguments are abstract objects (i.e., propositions, facts, utterances, etc.), (iv) and typically expressed as clauses. Typical examples for discourse connectives in English are ‘because’, ‘however’, or ‘in addition’. Despite this relatively concise definition, the class of discourse connectives is heterogeneous and cannot be easily approximated using standard lexicons or morphological/syntactic properties (such as part of speech). The reason for this is that discourse connectives are primarily defined by their semantic and pragmatic role. Discourse connectives can be subordinating or coordinating conjunctions (though not all instances of conjunctions are discourse connectives), but also many types of adverbs, particles, as well as complex phrases. We therefore follow previous research in starting with lists of discourse connectives from other languages and projecting them into the target language Ukrainian. This has the additional advantage that each Ukrainian connective comes with a cross-linguistic link to another language.

3.1. Semi-automatic Candidate Selection

In order to build a connective lexicon on a short time scale without any similar works found in Ukrainian NLP to base our research upon, we decided to start with a bilingual aligned corpus and a seed list of connectives from another language. As the most-used list of connectives in NLP, we chose the list of English connectives from the Penn Discourse Treebank corpus [12], which is widely used in discourse parsing. We extracted all Ukrainian words and phrases that have been matched to any of the English connectives in the automatically aligned English-Ukrainian dictionary based on the OpenSubtitles 2018 corpus [25, 26][3]. This seed extraction yielded 154 Ukrainian connective candidates.

3.2. Crowd-sourced Correction

We then proceeded to manually correct the list of candidates and enrich the list of connectives with metadata. To do this, Ukrainian native speakers were asked to review the list of candidates and decide on (1) whether or not they can be used as connectives according to the definition given above, and if so, (2) their syntactic properties and (3) their meaning. The native speakers were around 50 participants in a workshop on Shallow Discourse Parsing held at the offices of Grammarly, Kyiv, by the first author in February, 2020. They received brief training on the criteria for connectives [18] and lists of the available semantic and syntactic categories. In addition, we provided corpus examples for all connective candidates, automatically extracted by string-matching from the QED corpus [27]. The workshop participants worked in small groups to decide on the three questions given above, yielding an initial draft version of the annotated connective lexicon.

Apart from the connective candidate itself, this draft version contained the following information: “connective”, “syntax”, up to three semantic senses, as well as “comments”. The column “connective” with binary values “conn” vs. “no-conn” was the most important part of the work since we had to make a decision whether or not the candidate is a connective and will be part of the lexicon. In addition to true connectives, the first draft list contained a large number of russified variants of the connectives as well as a lot of duplicates.

The other time-consuming and very important part of the lexicon creation was assigning PDTB senses for each candidate. Since one connective can express several different meanings in different contexts, three sense columns were introduced to the lexicon draft to gather all the variants. For instance, the connective ‘a’ (‘and/but’), can provide the senses of comparison and/or contrast (e.g., “Один за вас життя віддасть, а інший також помре…”) or the sense of temporal synchronicity (e.g., “…моєю двоюрідною сестрою, в котрої під час важких боїв загинула мама, Хотина, а на фронті смертю хоробрих поліг її рідний брат…”). The first

[3] The dictionary is publicly available at: https://object.pouta.csc.fi/OPUS-OpenSubtitles/v2018/dic/en-uk.dic.gz

