Creating a POS Gold Standard Corpus of Modern Ukrainian

Vasyl Starko
Ukrainian Catholic University
Ukraine
vstarko@gmail.com

Andriy Rysin
Independent researcher
USA
arysin@gmail.com

Abstract

This paper presents an ongoing project to create the Ukrainian Brown Corpus (BRUK), a disambiguated corpus of Modern Ukrainian. Inspired by and loosely based on the original Brown University corpus, BRUK contains one million words, spans 11 years (2010–2020), and represents edited written Ukrainian. Using stratified random sampling, we have selected fragments of texts from multiple sources to ensure maximum variety, fill nine predefined categories, and produce a balanced corpus. BRUK has been automatically POS-tagged with the help of our tools (a large morphological dictionary of Ukrainian and a tagger). A manually disambiguated and validated subset of BRUK (450,000 words) has been made available online. This gold standard, the biggest of its kind for Ukrainian, fills a critical need in the NLP ecosystem for this language. The ultimate goal is to produce a fully disambiguated one-million corpus of Modern Ukrainian.

1 Introduction

Ukrainian has a growing ecosystem of NLP datasets and tools. Still, it falls into the category of low-resource languages, despite increasing interest in the language and the development of multiple resources and tools over the past couple of years. Most general-purpose corpora that are available for Ukrainian, such as the General Regionally Annotated Corpus of Ukrainian (GRAC) by Shvedova et al. (2017-2023), Zvidusil by Kotsyba et al. (2018), and the Ukrainian Language Corpus (KUM) by Darchuk (2003-2023) and her team, are only accessible via a web user interface. Among downloadable Ukrainian corpora, one project that stands out here thanks to its size and thoroughness is UberText 2.0 by Chaplynskyi (2023). However, one of the missing resources is a reliable, balanced, and disambiguated corpus of sufficient size.

Until recently, the only such resource was the treebank created within the Universal Dependencies project by Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko.[1] With the overall size of some 120,000 tokens, it is comprised of fiction (24%), essays (8%), legal acts (7%), fairytales (7%), analytical articles (6.5%), news (6%), commentary (5%), textbooks (5%), Wikipedia articles (5%), scholarly works (4%), letters (3%), and some other types.[2] The creators made a laudable effort to include a wide variety of texts, and their resource has been invaluable for Ukrainian NLP. Nevertheless, some aspects require improvement. For one thing, the texts in the UD Ukrainian treebank come both from modern sources (past 15–20 years) and the first half of the 20th century, which does not make the entire treebank representative of any one period. Second, the proportions of text types are far from reflecting either the production or the consumption of texts in modern Ukrainian society. For example, news is significantly more popular than its share in this treebank would suggest. Third, a bigger corpus would help achieve better quality of NLP models. Furthermore, the small proportions of all types, except fiction, in the treebank complicate the task of training or fine-tuning models for a specific type. Finally, the development of this treebank seems to have come to a halt several years ago.

2 Corpus Design

Perceiving the need for a more balanced and larger disambiguated corpus, we have developed the Ukrainian Brown Corpus (BRUK)[3] modeled on the original Brown University corpus. The Brown University Standard Corpus of Present-Day American English (Francis and Kucera, 1979) has been an indispensable resource for the development of computational linguistics. It has given rise to

[1] https://universaldependencies.org/treebanks/uk_iu/index.html
[2] https://mova.institute/
[3] https://github.com/brown-uk/corpus

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 91–95

an entire family of Brown corpora, including the Lancaster-Oslo/Bergen Corpus (LOB) (Johansson, 1978), the Freiburg-Brown Corpus of American English (FROWN) (Hundt et al., 1998) and Freiburg-LOB Corpus of British English (FLOB) (Hinrichs et al., 2007) (Leech and Smith, 2005). Similar corpora have also been constructed for other languages (Koeva et al., 2006) and successfully used for training NLP models.

