UDC 811.162.3'36
811.161.2'36
https://doi.org/10.18485/ms_zmss.2024.105.22

Jana Kockova
Institute of Slavonic Studies of the Czech Academy of Sciences
kockova@slu.cas.cz

Hanna Sytar
Institute of Slavonic Studies of the Czech Academy of Sciences
sytar@slu.cas.cz

THE MOST FREQUENT LEMMAS IN THE UKRAINIAN AND CZECH CORPUS AS A RESOURCE FOR FOREIGN LANGUAGE LEARNING AND TEACHING

The study examines the potential of language corpora in teaching foreign languages, specifically Czech and Ukrainian. By analysing the most common lemmas in both languages, we explore their potential for supplementing and creating teaching materials. The corpus data analysis can be used to add culturally and socially conditioned semantic groups, introduce productive word-formation models, and highlight differences in grammatical structures. The most frequent lemmas also provide suitable examples of various phenomena in language learning, such as frequently used abbreviations and proper names.

Key words: Language Corpora, Lemma, L2 Learning, Ukrainian, Czech.

Corpus material in foreign language teaching

Growing interest in cognitive science has led to a surge of scientific studies on foreign language learning, raising its importance in the contemporary world. A specific focus in research lies in the vocabulary necessary to attain a particular level of language proficiency. The establishment of standardized levels of foreign language proficiency was facilitated by the Council of Europe through the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (hereinafter CEFR), for the corresponding specification of CEFR for Czech see Hádková 2005, Čadská 2005, Holub 2005, and for the specifications for Ukrainian, refer to Стандартизовані 2018[1].

[1] For the Ukrainian language a detailed specification is still being created (Європейське 2008; Прокопчук/Гузар 2020; Вимоги 2021, etc.).

The impetus for conducting this research is the increased need for teaching materials for the Ukrainian learners of Czech. While most of the teaching materials work with thematically structured lexica (cf. Федонюк 2022; Anderš/Cholodová 2022; Holá/Bořilová 2011; 2014), this study aims to offer a different perspective on vocabulary and teaching materials and to demonstrate the possibilities of using the corpus approach in second language teaching. It is based on the most frequent lemmas represented in the given corpora. Along with semantic and grammatical parameters, frequency is employed to determine language units for CEFR, e. g. in Sohsah et al. 2015; Шведова 2022[2].

This study analyzes the most frequent lemmas in Czech and Ukrainian, obtained from the monolingual language corpora CNC and Grac (see below). Its primary focus is to delineate the principal lexical and morphological differences recorded in the lemma registers of both languages. Considering that the specific most frequent lemmas may vary depending on the corpus structure (see the Methodological Excursion section), our analysis primarily refers to phenomena extending beyond individual lemmas, emphasizing systemic aspects (e.g., word-formation models). These findings could potentially inspire the development of teaching materials. Our attention extends towards areas that are typically addressed intuitively in teaching materials, such as the selection of names and geographical terms, or those that are rather neglected, e.g., different uses of feminatives and patterns of word-formation. The texts within the corpora do not necessarily cover topics related to everyday life and casual conversation (e. g., shopping, traveling); instead, they offer insights into enduring social issues and societal discourse. Moreover, there are prominent differences between the list of lemmas obtained from a corpus of texts or compiled based on the selection of lexemes by experts (Pintard/François 2020: 87f.). In this respect, they provide a valuable resource for crafting teaching materials and represent an alternative educational resource, complementing traditional teaching materials.

Methodological excursion

The material for analysis includes 2000 most frequent lemmas obtained from the large corpus of modern Czech (syn v9 CNC, number of tokens: 5 692 729 712, number of lemmas: 7 375 002) and from the General Regionally Annotated Corpus of Ukrainian (Grac v.15, number of tokens: 889 097 859, number of lemmas: 4 089 165). Due to the predominantly post-1945 content in the Czech synchronous corpus, the search for data in the Ukrainian corpus is restricted to the period from 1945 to 2021, with some exceptions.

