Creating Parallel Corpora for Ukrainian: a German-Ukrainian Parallel Corpus (ParaRook||DE-UK)

Maria Shvedova, Arsenii Lukashevskyi
National Technical University "Kharkiv Polytechnic Institute", University of Jena
Kyrpychova 2, 61002, Kharkiv, Ukraine; Ernst-Abbe-Platz 8, 07743, Jena, Germany
Mariia.Shvedova@khpi.edu.ua, Arsenii.Lukashevskyi@sgt.khpi.edu.ua

Abstract

Parallel corpora are currently a popular and vibrantly developing category of linguistic resources, used both in literature and translation studies, as well as in the field of NLP. For Ukrainian, though, there are still not enough significant parallel corpora compiled within a single roof project and made available to the research community. In this paper we present a newly developed resource, the German-Ukrainian Parallel Corpus — ParaRook||DE-UK, searchable online. We describe various issues related to its compilation, text selection, and annotation. The paper also features several examples of how the corpus can be used in linguistic research and translation studies. Using the experience of the German-Ukrainian parallel corpus, parallel corpora for other languages with Ukrainian can be developed.

Keywords: parallel corpus, corpus annotation, Ukrainian, German, translation

1. Parallel Corpora for Ukrainian

Parallel corpora are a valuable linguistic resource that is applied primarily for translation research and practice as well as comparative linguistic studies, it can also be useful for monolingual studies. With the development of computer technologies, the role of parallel corpora as datasets for machine translation is becoming increasingly important. Though datasets of parallel sentences in different languages are often collected automatically from the Internet, it is still useful to create some parallel corpora semi-manually, especially for fiction texts that typically lack exact (word-for-word) match between the original and the translation, which makes it difficult to automatically collect and align them. Here are several references to the books on the use of parallel corpora in linguistic studies, translation studies and translation teaching (Anderman and Rogers, 2007; Hansen-Schirra et al., 2012; Enghels et al., 2020; Liu, 2020)

For the Ukrainian language, there are still few parallel corpora available online for searching. One of these projects is the Polish-Ukrainian parallel corpus (Kotsyba, 2016), which has size of about 4 million tokens in the Polish part and is searchable both via an older search manager (Kotsyba and Turska, 2005 - 2011) and on the NoSketchEngine platform on the website of the Laboratory of Ukrainian project. (Kotsyba, 2018) The site also published a parallel English-Ukrainian corpus of 1.5 million tokens in the English part and smaller French-, German-, Spanish-, and Portuguese-Ukrainian bilingual pairs (500, 190, 65 and 16 thousand tokens, respectively) containing literary texts, including some translated from a third language.

The largest collection of semi-manually aligned parallel texts with Ukrainian is now available for search as a part of the InterCorp parallel corpora collection (Čermák and Rosen, 2012). In InterCorp v.16, the volume of Ukrainian texts is over 18 million tokens with aligned originals or translations into Czech and other languages through Czech as a pivot language. The Ukrainian part of InterCorp consists mainly of fiction texts and a smaller dataset featuring subtitles and the Bible. (Čermák and Rosen, 2008 - 2023)

A one-million-tokens dataset of Ukrainian parallel fiction and medical texts with French, English, and Polish is available for download on Natalia Grabar's site (Grabar and Hamon, 2017).

A significant part of the existing Ukrainian parallel corpora is not currently available to the Ukrainian community for various reasons. Access to the Ukrainian-Russian parallel corpus within the Russian National Corpus (Sitchinava et al., 2011) is blocked in Ukraine since 2017 due to the war. ParaSol: a Slavic Parallel Corpus is currently under reconstruction (von Waldenfels, 2011). The following corpora have not been published: Bulgarian-Ukrainian parallel corpus KUB (Siruk and Derzhanski, 2013), Polish-Ukrainian and Ukrainian-Polish parallel corpus of Ivan Franko's self-translations (Buk, 2012), English-Ukrainian parallel corpus ParKUM (Darčuk et al., 2017), English-Ukrainian parallel corpus of Legal Texts (Matvieieva, 2019), English-Ukrainian parallel corpus compiled by Serhij Zasiekin (Zasiekin, 2020), English-Ukrainian parallel corpus of IT texts (Mandziy et al., 2022) etc. Smaller user collections of parallel texts are created by students at various Ukrainian universities, such as Lviv Polytechnic, Odesa National University, Kherson National Technical University, and others, for educational purposes, but there is no

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 14–22
25 May, 2024.

Original/Style	Fiction	Nonfiction
English (EN)	48,716,969	8,962,108
Russian (RU)	18,265,944	2,413,472
French (FR)	17,844,342	2,462,649
Polish (PL)	10,931,819	1,816,123
German (DE)	9,661,714	2,520,939
Czech (CS)	4,130,289	389,861
Spanish (ES)	3,641,073	450,012
Italian (IT)	3,413,198	598,576
Bulgarian (BG)	2,736,933	109,597

Table 1: The scope of translated texts in GRAC v.17 by original language and style.

coordinated system that would accumulate these materials and make them available for use.