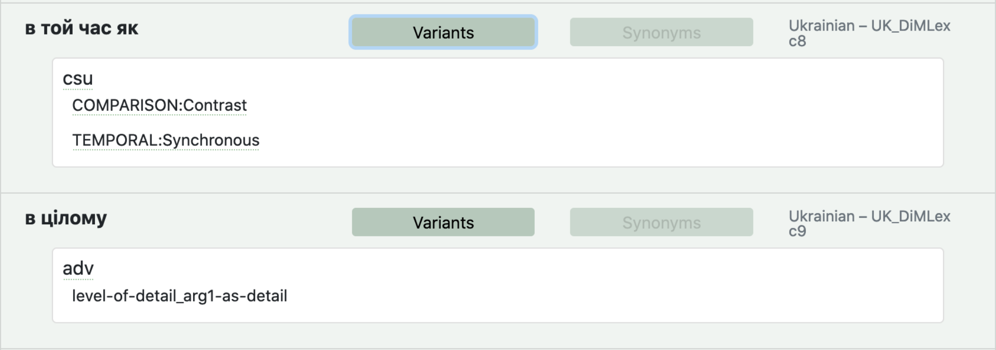
sense column represents the most common sense of the connective, and the third the most rare one.

3.3. Expert Correction

The work on the second draft consisted mostly of cleaning the duplicates and removing the so-called “Russisms” [28], which are russified words and expressions that are widely used by Ukrainian speakers, but are not considered part of standard Ukrainian, as for example connectives однако (‘however’) and ілі (‘or’). For each of the connectives, we also additionally reviewed the senses as well as the syntactic properties. This step was carried out by two trained linguists with experience in discourse annotation and native speaker competence in Ukrainian (the second and third author). The extra columns added in this stage were the English translation for each connective, the binary columns “continuous vs. discontinuous connective” and “single vs. phrasal connective” and “orthographic variants of the connective”. These last three new columns are complying with the format of the DiMLex data fields [14].

To complete the list with more connectives, additional candidates were automatically extracted from the the POS tag dictionary for the Ukrainian language[4]. We extracted all words with the ‘conj’ tag (for conjunctions) and manually excluded archaic expressions. Then, the extracted conjunctions were manually inserted into the existing list alongside with their senses and English translation. In total, 47 connectives were added in this step.

3.4. UK-Dimlex: The Lexicon of Ukrainian discourse markers



в той час як

Ukrainian – UK_DiMLex
c8

csu
COMPARISON:Contrast
TEMPORAL:Synchronous

в цілому

Ukrainian – UK_DiMLex
c9

adv
level-of-detail_arg1-as-detail

Figure 1: Screenshot of the lexicon as provided through connective-lex.info

The final validated lexicon of Ukrainian connectives contains 129 entries, and is provided in open access in xml format[5]. Example entries, shown in Figure 1, consist of the connective itself, possible spelling variants, its syntactic category (coordinating conjunction, subordinating conjunction, adverb, preposition, or other[6]), and its possible semantic senses according to the

[4] https://github.com/brown-uk/dict_uk

[5] https://github.com/TScheffler/UK_DiMLex

[6] The syntactic category ‘other’ currently only includes phrases (which do not have a single part of speech tag), such as не тільки…а й (‘not only…but also’).

PDTB3 hierarchy [29]. Some statistics are shown in Table 1. The xml lexicon further provides an English translation for each connective. The lexicon has been integrated into the multilingual online database connective-lex.info, which allows users to search connective lexicons across 10 languages in a common format.

Table 1
Basic statistics of connectives in the Ukrainian lexicon.

syntax[7]	
coordinating conjunction	14
subordinating conjunction	49
adverb	41
preposition	9
other	18
form	
single word	120
phrasal	35
total	129

4. Case Study

As described above, discourse connective lexicons can be useful for linguistic research as well as NLP applications, see also [18]. As a case study, we employ the lexicon to annotate a sample of Ukrainian spoken and written media and compare the results with each other and cross-linguistically with results from such comparisons in other languages.

4.1. Data

We chose a subsample from the corpus of Ukrainian spoken and written media discourse GRAK [6]. The corpus includes texts from 2000 to 2020. It is divided into two parts: conversational media discourse and written media discourse. The part with conversational media discourse has more than four hundred thousand words. It includes texts that are divided into two types: interviews and speeches. Texts from the interviews were collected from online versions of Ukrainian newspapers, magazines and other online publications. Texts with speeches consist of Ukrainian political figures' public speeches (congratulations on the new year, plans for the future).