The Czech and Ukrainian corpora exhibit structural differences. In syn v9, journalistic texts predominate, while scientific literature and fiction form a smaller portion (Cvrček/Richterová 2023). In contrast, Grac v.15 includes all

[2] The CEFRLex project also created word lists for passive knowledge (receptive vocabulary lists) for different languages (including English: Dürlich/François 2018; French: François et al. 2014).

functional styles of Ukrainian, with literary texts accounting for more than half. Moreover, the corpus encompasses a variety of text types, including journalistic, scientific, official business, religious, and folklore texts, as well as ego-texts such as memoirs, letters, diaries, blogs, and autobiographies. A distinctive feature of the Ukrainian corpus is the inclusion of recorded verbal public and oral private messages. These corpora also differ in lemmatization and morphological indexing[3]. However, there are currently no larger corpora in this language combination that are comparable in both structure and size. The parallel corpus InterCorp is not suitable for such an analysis due to its composition (mostly fiction) and its size.

To assess the impact of style representation in the corpus on lemma frequency, we compared the most frequent lemmas extracted from the entire corpus with those from subcorpora of fiction, nonfiction, and journalistic texts. The Czech corpus SYN v9 showed the highest agreement between the 2000 most frequent lemmas obtained from the whole corpus and from the subcorpus of journalism (93.7 %), reflecting the significant presence of journalistic texts in SYN v9. Notable agreement was also found between the whole corpus and the subcorpus of non-fiction (73 %), while the lowest agreement was observed between the whole corpus and the subcorpus of fiction (55.7 %). It's crucial to note that this ratio reflects variations in lemma frequency based on text types. Similarly, in the Ukrainian corpus Grac, the highest agreement was found between the whole corpus and the corpus of journalistic texts (85.15 %). Furthermore, there was substantial agreement between the whole corpus and the corpus of scientific, official, religious texts, and so-called ego-texts (diaries, memoirs, letters, blogs, etc.) with the rate of 78.25 % (the Ukrainian subcorpus is most similar to the content of the Czech non-fiction subcorpus). In Ukrainian as well, the lowest level of agreement was found between the whole corpus and the corpus of fiction with the rate of 65 %.

Creating language learning materials using the corpus

National corpora are typically composed of written texts subjected to normalization and correction, with a prevalence of literary language. However, materials for language learning often focus on everyday speech patterns. Despite this, textbooks may not always provide sufficient vocabulary for comprehending news and official discourse in a particular country or culture.

To determine the number of units for analysis, we follow the volume established for A2 English language learners (Milton 2010: 224). Obtaining a comparable sample of the 'most frequent' lexical units poses challenges due to the dependence on corpus texts, lemmatization, and search tools. Nevertheless, despite

[3] These differences include especially the non-separation of reflexive and non-reflexive verbs in Czech, the presence of separate tags for adjectival and adverbial participles in the Ukrainian corpus, and separate annotation of the continuous forms of the verb *být* and conjunctions *aby* 'to', *kdyby* 'if' in Czech (*abys* 'to', *kdybychom* 'if we', etc.).

some conventionality in comparing the obtained lists, lemmas with high frequency in corpora serve as valuable guides for language learners and educational material authors. The most frequent lemmas not only help to specify required vocabulary for study but also act as keys to understanding public discourse, including mass media. Additionally, corpora enable the consideration of specific needs within a learner group. For instance, sub-corpora based on certain styles can be employed. The following section aims to demonstrate the potential of using lemma lists for foreign language teaching. This includes areas such as semantic groups, proper names, word-building models, abbreviations, gender, and feminitives in the language.