A valuable resource for creating Ukrainian parallel corpora is GRAC (Shvedova, 2017 - 2024): the Ukrainian language reference corpus, which contains translations from 89 languages, mostly fiction, with a total size of 172 million tokens of texts translated from different languages (GRAC v.17). The size of the largest subcorpora of translated texts in GRAC by language and style is shown in the Table 1.

There are many parallel corpus projects where texts are collected and aligned automatically. One such project is ParaCrawl (Bañón et al., 2020), notably its MultiParaCrawl [1] corpus series, which includes 705 bilingual language pairs for 41 languages, including 36 pairs for Ukrainian with different European languages. The languages were identified with Google's Compact Language Detector 2, and neural network technologies were applied for text alignment and cleaning. Specifically, this corpus was prepared for the OPUS (Tiedemann, 2012) project by pivoting text documents through English to achieve a massive parallel corpus. It includes only the new language pairs built by this procedure and can be downloaded in TMX, XML, and Moses formats from the OPUS website. As of March 2024, it includes 34 million sentences in Ukrainian.

The ParaCrawl project itself is focused more on English-centric language pairs and is larger than MultiParaCrawl (compare 1.5 billion sentences of this project and 789 million of MultiParaCrawl). However, it includes 14 million sentences in Ukrainian. It is freely available for download in TMX, TXT, and raw formats. In addition, it is distributed via OPUS.

Another project is NLLb (Schwenk et al., 2019; Fan et al., 2021), a large dataset containing bitext for 148 English-centric and 1465 non-English-centric language pairs. The dataset was created based on metadata for mined bitext released by Meta AI. It was filtered for language identification, emoji-based filtering was performed, and, for some high-resource languages, a language model was applied. The data was processed using the stopes mining library and the LASER3 encoders (Costa-jussà et al., 2022). Currently, it includes 166 million tokens in Ukrainian.

MultiCCAligned (El-Kishky et al., 2020) is a parallel corpus comprising web-document pairs in 137 languages aligned with English. The corpus was created by performing language identification on raw web documents and ensuring that corresponding language codes match the URLs of web documents. More than 100 million aligned documents were paired with English. Some English documents were aligned to multiple documents in different target languages. Sentence pairs were extracted using similarity scores of LASER embeddings from the document pairs. The latest release of MultiCCAligned is v1.1, created from 68 Commoncrawl Snapshots up until March 2020. It includes 62 million sentences in Ukrainian.

MaCoCu (Bañón et al., 2022) is a multilingual parallel corpus built by crawling national internet top-level domains. The corpus was processed using the Bitextor tool, with considerable effort put into cleaning the extracted text. Accordingly, the MaCoCu-uk-en 1.0 was created based on scanned data from sites on the .ua domain and includes 238,841,101 tokens.

Also, an essential source of parallel texts is Wikipedia. One project that has put this into practice is WikiMatrix. The project focuses on languages with low resources, making it a valuable dataset for researchers and developers working with less commonly studied languages. Currently, WikiMatrix provides parallel data for over 1620 language pairs. The authors state that their project makes 135 million parallel sentences available in 96 languages, of which only 34 million are aligned with English. One of the largest pairs is the Ukrainian-Russian one, amounting to 2.5 million sentences (Schwenk et al., 2019). It should be noted that the source of Ukrainian-Russian sentences could be numerous Ukrainian sites with parallel language versions, which tend to use automatic translation.

The Ukrainian language is represented in two multilingual parallel corpora on Sketch Engine, namely OPUS parallel corpus covering 40 languages (the size of Ukrainian texts is 2.5 million tokens) and OpenSubtitles: multilingual corpus in 58 languages (the size of the Ukrainian part is 5 million tokens) (Lison and Tiedemann, 2016).

Only some automatically built parallel corpus projects are listed in this section. It is essential to mention, for example, the OpenSubtitles corpus practices. It differs from other automatic cor-

[1] https://paracrawl.eu/news/item/18-multiparacrawl-9-including-ukrainian

pora by using a time-based approach and intra-language alignment as subtitles in one language often have many variants, which allows for more accurate learning of nuances and variations in the language (Lison and Tiedemann, 2016).

Most of the listed projects are available for download through the OPUS website or in different formats. One is the TXT format ParaCrawl, a bilingual text where sentences are aligned in a one-line per sentence format in 2 columns. For users who need to become more familiar with technology, the Sketch Engine platform will be helpful, featuring both OpenSubtitles and parallel corpora in 40 languages from OPUS in a search interface.

However, it is essential to note that they often lack precision in alignment and data cleanliness, which can impact the quality of the results. For instance, due to the automated nature of the alignment process, there may be instances where sentences or phrases are not accurately matched (Zariņa et al., 2015). Similarly, cleaning may sometimes leave irrelevant or noisy data.

Such projects are useful for purposes such as training machine translation models (Tiedemann and Thottingal, 2020), but are not completely suitable for linguistic research due to their frequent noisiness and lack of accuracy, which is impossible on such large arrays of text. This is why, despite advances in technology and smaller size, manually collected corpora are extremely useful in literature research, translation studies, comparative and typological studies.