The part with written media discourse has more than a million words. The texts in this section are divided into discourse from the social network Facebook and newspaper texts. It includes texts from the Ukrainian newspapers "Krymska Svitlytsia", "Galnet" and "Ukrainska Pravda". A significant part of this section is texts from Facebook posts of Ukrainian politicians, activists, artists and scientists.

[7] Note that some connectives can be ambiguous between more than one syntactic category.

For this research we left interviews outside of the scope of this paper, and chose 10 official speeches for our analysis. As for the written discourse we only consider 10 articles from "Krymska Svitlytsia", leaving social media material for future work.

4.2. Annotation

We pre-annotated the corpus in focus with the help of the Ukrainian Connective Lexicon to facilitate the manual annotation process. All occurrences of items from our lexicon were automatically marked as discourse connective candidate instances. In case of discourse markers, and Ukrainian language in particular, this approach has many limitations. On the one hand, although Ukrainian already has several open-source tools for tokenization and sentence-segmentation[8], their application on these data shows lower quality. On the other hand, many connectives coincide in form with conjunctions without discourse function (as the word і ('and') in the phrase письменник Микола Гоголь *і* художник Андрій Вархола, 'the writer Nikolay Gogol *and* the artist Andrew Warhol'). These items create a large number of false positives. This means that annotators need to correct more false instances than annotate new ones.

We chose WebAnno[9] to manually correct the pre-annotated instances, re-creating WebAnno TSV v3.2 format [30]. Three trained linguists, from different regions of Ukraine (authors 2, 3, 4), each annotated up to 10 of the selected texts. In order to calculate the inter-annotator agreement, 8 texts (5 articles, 3 speeches) were annotated in overlap between two annotators. In addition to disambiguating connective from non-connective occurrences, the annotators also added semantic PDTB3 senses to each of the connective uses. As a post-processing step, we manually corrected disjoint continuous connectives, such as 'not only…but also' and its variants, in order to combine annotations between the two spans of the connective into one.

Since all annotators only worked on a subset of the data, and since we have more than two annotators, we report Fleiss κ for inter annotator agreement. For the binary decision between connective or non-connective, we reached reasonable token-wise agreement for the news articles ($\kappa = 0.67$) but not the speeches ($\kappa = 0.21$) The overall total agreement for the overlapping texts was $\kappa = 0.62$. Due to the small sample annotated, these scores actually reflect a relatively small number of absolute overall disagreements; the overall absolute agreement is 0.97.

Annotation of connective senses is notoriously difficult, since they are often ambiguous or vague in context. Major annotation efforts even allow the parallel assignment of two senses to the same connective instance [12]. We reached a moderate agreement of $\kappa = 0.48$ on the top-level senses in the articles in the first run, which include Contingency, Comparison, Expansion, Temporal, or "None" (for all texts, $\kappa = 0.44$). We believe that further training of the annotators and discussion and adjudication of difficult examples can further improve the agreement between them.

[8] see Section 2.1

[9] https://webanno.github.io/webanno/

4.3. Results

We measured the annotation results, to compare the distribution of connectives quantitatively and qualitatively in speeches and articles. We found 597 connective instances in the speeches, which is 3.8% of all tokens, and 740 connective instances or 3.6% of tokens in articles. This means that the articles, on average, have slightly longer texts with slightly smaller proportion of connectives. Based on previous research, we expected speeches to have fewer discourse relations, but to use relatively more explicit connectives to mark them. In our data, these two tendencies seem to balance out to yield about equal numbers of explicit discourse relations.