In order to establish the categorial structure of BRUK, we have used the same method of an expert poll with averaged results as did Henry Kucera and W. Nelson Francis and kept the overall split into informative and imaginative types. However, further subdivision into categories is different as it is aimed at reflecting the prevalence of each category of texts in modern Ukrainian society. This is in line with the established practice as corpora derived from the Brown University corpus include modifications on the original design and adjustments to account for the specific features of the language and country in question. The categories thus established for BRUK are as follows (percentages represent proportions of the total size):

A. Press, 25%. While BRUK has no formal subdivision into reportage, editorials, and reviews, a special effort has been made to represent these subcategories and ensure topical diversity (politics, society, finances, sports, culture, and environment). This category includes texts selected from national, regional, and local (city or district-level) mass media outlets in both printed and electronic form.
B. Religion, 3%. Importantly, texts representing different religions have been included.
C. Skills and Hobbies, 7%. Popular topics, such as household, crafts, farming, gardening, and construction, are represented.
D. Essays, Biography, Memoirs, etc., 7%. This is a catch-all category for informative texts that do not fit elsewhere, including forewords, personal letters, and literary and art criticism.
E. Administrative Documents, 3%. Laws, government regulations, reports, and official letters comprise this category.
F. Popular Science, 5%. Experts agreed that these texts required a separate category due to their linguistic characteristics.
G. Science, 10%. A balanced selection of texts in natural sciences and the humanities has been made.
H. Textbooks, 15%. This sizable category reflects the important role such texts play in Ukraine, where a wide audience of students reads them.
I. Fiction, 25%. While no formal subdivision has been adopted, variety is ensured by selecting works of different lengths (from short stories to novels) and genres.

In filling each category in the corpus with texts, we employed random sampling through crowdsourcing: more than a hundred individuals were involved in sample selection. Submitted samples were verified and filtered by corpus creators, for example, to remove duplicates and avoid overrepresentation of a particular newspaper, author, or topic.

Each text fragment in the corpus is supplied with metadata identifying the author(s), title, book/journal title (if applicable), place and year of publication, publisher, page range, length in tokens, orthography (official or alternative), and detected errors. Metadata information is stored separately from texts in a .csv file available for download and processing. Each file containing a text fragment is given a name that begins with a letter (A–I) for the respective category, enabling users to quickly separate the necessary category from the entire corpus.

3 Text Requirements

Texts in BRUK must meet a set of requirements, some of which mirror those for the original Brown University Corpus, while others represent a conscious departure from its model to match the realities of modern Ukrainian better:

1 Original (not translated) and human-written texts. The primary challenge here was to weed out texts surreptitiously translated from Russian (a common practice among some publishers and mass media outlets in Ukraine) and products of machine translation. In doubtful cases, we opted to err on the side of exclusion.

2 Edited prose only. Non-prose works, e.g., poems and drama pieces, are excluded, as are non-edited texts. In dubious cases, we rejected texts that clearly needed editing.

3 Written, rather than spoken, texts. BRUK generally represents written Ukrainian with only a sprinkle of “quasi-spoken” texts. Fiction may include dialogue, and some news articles contain interviews. Several texts selected for the corpus

were first spoken and then written down, such as public speeches and sermons.

4 Texts first published in 2010–2020. We excluded texts with the publication date within this period but written much earlier. The original Brown corpus represents one year. This narrow focus led to certain entities and topics being overrepresented, such as U.S. President John F. Kennedy and the tense U.S. relations with the Soviet Union before the Cuban Missile Crisis. For BRUK, we decided to draw samples from a longer period (11 years) in an effort to overcome this issue and ensure a better topical balance.

5 Texts published in mainland Ukraine. While diaspora texts are essential for the Ukrainian language, they are characterized by a number of divergencies in spelling, grammar, and lexis. They need to be collected in a separate corpus, which would make a valuable complement to BRUK.

6 Up to 2,000 words in total from one source. While the original Brown Corpus contained 500 continuous samples of text, each around 2,000 words long, BRUK is more fragmented as it is comprised of more fragments that are smaller in size. Most fragments contain less than 1,000 words of running text, and just a handful reach the 2,000-word mark. This approach has made it possible to include a greater variety of sources.

Detailed annotation guidelines[4] have been used by all contributors to BRUK.