In this paper, we present ParaRook||DE-UK (Shvedova and Lukashevskyi, 2023-2024), which is the first large German-Ukrainian corpus collected and verified manually, with detailed meta-annotation and morphosyntactic annotation, and searchable online. The title refers to the Ukrainian monolingual reference corpus GRAC (*grak* is the Ukrainian name for rook) and also sounds like "pair of hands" in Ukrainian.

2. Composition and structure of ParaRook||DE-UK

2.1. Texts

The history of German-Ukrainian literary translation is a complex and interesting field (Ivanytska, 2015). We aimed to show samples of German-Ukrainian translation from different periods, namely Soviet, with specific features of the time, and contemporary.

As shown by M. Ivanytska, German-Ukrainian translations were sometimes made not directly between two languages, but through the mediation of a Russian translation. We tried to reduce the amount of such texts in the corpus, because the influence of the intermediary language is often very noticeable in them. In the example below, the Russian translator did not render the author's idiom, but instead used expressive syntax. The Ukrainian translator calqued this syntactic construction, which is not very frequent in Ukrainian.

- *(de) Ich bin ein ausgewichster Panzermann, aber die sind doch keine halbe Nase weniger schlau! [I am a good tankman, but they are not less smart!] (Dieter Noll. Die Abenteuer des Werner Holt. 1960)*

  (ru) Už na čto ja byvalyj tankist, no oni ničut' ne glupee! (Translation by V. Kurilla, R. Galperin. 1962)[2]

  (uk) Naščo vže ja buvalyj tankist, ale vony anitroxy ne durniši! (Translation by Y. Mykhailyuk. 1965)

According to M. Ivanytska, censorship did occur in Ukrainian translations from German under the Soviet regime, and our material also shows this. In such cases, we keep the untranslated text in the corpus without a Ukrainian version (Table 2).

In literary translations from German, we also often find just omitted and shortened fragments, cases of inaccurate translation, rearranged sentences, etc. that are not related to censorship.

Lion Feuchtwanger. Erfolg. 1929	Ukrainian translation. Oleksa Oleksa Synyčenko. 1980
Möglich, daß in Bayern die Justiz besonders bösartig und verbohrt gehandhabt wurde, aber viel anders war es ringsum auch nicht. [It is possible that justice was administered in Bavaria in a particularly malicious and biased manner, but it was not much better in other countries.]	Možlyvo, ščo v Bavariï pravosuddja čynyly osoblyvo zlisno j uperedženo, ale ne nabahato krašče bulo i v inšyx kraïnax.
In Ungarn, auf dem Balkan, in Rußland stand es vielleicht noch schlimmer als auf der bayrischen Hochebene. [In Hungary, the Balkans, and Russia, the situation was probably even worse than on the Bavarian Plateau.]	—

Table 2: Soviet Censorship in German-Ukrainian Translation.

[2] Hereinafter, examples in Cyrillic are transliterated.

ParaRook||DE-UK has size of 382 thousand sentences and 6,3 million tokens in the German-language part. The core of the corpus currently consists of 20th-century fiction translated from German into Ukrainian. The corpus contains 58 texts: 53 translated from German and 5 from Ukrainian. The corpus features works by 29 famous German-speaking authors from different countries, which makes it possible to compare regional variants of the German language. The corpus includes novels by Erich Maria Remarque, Thomas Mann, Heinrich Mann, Hermann Hesse, Alfred Döblin, Dieter Noll, Heinrich Böll, Günter Grass, Patrick Süskind (Germany), Franz Kafka, Stefan Zweig, Robert Musil, Gustav Meyrink, Joseph Roth (Austria), Friedrich Dürrenmatt (Switzerland), and other writers (Appendix B).

2.2. Annotation and Technical Details

The texts for the corpus were collected manually from public libraries on the Internet (the sources are given in the metadata), most Ukrainian texts were taken from GRAC. The original texts and translations were aligned using the InterText program (Vondřička, 2014), and the alignment of all texts was checked and corrected manually. All the cases of inaccurate translation were saved in the corpus for research, the texts were aligned without changing the structure of the original text or translation. Aligned parallel texts are saved in tmx format, e.g.:

<tu><prop type="x-sentbreak">|#|</prop>

<tuv xml:lang="de"><seg>Der Knabe war klein, die Berge waren ungeheuer.</seg></tuv>

<tuv xml:lang="uk"><seg>Xlop'ja bulo male, hory – vysočezni.</seg>

</tuv>

</tu>

The parallel corpus was annotated with UDPipe2 (Straka, 2018) using Universal Dependencies models, namely GSD for German (Petrov, 2023) and IU for Ukrainian (Kotsyba, 2016). The choice of the German model was based on model evaluations on the official UD website (Nivre, 2015 - 2024), while the Ukrainian model was the only one presented. Ukrainian's current Universal Dependencies model achieves an accuracy rate of 97.5% for POS tagging, 91.6% for morphological features, and 81.7% for syntactic relation (Kotsyba, 2018).

The Universal Dependencies were chosen because their annotation is universal regardless of the language of the analyzed text, and the process can be optimized using graphics processors, particularly NVIDIA CUDA technology, which significantly speeds up computation, as opposed to using a traditional CPU.