Semantic groups

Frequency lists serve multiple purposes, including identifying matches within discourses, recognizing gaps, and pinpointing culturally and historically stipulated areas. A significant portion of the lemma lists in both Czech (hereinafter — Cz.) and Ukrainian (hereinafter — Ukr.) languages coincide. Notably, both samples include nouns denoting **kinship**: Ukr./Cz. мати[4]/*matka* 'mother', дитина/*dítě* 'child', батько/*otec* 'father', син/*syn* 'son'etc.; **body parts**: голова/*hlava* 'head', рука/*ruka* 'hand', око/*oko* 'eye', etc.; **parts of the day**: день/*den* 'day', ніч/*noc* 'night', вечір/*večer* 'evening', etc.; **verbs associated with human activity**: *працювати/pracovat* 'to work', знати/*vědět* 'to know', казати/říkat 'to say, tell', etc. On the other hand, many lemmas, present in teaching materials for beginner levels (clothes and shoes, appliance names, etc.), do not show high frequency in language corpora.

However, there are also structural differences in vocabulary that are crucial for learners, identifiable through lemma list analysis. This includes gaps in vocabulary, such as the absence of a one-word equivalent for the Czech *odpoledne/dopoledne* 'p. m./a. m.' and *ročně* adv. 'yearly'and *loňský* adj. 'last year' in Ukrainian, or missing of the equivalent for the Ukrainian *доба* '24-hours' in Czech, etc[5]. Additionally, there are some partial structural differences to note; for example, the Czech equivalent *tam* 'there' corresponds to both Ukrainian expressions *там* 'there' and *туди* 'towards there', i.e. no distinction is made between the adverb of location 'there' and direction 'towards there' in Czech, etc.

Proper names

The selection of proper names in foreign language teaching often does not receive sufficient attention, yet it constitutes a significant part of vocabulary.

[4] The high frequency of *мати* is due to unremoved homonymy in the corpus: *мати* 'mother' as a noun ('mother') and *мати* 'have' as a verb.

[5] This does not mean, of course, that the languages are incapable of expressing a given meaning, but merely that they do not have a corresponding one-word expression. Cf. Czech: *Cesta trvá den*. 'The journey takes a day'; *Подорож триває добу*. 'The journey takes 24 hours.'; Cz. *Cesta trvá dva dny*., Ukr. *Подорож триває дві доби*. 'The journey takes 48 hours.'

This decision, of which names to include, should be grounded in the broader social discourse of the country. In this regard, corpora play a crucial role in guiding the selection of proper names for educational materials.

Proper names that are part of the list of the most frequent lemmas encompass a variety of entities, including names of individuals, cities, regions, countries, continents, rivers, and more. While the most common **personal proper names** may evolve over time, reflecting preferences in choosing names and also including those of famous politicians and cultural figures (for example, in Czech *Jan, Petr*, *Jiří*, as well as *Miloš* (the name of the former President and Prime Minister of the Czech Republic Zeman); in Ukrainian *Володимир* 'Volodymyr', *Іван* 'Ivan', *Олександр* 'Oleksandr', but also *Віктор* 'Viktor' (the name of the former Ukrainian Presidents — Yushchenko and Yanukovych), *Леонід* 'Leonid' (Kravchuk and Kuchma).

The list of 38 most frequent Czech names includes only seven female names, including Jana, Marie, Eva, etc. Similarly, in the Ukrainian samples, there are eight female names out of 36, such as (*Марія* 'Mariia', *Олександра* 'Oleksandra', *Юлія* 'Yuliia', *Люда* 'Liuda' (a short version of the name *Людмила* 'Liudmila') etc.

A notable characteristic of the Ukrainian onomasticon is the prominence of patronymics (Ukr. *по батькові*), traditional middle names derived from the father's first name. However, among the most frequent lemmas, only two male patronymic names, *Іванович* 'Ivanovich' and *Михайлович* 'Mykhailovich', are found.