4 POS tagging

4.1 Tools

BRUK has been automatically part-of-speech tagged using VESUM[5], a Large Electronic Dictionary of Ukrainian, and the TagText tagger for Ukrainian, part of the NLP UK toolkit for Ukrainian[6]. For proofreading the disambiguated part of BRUK, we used a modified Ukrainian module of LanguageTool[7], particularly its token agreement and case government rules. This allowed us to automatically detect a number of POS tagging errors that are hard to catch for human annotators. One of the determining factors in favor of these tools is that VESUM is the largest machine-readable morphological dictionary of Ukrainian. Its current version (6.1.1) comprises over 418,000 lemmas from which more than 6.5 million word-forms are generated. The dictionary achieves 97–99% word coverage on non-encyclopedic texts. Moreover, the TagText tagger includes a dynamic tagging component to recognize and tag words not found in VESUM, reaching 95% accuracy on these out-of-vocabulary items (Starko and Rysin, 2022). This combination of tools has been successfully utilized to tag successive iterations of GRAC, a large reference corpus of Ukrainian (Shvedova, 2020) (Starko et al., 2021). Second, unlike other morphological dictionaries of Ukrainian, VESUM includes numerous proper nouns and nonstandard lemmas, such as alternative spellings, slang, deprecated lexical items, dialectal words, and substandard word-forms, which are not to be found in other lexicographic resources. These linguistic items occur in modern texts and need to be duly recognized.

4.2 POS Tagset

BRUK has been tagged using the POS tagset of 21 tags, some of which are supplied by the VESUM dictionary and others assigned by TagText dynamically as it processes texts:

1 Inflection classes from VESUM: noun, verb, adj(ective), adv(erb), advp (adverbial participle), numr (numeral), conj(unction), prep(osition), part(icle), int(erjection), onomat(opoeic word), foreign (transliteration into Ukrainian), and noninfl(ected word that does not fit elsewhere).

2 Dynamic tags: number, date, time, hashtag, punct(uation), symb(ol), unknown (word written in Ukrainian letters but not recognized), and unclass (word that cannot belong to the Ukrainian lexicon, e.g., alphanumeric abbreviations, words in Latin script, non-Ukrainian words in Cyrillic, etc.).

Additional tags found in BRUK that describe, among other things, specific morphological features of Ukrainian words, such as case, number, and gender for nouns, can be looked up online[8].

[4] https://github.com/brown-uk/corpus/blob/master/doc/vymohy_do_frahmentiv.md

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

[6] https://github.com/brown-uk/nlp_uk

[7] https://github.com/languagetool-org/languagetool/tree/master/languagetool-language-modules/uk

[8] https://github.com/brown-uk/dict_uk/blob/master/doc/tags.txt

Texts tagged with the tools described above will contain part-of-speech ambiguity, with merely several hundred cases of ambiguity resolved automatically (Starko and Rysin, 2022). Thus, the next step in preparing BRUK was the manual disambiguation of automatically POS-tagged texts.

5 Disambiguation

Ukrainian is a highly inflected language with ubiquitous lexical and morphological ambiguity. In BRUK, an ambiguous word may have from 2 to over 30 homonymic readings.

As of this writing, ambiguity has been resolved for 450,000 Ukrainian words (560,000 tokens), making the disambiguated subset of BRUK the biggest such resource for Ukrainian. This part comprises 80,000 Ukrainian types and over 37,000 lemmas. Morphological ambiguity (58% of the words in the disambiguated subset of BRUK) is much more prevalent in Ukrainian than lexical ambiguity (13%), and a Ukrainian word has 2.88 homonym interpretations on average.

After automatic tokenization, lemmatization, and POS tagging (all performed by TagText), BRUK texts were subjected to a two-stage (in some cases, three-stage) disambiguation process. Initially, ambiguity was resolved by trained individuals (students), and these results were then verified by an expert linguist. Another expert was consulted in difficult cases. The nuances of tagging were communicated to students during training, and a number of challenging cases are explained in tagging guidelines[9]. The outcome of this process is a set of disambiguated texts in which each token has one correct and verified reading.

6 Conclusions and Future Work

The Ukrainian Brown Corpus (BRUK) is a one-million balanced corpus of modern Ukrainian covering 2010–2020. It is loosely modeled on the original Brown University corpus and consists of small fragments (mostly up to 1,000 but no longer than 2,000 words of running text) divided between 9 categories. The creators have made a concerted effort to ensure variety in the corpus along different dimensions. The corpus has been automatically tokenized, lemmatized, and POS-tagged. A subset of BRUK (450,000 words) has been manually disambiguated, validated through a multi-stage process, and made available for download.