The annotation of documents in the parallel corpus also includes syntactic relations between words within a sentence, which can serve as a source of data for contrastive syntactic analysis (Poiret et al., 2021). During the preprocessing of the corpus materials, the text was segmented into sentences using the SpaCy models appropriate to the language of the text: uk_core_news_sm (Kurnosov, 2022) and de_core_news_sm (Brants, 2023); this step was necessary to improve accuracy in morphosyntactic annotation using UD.

The corpus manager used was NoSketch Engine, one of the most featureful open-source corpus manager solutions available. The corpus is accessible for search on the website: https://uacorpus.org/Kyiv/ua/pararook

Besides morphosyntactic annotation, the corpus provides extensive metadata. The list below presents all the necessary information regarding metadata and tag descriptions.

word: Token attribute for a word.

lemma: Token attribute for lemma.

upos: Token attribute for UD part-of-speech tag.

xpos: Language-specific grammatical annotation token attribute.

morphology: Morphological annotation.

head: Syntactically the main word in a sentence.

dependency_tag: Syntactic relationship of a word in a sentence.

extra_dependency: Additional information about the syntactic role of a word in a sentence.

authors_names_{uk|de}: Authors' name in Ukrainian/German.

translators_names_{uk|de}: Translators' name in UK/DE.

authors_born: Authors' birth year.

authors_sex: Authors' gender.

authors_regionCode: Authors' region.

translators_regionCode: Translators' region.

translators_born: Translators' birth year.

translators_sex: Translators' gender.

title_{uk|de}: Document title in UK/DE.

original_language: Original language.

date_{uk|de}: Year of creation in UK/DE.

pub_city_{uk|de}: City of publication in UK/DE.

publisher_{uk|de}: Publisher in UK/DE.

pub_year_{uk|de}: Year of publication in UK/DE.

publication_{uk|de}: Title of publication in UK/DE (magazine number, title of collection).

url_{uk|de}: Reference to the source of the document in UK/DE.

An example of parallel sentences with metadata is provided in Appendix A.

3. Using ParaRook||DE-UK

Since ParaRook||DE-UK is only available for online search, it is intended primarily for academic linguistic and translation studies, for compiling dictionaries, as well as for use in the process of human translation.

A parallel corpus not only provides a richer range of translation options in context than a dictionary, but also enables research on phenomena that do not have a well-established translation: it can be used to study lacunarity and non-equivalent linguistic patterns (Sitchinava, 2016; Dobrovol'skij and Pöppel, 2017; Mellado Blanco, 2019; Grabowski and Groom, 2022)

For example, the German construction *immer noch* 'still' may be translated into Ukrainian in many different ways, or it may be omitted in translation at all. In a random sample of one hundred parallel sentences from ParaRook||DE-UK, the following translation variants were found: *i(j) dosi (21 times), vse(use) šče (15), šče(išče) (15), i(j) dali (7), tak samo (7), j (4), vse (2), vse(use) ž taky (2), vse odno (1), dali (1), zavždy (1), i vse odno (1), i dosi šče (1), j tak samo (1), poky ščo (1), skil'ky zavhodno (1), tak use j (1), teper (1), u krajn'omu razi (1), šče j dovho (1), šče j dosi (1), šče raz (1), jak i raniše (1).* In 12 cases of 100, the German construction had no equivalent in Ukrainian translation at all, e. g.

- *(de) Wir tranken, und* ***immer noch*** *standen die Uhrzeiger, wie sie schon seit drei Wochen standen: auf halb elf. (Heinrich Böll. Irisches Tagebuch. 1957)*

  (uk) My pyly, a strilky hodynnyka stojaly na misci, jak i ves' čas protjahom ostannix tr'ox tyžniv, — na piv na odynadcjatu. (Ukr. translation by Volodymyr Šelest. 1989)

When working with one language, the translation presented in the parallel corpus can be used as an additional layer of annotation, which makes it possible to search by semantics. This advantage of parallel data is already being extensively used for automatic word sense disambiguation (Yee Seng Chan and Zhong, 2007; Hwee Tou Ng and Chan, 2003; Banea and Mihalcea, 2011; Shahid and Kazakov, 2013), and it can be useful for a manual lexical research as well. Below is an example of search results in ParaRook||DE-UK of a Ukrainian word *kaminec'* that has two meanings, a commonly used 'small stone' and a rarely used 'fruit bone'. To find examples in the second meaning only, the German equivalent *Kern* was used, which helped to specify the required sense.

- *(de) Weißrot klappern Störche auf Dächern, daß Kirschen die* ***Kerne*** *ausspucken... (Günter Grass. Blechtrommel. 1959)*

  (uk) Bilo-červoni busly triskotjat' na daxax pro te, ščo vyšni vypl'ovujut' svoï ***kaminci****... (Ukr. translation by Oleksa Lohvynenko. 2005)*

- *(de) Sie brach eine der überreifen Früchte auf, warf den* ***Kern*** *zu Boden und reichte ihm eine der Hälften. (Dieter Noll. Die Abenteuer des Werner Holt. 1960)*

  (uk) Potim rozlomyla najspilišyj plid i, vykynuvšy ***kamінčyka****, prostjahla polovynu Hol'tovi. (Ukr. translation by Jurij Myxajljuk. 1965)*

More examples of the use of parallel corpora for manual research and teaching can be found in the relevant work presented in our bibliography.