Onyms in both languages represent well-known personalities, particularly politicians, writers, and artists, reflecting the social discourse. Examples include: Ukr. *Шевченко* 'Shevchenko'[6], *Янукович* 'Yanukovych', *Ющенко* 'Yushchenko' and *Порошенко* 'Poroshenko' (former Presidents of Ukraine), *Путін* 'Putin' (the President of Russia) and *Ленін* 'Lenin' (the chairman of the Council of People's Commissars of the USSR), Cz. *Sokol*[7], *Zeman* (the former President of the Czech Republic), *Dvořák* (composer), among others.

Geographical names

Geographical names constitute a relatively stable part of discourse. The lists of lemmas highlight differences in the most frequent names of states, continents, supra-state formations, and nationalities, partially reflecting the focus of the broader discourse in each language. Notably, neighboring states and nationalities dominate in both languages, such as *Evropa* 'Europe'*, Německo* 'Germany'*, USA, EU, Rusko* 'Russia'*, Slovensko* '*Slovakia'* etc. for Czech, and *Росія* 'Russia'*, Європа* 'Europe'*,*, *США* 'USA'*, Польща* 'Poland'*, Німеччина* 'Germany' etc. for Ukrainian.

[6] This lemma combines namesakes: the outstanding Ukrainian writer and artist Taras Shevchenko, the famous football player Andrii Shevchenko and ordinary citizens bearing this common surname.

[7] Besides the surname, a sports organization is called like that.

In the Czech material, there are three different names for the territory of the Czech Republic: *Česko, Čechy* (where the second mentioned word is used either for the name of the country or the region of Bohemia), and the region *Morava*; The Ukrainian list includes three regions *Крим* ‘Crimea’, *Донбас* ‘Donbas’ and *Галичина* ‘Halychyna’, along with the historical term *СРСР* ‘USSR’. The most frequent lemmas also include the corresponding adjectives of geographical names. Notably frequent are adjectives of city names. Hence, considering their high representativeness, it is advisable, for instance, to include the possessive adjective *Karlův* in the passive vocabulary when teaching Czech (*Karlův most* ‘Charles Bridge’, city *Karlovy Vary*, *Univerzita Karlova* ‘Charles University’, etc.) or proper names with the words *Ústí* (town *Ústí nad Orlicí, Ústí nad Labem,* etc.*)* or *Hradec* (town: *Jindřichův Hradec, Hradec Králové,* etc.) that appear in a number of place names.

Specific semantic areas

In addition to shared areas, the comparison of the most frequent lemmas reveals semantic groups reflecting specific discourse of each language area. One notable difference is the high frequency of words related to **war, peace**, military affairs, and revolution in Ukrainian, including Ukr. *війна* ‘war’, *військовий* ‘military’, *армія* ‘army’, *військо* ‘army’, *революція* ‘revolution’, *бій* ‘battle’, *фронт* ‘front’, *бойовий* ‘combat’, *солдат* ‘soldier’, *мир* ‘peace’, *капітан* ‘caption’, *командир* ‘commander’, *революційний* ‘revolutionary’*, Майдан*[8] ‘*Maidan*’ etc. Notably, such vocabulary is found not only in texts related to the Russian-Ukrainian war (since 2014), but also in other contexts. In Czech, only a few words represent this realm, such as *válka* ‘war’, *vojenský* ‘military’, *armáda* ‘army’, *voják* ‘soldier’.

Another distinctive feature of the Ukrainian list is the prevalence of religious vocabulary, which is high across different styles. Religion-related words and phrases are among the most frequent lemmas in non-fiction and journalistic subcorpora, including Ukr. *бог* ‘god’, *душа* ‘soul’, *церква* ‘church’, *святий* ‘saint’, *дух* ‘spirit’, *божий* ‘divine’, *віра* ‘faith’, *духовний* ‘spiritual’, *Господь* ‘God’, *храм* ‘temple’, *православний* ‘orthodox christian’, *гріх* ‘sin’. The Czech register only includes the lemmas *kostel* ‘kostel’, *svatý* ‘saint’ and *duch* ‘spirit’.