For both speeches and articles, even after correction of the non-connective occurrences, і (‘and’) is the most frequent connective (162 for speeches and 135 for articles). А (‘but’, 44 sp. / 90 art.), як (‘as’, 37 sp./ 38 art.), але (‘but’, 53 sp. / 28 art.), and та (‘and’, 19sp. / 34 art.), are also among the most frequent ones for both genres. In articles, however, ’то’ (19) and ’ще ’(28) are also comparatively high in frequency, while ’ще’ in speeches occurs only 6 times, with no uses of ’то’ found at all. Another curious difference, is that in writing, ‘не тільки...але/а й/і’ is more frequent, than ‘не лише...але/а й/і’ (5 to 1), with the totally opposite situation in spoken language (1 to 7). Articles also seem to use more connectives related to justifications, as ‘бо’, ‘щоб’, ‘адже’, ‘тому що’, ‘наприклад’, ‘оскільки’ (51 to 41 in speeches). The distributions of connectives in the two types of media are shown in Figure 2.

It is worth noting that our annotators were able to add connective annotations, even when a word or phrase was not previously present in the connective lexicon. In this data-driven way, the annotators identified 50 additional connective candidates, listed in Table 2 (note that some of these candidates are spelling variants of existing connectives). Thus, the annotation process can also serve as validation and improvement of the lexicon itself. After a quality check, we will add these new connectives to the lexicon to greatly improve its coverage, releasing a new version.

Table 2
New connective candidates identified during annotation.

‘а саме’, ‘аж’, ‘аніж’, ‘більше ніж’, ‘більше того’, ‘більше’, ‘в тому’, ‘вже’, ‘відтак’, ‘для того щоб’, ‘до речі’, ‘доти’, ‘дотого ж’, ‘коли ще’, ‘лише’, ‘між тим’, ‘не лише’, ‘не лише...а й’, ‘не лише...але й’, ‘не лише...але і’, ‘не тільки’, ‘не тільки...але і’, ‘не тільки..але й’, ‘незалежно’, ‘незважаючи’, ‘ось чому’, ‘по-друге’, ‘по-перше’, ‘поки що’, ‘попри це’, ‘при цьому’, ‘так само’, ‘так само...як і’, ‘таки’, ‘тепер’, ‘тим більше’, ‘тим не менше’, ‘тим’, ‘у зв’язку з’, ‘у свою чергу’, ‘у той час як’, ‘час’, ‘чим’, ‘чому’, ‘що’, ‘щоб’, ‘якби не’, ‘яку’, ‘і ще’, ‘–’

Turning to the senses, as it can be seen in Figure 3, connectives indicating Expansion (such as Conjunction, Elaboration) are the most frequent for both speeches and articles (16.2 and 11.3 connectives per one thousand tokens). The Comparative sense class is the second most used, and curiously articles outrun speeches with 10 to 9/1K tokens. Contingency (Cause and Condition) and Temporal connectives are almost equal in distribution for both genres and represent third (both around 8.1/1K toks.) and forth (4.5 and 4.7/1K toks.) most frequent senses respectively. It can be seen that overall, Expansion relations are much more frequent in