4. Conclusions and Future Plans

The first representative German-Ukrainian parallel corpus has been created and is available to search online. This is an important language resource that provides parallel texts for linguistic and translation research. With this work, we would like to draw attention to the importance of making computational linguistic resources more inclusive for philologists, not only for wider use of such resources in academic work, but also for involving professional linguists, translators, and texts experts in the development of quality textual data.

In the future, we plan to add more texts translated from Ukrainian into German and to develop parallel corpora for other languages with Ukrainian, primarily English and French.

It is possible to create a much larger German-Ukrainian corpus based on ParaRook||DE-UK by adding non-fiction texts, such as legal, news, and subtitles, which are usually translated quite literally and require less manual alignment checking. They can be downloaded from the Internet and automatically aligned.

As currently a single Universal Dependencies model is available, we plan to expand the range of models at hand for the Ukrainian language in UD. This expansion aims to improve the accuracy of morphosyntactic analyses and contribute to developing more robust and diverse linguistic tools.

5. Limitations

Since we check the alignment manually, it would be a challenge to collect a corpus larger than several millions of tokens. Manual alignment checking is highly desirable for fictional texts, where the translation is often not quite literal, but it takes a lot of time.

Based on Universal Dependencies, the current morphosyntactic analysis of the Ukrainian language needs to yield optimal accuracy. Improving this system is of great importance for the further development of parallel corpora. While the current accuracy is promising, more is required for large corpora. Optimal accuracy is critical in syntax studies that use parallel corpora to ensure reliable and meaningful findings.

6. Acknowledgements

We owe our acknowledgements to the German academic community for their strong support of Ukrainian researchers and students.

We would like to thank the reviewers for their useful comments on the paper.

7. Bibliography

References

G. Anderman and M. Rogers. 2007. *Incorporating Corpora: The Linguist and the Translator*. Multilingual Matters.

C. Banea and R. Mihalcea. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. Forcada, A. Kamran, F. Kirefu, P. Koehn, et al. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. Association for Computational Linguistics (ACL).

M. Bañón, M. Esplà-Gomis, M. L. Forcada, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. P. Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, T. van der Werff, and J. Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

S. Buk. 2012. Arxitektura pol's'ko-ukraïns'koho ta ukraïns'ko-pol's'koho paralel'noho korpusu avtoperekladiv Ivana Franka. *Slavia Orientalis*, LXI(2):213–230.

M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

N. P. Darčuk, M. O. Lanhenbax, V. M. Sorokin, and Ja. V. Xodakivs'ka. 2017. Paralel'nyj korpus tekstiv ParKUM. *Naukovyj časopys Nacional'noho pedahohičnoho universytetu imeni M. P. Drahomanova. Serija 9 : Sučasni tendenciï rozvytku mov : zb. nauk. prac'*, 15:28–35.

D. Dobrovol'skij and L. Pöppel. 2017. Constructions in parallel corpora: A quantitative approach. In *Computational and Corpus-Based Phraseology*, volume 10596.

A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

R. Enghels, B. Defrancq, and M. Jansegers. 2020. *New approaches to contrastive linguistics: empirical and methodological challenges*. De Gruyter Mouton.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

N. Grabar and T. Hamon. 2017. Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In *COLINS 2017*, Kharkiv, Ukraine.

Ł. Grabowski and N. Groom. 2022. Functionally-defined recurrent multi-word units in English-to-Polish translation. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 35(1):1.

S. Hansen-Schirra, S. Neumann, and E. Steiner. 2012. *Cross-linguistic corpora for the study of translations: insights from the language pair English-German*. de Gruyter Mouton.

Bin Wang Hwee Tou Ng and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. pages 455–462.

M. Ivanytska. 2015. *Osobystist' perekladača v ukraïns'ko-nimec'kyx literaturnyx vzajemynax*. Knyhy – XXI, Černivci. [The Personality of the Translator in Ukrainian-German Literary Relations].

N. Kotsyba. 2016. Polsko-ukraiński korpus równoległy PolUKR i jego następca PolUKR-2. pages 133–142.

P. Lison and J. Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*.

Kanglong Liu. 2020. *Corpus-Assisted Translation Teaching: Issues and Challenges*, 1st edition. Springer Singapore.

K. S. Mandziy, U. V. Yurlova, and M. P. Dilai. 2022. English-Ukrainian parallel corpus of IT texts: Application in translation studies.

S. Matvieieva. 2019. Selection criteria and initial processing of empirical material for a parallel corpus of legal texts. *Forum Filologiczne Ateneum*, 1(7):167–181.

C. Mellado Blanco. 2019. Phrasem-konstruktionen kontrastiv Deutsch–Spanisch: ein korpusbasiertes beschreibungsmodell anhand ironischer vergleiche. *Yearbook of Phraseology*, 10(1):65.