Several factors make BRUK a unique resource compared to other Ukrainian corpora: it is a balanced downloadable corpus comprised of Modern Ukrainian text samples that vary along several dimensions and is currently the largest corpus representing a POS gold standard for Ukrainian. BRUK has the potential to become a key resource in solving the foundational problem of POS disambiguation for a wide variety of practical projects. Other applications are also possible, such as testing spellchecking systems, NER models, and so on. BRUK has been used to build a stochastic model for POS tagging, generating a strong baseline. On the theoretical side, BRUK provides insights into Ukrainian morphology that have already helped us improve its formal description for the purposes of NLP and computational linguistics research.

Our immediate plans include the semiautomatic disambiguation of the rest of BRUK (550,000 words). It is desirable to complement BRUK with later publications to cover a rapidly growing number of texts about the unprovoked war Russia unleashed against Ukraine on 24 February 2022. Further plans include adding a dependency annotation layer to the corpus.

Another line of activity is training language models. Even if trained on the released subset of BRUK rather than the entire corpus, they can be instrumental in solving various computational linguistics and NLP tasks, bringing the Ukrainian language a step closer to the status of a mid-resource language.

Limitations

No corpus is fully representative of the language in question. By design, BRUK represents only modern written Ukrainian focusing on edited texts. Even though BRUK includes texts referring to the COVID-19 pandemic, a separate collection may need to be added to better represent this widely discussed topic. Furthermore, new official orthographic rules for Ukrainian were introduced in mid-2019. The spelling novelties are reflected in BRUK texts published in 2019–2020, but their proportion is relatively small compared to the pre-2019 texts. Even though the orthographic changes are not drastic, it might be advisable to complement the corpus with more after-reform texts.

[9] https://github.com/brown-uk/corpus/blob/master/doc/skladni_momenty_tegiv.md

Ethics Statement

Our work aims to enrich the ecosystem of NLP resources and tools for the Ukrainian language. By making the BRUK corpus downloadable, we hope to stimulate research into Ukrainian both inside Ukraine and worldwide. The broader impact of our project lies in the fact that BRUK can be used to train Ukrainian language models and utilize them in various other NLP projects, specifically to tag and disambiguate much larger Ukrainian corpora.

Acknowledgements

This research has been supported by a grant from the Humanities Faculty of the Ukrainian Catholic University and a private donation. We thank Olha Havura, Nastia Osidach, Natalia Olishkevych, Natalia Cheilytko, Mariana Romanyshyn, and many other colleagues and UCU students who have contributed to BRUK.

References

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: a corpus of modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Nataliia Darchuk. 2003-2023. Korpus ukrayinskoyi movy.

Nelson W. Francis and Henry Kucera. 1979. BROWN CORPUS MANUAL MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.

Lars Hinrichs, Nicholas Smith, and Birgit Waibel. 2007. The part-of-speech-tagged 'Brown' corpora: a manual of information, including pointers for successful use.

Marianne Hundt, Andrea Sand, and Rainer Siemund. 1998. Manual of Information to Accompany the Freiburg-Brown Corpus of American English (FROWN).

Stig Johansson. 1978. Manual of Information to Accompany the Lancaster- Oslo/Bergen Corpus of British English (FROWN).

Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova, Ekaterina Tarpomanova, and Maria Todorova. 2006. Bulgarian Tagged Corpora. *In Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages*, pages 78–86.

Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko et al. 2018. Laboratoriya ukrayins'koyi.

Geoffrey Leech and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal*, 29:83–98.

Maria Shvedova. 2020. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality. In *International Conference on Computational Linguistics and Intelligent Systems*, pages 489–506, Lviv.

Maria Shvedova, Ruprecht von Waldenfels, Serhii Yaryhin, Andriy Rysin, Vasyl Starko, and Tymofij Nikolajenko et al. 2017-2023. GRAC: General Regionally Annotated Corpus of Ukrainian.

Vasyl Starko and Andriy Rysin. 2022. VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. In *Computational Linguistics and Intelligent Systems.*, volume 6th Int. Conf, pages 71–80, Gliwice. COLINS.

Vasyl Starko, Andriy Rysin, and Maria Shvedova. 2021. Ukrainian Text Preprocessing in GRAC. In *IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, pages 101–104, Lviv.