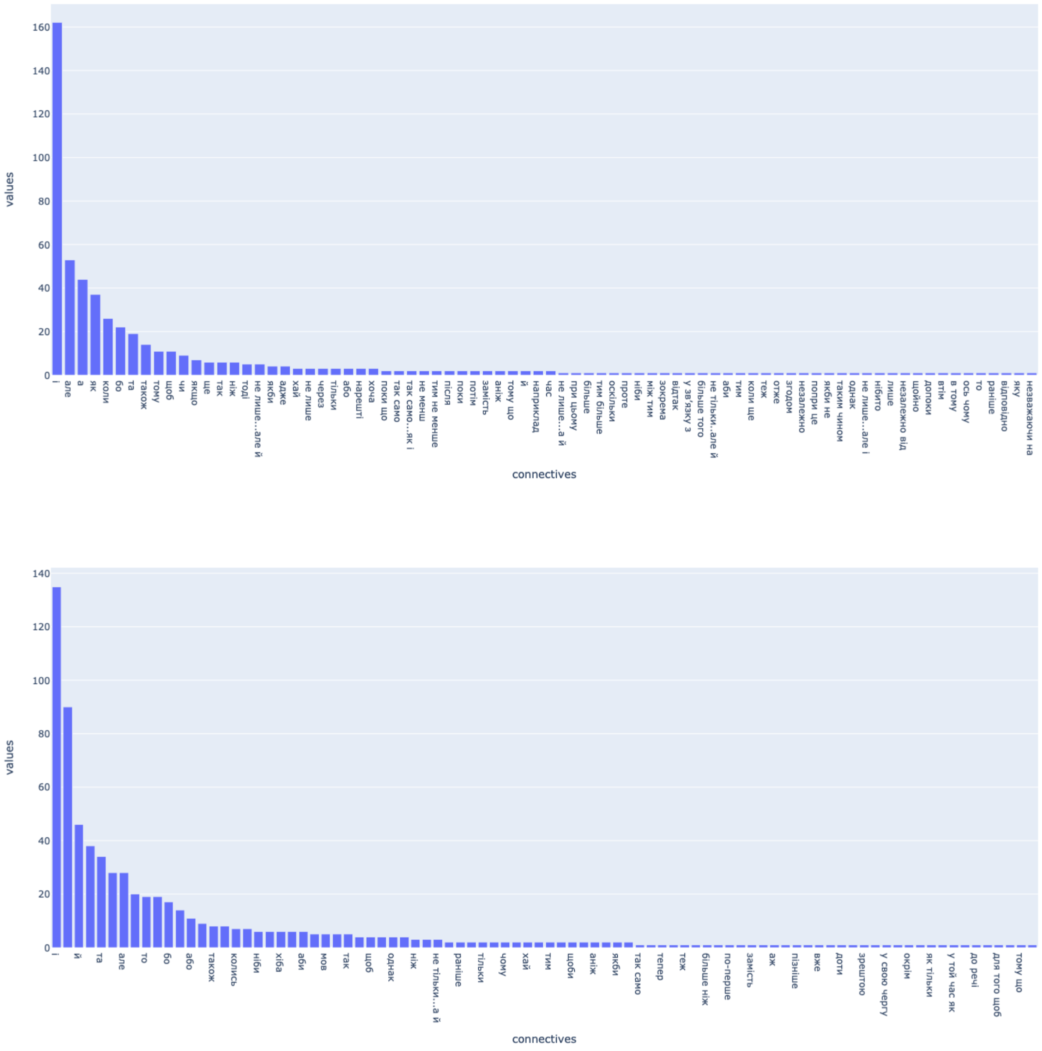


Figure 2: Distribution of connectives in speeches (top) and articles (bottom).

speeches than articles, which differs from previous accounts such as [22] for Italian and [23] for English, who observe higher frequency of Temporal and Causal relations in spoken language.

5. Conclusion and Outlook

In this paper, we presented the first lexicon of Ukrainian discourse connectives. We have discussed our process for constructing this lexicon with the help of existing resources for another language (English), automatically aligned bilingual corpora, crowd-sourcing, and a