Ra Poiret, S. Mille, and Haitao Liu. 2021. Paraphrase and parallel treebank for the comparison of French and Chinese syntax. *Languages in Contrast*, 21(2):298–322.

H. Schwenk, G. Wenzek, S. Edunov, E. Grave, and A. Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

A. R. Shahid and D. Kazakov. 2013. Using parallel corpora for word sense disambiguation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 336–341, Hissar, Bulgaria.

O. Siruk and I. Derzhanski. 2013. Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. In *Digital Presentation and Preservation of Cultural and Scientific Heritage*, volume 3, pages 91–98.

D. Sitchinava. 2016. Parallel corpora as a source of defining language-specific lexical items. In *Proceedings of the XVII EURALEX International Congress*, pages 394–401.

D. V. Sitchinava, O. O. Tyshchenko-Monastyrska, and M. O. Shvedova. 2011. Paralel’ni ukrayins’ko-rosiys’kyy ta rosiys’ko-ukrayins’kyy korpusy. *Leksykohrafichnyy byuleten*, 20:35–38.

V. Starko and A. Rysin. 2022. VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In *COLINS*, volume 6th Int. Conf, pages 71–80, Gliwice.

M. Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

J. Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC’2012)*.

J. Tiedemann and S. Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

R. von Waldenfels. 2011. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011*, pages 156–162, Bratislava. Tribun.

R. von Waldenfels. 2012. ParaSol: Introduction to a Slavic parallel corpus. *Prace Filologiczne*, LXIII:293–302.

P. Vondřička. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1875–1879. European Language Resources Association (ELRA).

Hwee Tou Ng Yee Seng Chan and Zhi Zhong. 2007. NUS-PT: Exploiting parallel texts for word sense disambiguation in the English all-words tasks. *Proc. 4th Int. Workshop Semantic Eval.*, pages 253–256.

I. Zariņa, P. Ņikiforovs, and R. Skadiņš. 2015. Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192.

S. V. Zasiekin. 2020. Psycholinguistic regularities of reproducing literary texts in translation (based on the English and Ukrainian languages).

F. Čermák and A. Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

8. Language Resource References

Brants, S., et al. 2023. *de_core_news_sm*.

Kotsyba, N., et al. 2016. *UD Ukrainian IU*. Institute for Ukrainian, NGO.

Kotsyba, N., et al. 2018. *https://mova.institute/*.

Kotsyba, N. and Turska, M. 2005 - 2011. *Polsko-ukrainski korpus rownolegly*.

Kurnosov, V., et al. 2022. *uk_core_news_sm*.

Nivre, J. et al. 2015 - 2024. *Universal Dependencies*.

Petrov, S., et al. 2023. *UD German GSD*.

Shvedova, M., et al. 2017 - 2024. *GRAC*.

Shvedova, M. and Lukashevskyi, A. 2023-2024. *ParaRook||DE-UK*.

Čermák, F. and Rosen, A. 2008 - 2023. *Intercorp*.

9. Appendix A: Example of parallel sentences with metadata

```
Ukrainian Text:
<doc authors_names_uk="Maks Friš"
    ↪ authors_names_de="Max Frisch"
    ↪ translators_names_uk="Jevhen
    ↪ Popovyč" translators_names_de="
    ↪ Jevhen Popovyč" authors_born
    ↪ ="1911" authors_sex="M"
    ↪ translators_born="1930"
    ↪ translators_sex="M"
    ↪ authors_regionCode="D-Z-CH"
    ↪ translators_regionCode="UA-C-CRK &
    ↪  UA-KYV-KYV" title_uk="Štiller"
    ↪ title_de="Stiller"
    ↪ original_language="DE" date_uk
    ↪ ="1968" pub_city_uk="Kyv"
    ↪ pub_year_uk="1970" publication_uk
    ↪ ="" url_uk="http://chtyvo.org.ua/"
    ↪  pub_city_de="Berlin" publisher_de
    ↪ ="Suhrkamp Verlag" publication_de
    ↪ ="" date_de="1954" url_de="library
    ↪ .lol/fiction/447
    ↪ EC7654424E50DD38BA58324825B29"
    ↪ orthography="sučasnyj pravopys"
    ↪ genre="" source="PRI" theme=""
    ↪ media="" style="FIC">
<align>
<s>
1 C’oho ce PRON Pd--nnsgn Animacy=Inan|
    ↪ Case=Gen|Gender=Neut|Number=Sing|
    ↪ PronType=Dem 2 obj _ _
2 vystačylo vystačyty VERB Vmeis-sn
    ↪ Aspect=Perf|Gender=Neut|Mood=Ind|
    ↪ Number=Sing|Tense=Past|VerbForm=
    ↪ Fin 0 root _ SpaceAfter=No
<g/>
3 . . PUNCT U _ 2 punct _ SpacesAfter=\r
    ↪ \n
</s>
<s>
1 Vin vin PRON Pp-3m-snn Case=Nom|Gender
    ↪ =Masc|Number=Sing|Person=3|
    ↪ PronType=Prs 2 nsubj _ _
2 zasmijavsja zasmijatysja VERB Vmeis-sm
    ↪  Aspect=Perf|Gender=Masc|Mood=Ind|
    ↪ Number=Sing|Tense=Past|VerbForm=
    ↪ Fin 0 root _ SpaceAfter=No
<g/>
3 . . PUNCT U _ 2 punct _ SpaceAfter=No
</s>
</align>
</doc>

German Text:
<doc authors_names_uk="Maks Friš"
    ↪ authors_names_de="Max Frisch"
    ↪ translators_names_uk="Jevhen
    ↪ Popovyč" translators_names_de="
    ↪ Jevhen Popovyč" authors_born
    ↪ ="1911" authors_sex="M"
    ↪ translators_born="1930"
    ↪ translators_sex="M"
    ↪ authors_regionCode="D-Z-CH"
    ↪ translators_regionCode="UA-KYV-KYV
    ↪  & UA-C-CRK" title_uk="Štiller"
    ↪ title_de="Stiller"
    ↪ original_language="DE" date_uk
    ↪ ="1968" pub_city_uk="Kyv"
    ↪ pub_year_uk="1970" publication_uk
    ↪ ="" url_uk="http://chtyvo.org.ua/"
    ↪  pub_city_de="Berlin" publisher_de
    ↪ ="Suhrkamp Verlag" publication_de
    ↪ ="" date_de="1954" url_de="library
    ↪ .lol/fiction/447
    ↪ EC7654424E50DD38BA58324825B29"
    ↪ orthography="sučasnyj pravopys"
    ↪ genre="" source="PRI" theme=""
    ↪ media="" style="FIC">
<align>
<s>
1 Das der PRON PDS Case=Nom|Gender=Neut|
    ↪ Number=Sing|PronType=Dem,Rel 2
    ↪ nsubj _ _
2 genügte genügen VERB VVFIN Mood=Ind|
    ↪ Number=Sing|Person=3|Tense=Past|
    ↪ VerbForm=Fin 0 root _ SpaceAfter=
    ↪ No
<g/>
3 . . PUNCT $. _ 2 punct _ SpacesAfter=\
    ↪ r\n
</s>
<s>
1 Er er PRON PPER Case=Nom|Gender=Masc|
    ↪ Number=Sing|Person=3|PronType=Prs
    ↪ 2 nsubj _ _
2 lachte lachen VERB VVFIN Mood=Ind|
    ↪ Number=Sing|Person=3|Tense=Past|
    ↪ VerbForm=Fin 0 root _ SpaceAfter=
    ↪ No
<g/>
3 . . PUNCT $. _ 2 punct _ SpaceAfter=No
</s>
</align>
</doc>
```