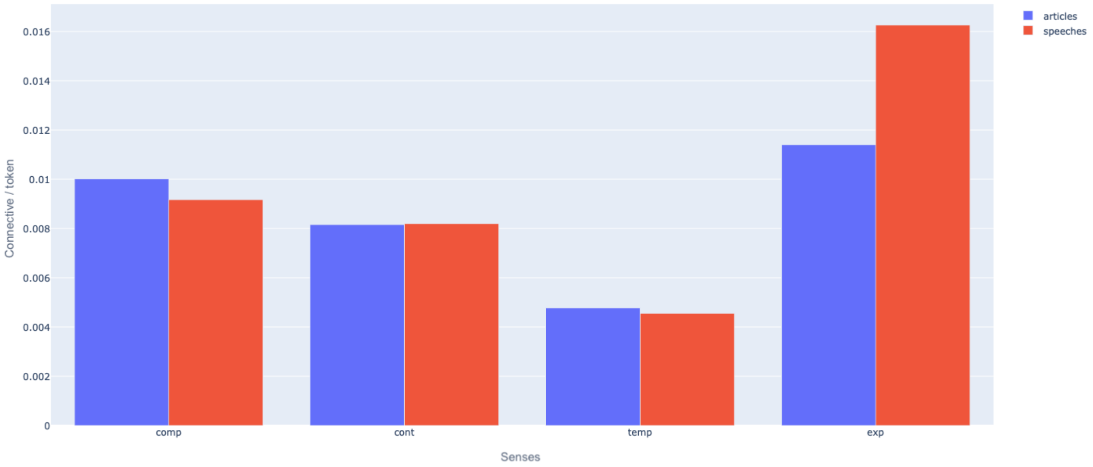


Figure 3: Senses of the annotated connectives normalized by the overall number of tokens per genre.

relatively small amount of expert manual correction. We believe that this approach can be useful for constructing other similar resources for new languages.

Further, we have used our lexicon to pre-annotate a small corpus sample of spoken and written texts, and we have carried out manual corrections on the annotations. Our analysis shows interesting differences in the usage of discourse connectives in speech vs. writing in Ukrainian, which partially reconfirm earlier findings in other languages.

We believe that our lexicon and corpus sample will be useful for linguists and computational linguists interested in studying the Ukrainian language at the discourse level and NLP practitioners who need access to discourse context. We, therefore, make our lexicon publicly available in a standard format. A connective lexicon is often the first step towards discourse parsing, which allows the identification of textual relations beyond the sentence. In future work, we are planning to build a shallow discourse parser for detecting explicit discourse relations in Ukrainian.

Acknowledgments

TS would like to thank the organizers and participants of the 2020 Grammarly workshop on Shallow Discourse Parsing, and in particular Mariana Romanyshyn for the invitation to Kyiv.

References

[1] N. Xue, H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, A. Rutherford, The CoNLL-2015 shared task on shallow discourse parsing, in: Proceedings of the Nineteenth Conference on Com-

putational Natural Language Learning - Shared Task, Association for Computational Linguistics, Beijing, China, 2015, pp. 1–16. URL: https://www.aclweb.org/anthology/K15-2001. doi:10.18653/v1/K15-2001.

[2] N. Xue, H. T. Ng, S. Pradhan, A. Rutherford, B. Webber, C. Wang, H. Wang, CoNLL 2016 shared task on multilingual shallow discourse parsing, in: Proceedings of the CoNLL-16 shared task, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1–19. URL: https://www.aclweb.org/anthology/K16-2001. doi:10.18653/v1/K16-2001.

[3] B. Verhoeven, W. Daelemans, Discourse lexicon induction for multiple languages and its use for gender profiling, Digital Scholarship in the Humanities 34 (2018) 208–220. URL: https://doi.org/10.1093/llc/fqy025. doi:10.1093/llc/fqy025.

[4] X. Wang, K. Evanini, K. Zechner, M. Mulholland, Modeling discourse coherence for the automated scoring of spontaneous spoken responses, in: Proc. 7th ISCA Workshop on Speech and Language Technology in Education, 2017, pp. 132–137.

[5] E. Cabrio, S. Tonelli, S. Villata, From discourse analysis to argumentation schemes and back: Relations and differences, in: J. Leite, T. C. Son, P. Torroni, L. van der Torre, S. Woltran (Eds.), Computational Logic in Multi-Agent Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 1–17.

[6] Шведова, М., фон Вальденфельс, Р. , Яригін, С., Рисін, А. , Старко, В., та ін., Генеральний регіонально анотований корпус української мови (ГРАК), 2017–2021. URL: http://uacorpus.org.

[7] Старко, В. Ф., Формування Браунського корпусу української мови, Мовні і концептуальні картини світу 48 (2014) 415–421.

[8] O. Syvokon, O. Nahorna, Ua-gec: Grammatical error correction and fluency corpus for the ukrainian language, 2021. arXiv:2103.16997.

[9] V. Dyomkin, lang-uk, https://github.com/lang-uk, 2015.

[10] B. Webber, M. Egg, V. Kordoni, Discourse structure and language technology, Natural Language Engineering 18 (2012) 437–490.

[11] A. Knott, A data-driven methodology for motivating a set of coherence relations, Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

[12] B. Webber, R. Prasad, A. Lee, A. Joshi, The Penn Discourse Treebank 3.0 Annotation Manual, Technical Report, University of Pennsylvania, 2019. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.

[13] G. Riccardi, E. A. Stepanov, S. A. Chowdhury, Discourse connective detection in spoken conversations, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6095–6099. doi:10.1109/ICASSP.2016.7472848.

[14] M. Stede, DiMLex: A lexical approach to discourse markers, in: A. Lenci, V. D. Tomaso (Eds.), Exploring the Lexicon: Theory and Computation, Edizioni dell’Orso, 2002.

[15] T. Scheffler, M. Stede, Adding semantic relations to a large-coverage connective lexicon of German, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 1008–1013.

[16] D. Debopam, T. Scheffler, P. Bourgonje, M. Stede, Constructing a lexicon of English discourse connectives, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Melbourne, Australia, 2018, pp.

360–365.

[17] A. Feltracco, E. Jezek, B. Magnini, M. Stede, LICO: A lexicon of Italian connectives, in: Proceedings of CLiC it, 2016. URL: http://ceur-ws.org/Vol-1749/paper24.pdf.

[18] M. Stede, A. Mendes, T. Scheffler, Connective-lex: A web-based multilingual lexical resource for connectives, Discours. Revue de linguistique, psycholinguistique et informatique (2019).

[19] J. Mírovskỳ, P. Synková, M. Rysová, L. Poláková, CzeDLex: A lexicon of czech discourse connectives, The Prague Bulletin of Mathematical Linguistics 109 (2017) 61.

[20] D. Zeyrek, A. Mendes, M. Kurfalı, Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 1913–1919. URL: https://www.aclweb.org/anthology/L18-1301.

[21] D. Samy, A. González-Ledesma, Pragmatic annotation of discourse markers in a multilingual parallel corpus (Arabic- Spanish-English), in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, pp. 3299–3305.

[22] S. Tonelli, G. Riccardi, R. Prasad, A. Joshi, Annotation of discourse relations for conversational spoken dialogs, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010, pp. 2084–2090.

[23] I. Rehbein, M. Scholman, V. Demberg, Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 1039–1046. URL: https://www.aclweb.org/anthology/L16-1165.

[24] L. Crible, M. Cuenca, Discourse markers in speech: Characteristics and challenges for corpus annotation, Dialogue and Discourse 8 (2017) 149–166. doi:10.5087/dad.2017.207.

[25] P. Lison, J. Tiedemann, OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 923–929. URL: https://www.aclweb.org/anthology/L16-1147.

[26] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218.

[27] A. Abdelali, F. Guzman, H. Sajjad, S. Vogel, The AMARA corpus: Building parallel language resources for the educational domain, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 1856–1862.

[28] J. Ajdukovic, An Introduction to Lexical Contact: The Theory of the Adaptation of Russisms In South and West Slavic Languages, Foto Futura, 2004.

[29] R. Prasad, B. Webber, A. Lee, Discourse annotation in the PDTB: The next generation, in: Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 87–97. URL: https://www.aclweb.org/anthology/W18-4710.

[30] R. Eckart de Castilho, É. Mújdricza-Maydt, S. M. Yimam, S. Hartmann, I. Gurevych, A. Frank, C. Biemann, A web-based tool for the integrated annotation of semantic and syntactic structures, in: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 76–84. URL: https://www.aclweb.org/anthology/W16-4011.