10. Appendix B: Corpus content and statistics

Author	Date (original)	Date (translation)	Original language	Title	Translator	Style (Fiction/Nonfiction/Ego-text)	Tokens (in the German-language part)
Alfred Döblin	1928	2020	DE	Berlin Alexanderplatz	Roman Osadčuk	FIC	200369
Andreas Kappeler	1995	2007	DE	Kleine Geschichte der Ukraine	Oleh Blaščuk	NOF	79781
Andreas Kappeler	2017	2018	DE	Ungleiche Brüder: Russen und Ukrainer vom Mittelalter bis zur Gegenwart	Volodymyr Kam'janec'	NOF	68549
Bernhard Kellermann	1913	1986	DE	Der Tunnel	Oleksa Lohvynenko	FIC	123092
Bernhard Schlink	1995	2016	DE	Der Vorleser	Petro Taraščuk	FIC	50080
Bertolt Brecht	1934	1973	DE	Dreigroschenroman	Jurij Lisnjak	FIC	142529
Bertolt Brecht	1943	1968	DE	Das Leben Des Galilei	Vasyl' Stus & Zinaïda Joffe	FIC	37205
Bertolt Brecht	1939	1973	DE	Mutter Courage und ihre Kinder	Marko Zisman	FIC	28711
Bohdan Scholdak	2000	1991	UK	Der Steinzeitmensch	Anna-Halja Horbatsch	FIC	2878
Christoph Ransmayr	1988	1992	DE	Die letzte Welt	Oleksa Lohvynenko	FIC	70374
Dieter Noll	1960	1965	DE	Die Abenteuer des Werner Holt. Roman einer Heimkehr. I	Jurij Myxajljuk	FIC	212573
Dieter Noll	1963	1965	DE	Die Abenteuer des Werner Holt. Roman einer Heimkehr. II	Jakiv Prylypko	FIC	188034
Elias Canetti	1935	2003	DE	Die Blendung	Oleksa Lohvynenko	FIC	228945
Erich Maria Remarque	1945	1986	DE	Arc de Triomphe	Jevhen Popovyč	FIC	185739
Erich Maria Remarque	1962	1963	DE	Die Nacht von Lissabon	Mykola Djatlenko & Arkadij Pljuto	FIC	93365
Erich Maria Remarque	1929	1986	DE	Im Westen nichts Neues	Kateryna Hlovac'ka	FIC	70858
Franz Kafka	1922	2006	DE	Das Schloss	Natalka Snjadanko	FIC	132233
Franz Kafka	1919	2012	DE	Brief an den Vater	Oleksa Lohvynenko	EGO	19412
Friedrich Dürrenmatt	1985	1987	DE	Justiz	Oleksa Lohvynenko	FIC	61489
Friedrich Dürrenmatt	1951	1989	DE	Der Richter und sein Henker	Kateryna Hlovac'ka	FIC	28828
Friedrich Glauser	1938	1994	DE	Wachtmeister Studer	Jurij Lisnjak	FIC	64239
Günter Grass	1959	2005	DE	Blechtrommel	Oleksa Lohvynenko	FIC	246103
Günter Grass	1961	2008	DE	Katz und Maus	Natalka Snjadanko	FIC	46170
Gustav Meyrink	1914	2011	DE	Der Golem	Natalja Ivanyčuk	FIC	88952
Heinrich Böll	1971	1972	DE	Gruppenbild mit Dame	Jevhen Popovyč & Jurij Lisnjak	FIC	164718
Heinrich Böll	1963	1965	DE	Ansichten eines Clowns	Mykola Djatlenko	FIC	92583
Heinrich Böll	1974	1989	DE	Die verlorene Ehre der Katharina Blum	Petro Sokolovs'kyj	FIC	36765
Heinrich Böll	1957	1989	DE	Irisches Tagebuch	Wladimir Schelest	FIC	35165
Heinrich Böll	1950	1969	DE	Wanderer, kommst du nach Spa…	Jevhenija Horeva	FIC	4001
Heinrich Mann	1935	1985	DE	Die Vollendung des Königs Henri Quatre	Jurij Lisnjak	FIC	321930
Heinrich Mann	1935	1975	DE	Die Jugend des Königs Henri Quatre	Jurij Lisnjak	FIC	244916
Heinrich Mann	1914	1969	DE	Der Untertan	Marko Zisman	FIC	163934
Hermann Hesse	1927	1977	DE	Steppenwolf	Jevhen Popovyč	FIC	81745
Ingrid Noll	1994	2019	DE	Die Apothekerin	Svjatoslav Zubčenko	FIC	62933
Joseph Roth	1930	2010	DE	Hiob	Jurij Proxas'ko	FIC	62317
Joseph Roth	1937	2010	DE	Das falsche Gewicht	Jurij Proxas'ko	FIC	41473
Joseph Roth	1939	2011	DE	Die Legende vom heiligen Trinker	Ol'ha Sydor	FIC	12842
Jurij Wynnytschuk	2000	1990	UK	Das Leuchten	Anna-Halja Horbatsch	FIC	12942
Lesja Ukrainka	2014	1900	UK	«Deine Briefe duften immer nach abgeblühten Rosen…»	Stanislaw Matijtschyn & Michael Beck	FIC	616
Lion Feuchtwanger	1929	1980	DE	Erfolg	Oleksa Synyčenko	FIC	308328
Martin Walser	1957	1975	DE	Ehen in Philippsburg	Jevhen Popovyč & Jarema Polotnjuk	FIC	116209
Max Frisch	1954	1968	DE	Stiller	Jevhen Popovyč	FIC	172306
Oleksander Denyssenko	2000	2000	UK	Die Seele des Flusses	Anna-Halja Horbatsch	FIC	3086
Otfried Preußler	1980	2006	DE	Krabat	Volodymyr Vasyljuk	FIC	73575
Patrick Süskind	1985	1993	DE	Das Parfum: Die Geschichte eines Mörders	Iryna Fridrix	FIC	89643
Patrick Süskind	1991	1995	DE	Die Geschichte von Herrn Sommer	Iryna Fridrix	FIC	19579
Patrick Süskind	1981	1996	DE	Der Kontrabaß	Iryna Fridrix	FIC	14387
Peter Handke	1976	1980	DE	Die linkshändige Frau	Oleksa Lohvynenko	FIC	23956
Robert Musil	1942	2010	DE	Der Mann ohne Eigenschaften. I.	Oleksa Lohvynenko	FIC	191592
Siegfried Lenz	1968	1976	DE	Deutschstunde	Oleksa Lohvynenko	FIC	193369
Stefan Zweig	1932	2017	DE	Marie Antoinette	Petro Taraščuk	FIC	183739
Stefan Zweig	1935	2018	DE	Maria Stuart	Petro Taraščuk	FIC	144602
Stefan Zweig	1929	2017	DE	Joseph Fouché	Petro Taraščuk	FIC	87781
Stefan Zweig	1911	1981	DE	Die Gouvernante	Iryna Stešenko	FIC	6512
Thomas Mann	1924	2008	DE	Der Zauberberg	Roman Osadčuk	FIC	375979
Thomas Mann	1900	1973	DE	Buddenbrooks	Jevhen Popovyč	FIC	281941
Thomas Mann	1954	2011	DE	Bekenntnisse des Hochstaplers Felix Krull	Roman Osadčuk	FIC	150007
Vasyl Barka	2007	1961	UK	Der gelbe Fürst	Maria Ostheim-Dzerowycz	FIC	135734