Historicisms denoting the social status of people in ancient times are present in the Ukrainian material, including Ukr. *князь* ‘duke’, *козак* ‘cossack, *цар* ‘tsar’, etc.; sovietisms, productive as a result of a rather long period of Ukraine as a part of the Soviet Union (*радянський* ‘soviet’, *СРСР* ‘USSR’, *колгосп* ‘kolgosp’ (collective farm), *комуніст* ‘komunist’, *міліція* ‘militia, police’ (compare now — *поліція* ‘police’). Beside this, there occurred a concurrency of some Russianisms, c. f. *головуючий* and *глава* ‘chairperson’ competing with the Ukrainian normative unit *голова* ‘chairperson, leader, head’ (which is homonymous with голова ‘head’).

[8] Capitalised, refers to the mass protest against the authorities in 2004.

Certain differences in the thematic representation of the most frequent lemmas could be attributed to the different structure of the two corpora. For instance, the significant presence of journalism in the main Czech corpus influences the frequency of words related to sports in Czech, such as *utkání* ‘match’, *trenér* ‘coach’, *liga* ‘league’, *gól* ‘goal’, *branka* ‘goal’. The vocabulary related to sport includes the names of sports clubs and organizations (*Sparta*, *Slavie, Sokol).* However, the only exceptions, which are also represented among the 2000 most frequent lemmas in fiction, are the lemmas *klub* ‘club’ and *vyhrát* ‘win’. Moreover, many words from this lexical area are also used in other areas of discourse, often appearing in intersecting areas like politics, economics, and sports, such as Cz. *soutěž* ‘competition’ (sports, politics, economics), *vyhrát* ‘win’ (elections, tender, match), *finále* ‘finale’ (championship, elections, the end in any sense). The sports vocabulary is also represented in the Ukrainian register of the most frequent lemmas, however, with lower number of lemmas: *матч* ‘match’, *спорт* ‘sport’, *спортивний* ‘sportive’, *чемпіонат* ‘championship’, *klub* ‘club’. Most of these sports-related words are found in Ukrainian journalistic texts, with only a small proportion also appearing in the corpus of fiction.

Gender

Ukrainian and Czech demonstrate differences in the productivity and frequency of using male and female names for professions and positions. Feminitives are actively used in both languages, such as Cz. / Ukr. *ředitelka/директорка* ‘directress’, *lékařka, doktorka/лікарка* ‘a female doctor’, *spisovatelka/письменниця* ‘a female writer’. However, some Czech forms lack normative equivalents in modern Ukrainian, leading to potential errors in language learning (compare Cz. *svědkyně* ‘a female witness’, *kapitánka* ‘a female caption’, Ukr.: *свідок* ‘witness’, *капітан* ‘caption’ which can denote both, a man and a woman).[9] This linguistic difference poses challenges for learners, especially considering the ongoing development and change in the Ukrainian language.

The use of feminine nouns in Ukrainian has a long, albeit interrupted, tradition. (cf. Уманець/Спілка 1893–1898; Сулима 1928: 11–12; СУМ 1907–1909; Дорошенко та ін. 1930; РУМ 1924–1933;). Words denoting professions or positions were intentionally phased out of active use during Ukraine’s period within the borders of the USSR. This trend was driven by a general tendency to educate citizens without emphasising gender affiliation and to bring Ukrainian and Russian languages closer together (many feminitives in Russian are limited by the colloquial speech, often with the estimated meaning: *врачиха* ‘a female doctor’, *директриса* ‘directress’, *профессорша* ‘a female professor’ and others). The be-

[9] Czech is remarkable for the only form of the masculine *host* ‘guest, visitor’, while Ukrainian uses distinct words to show gender distinction: masculine *гість* ‘guest, visitor’ and feminine *гостя* ‘a female guest’. Nevertheless, the feminitive *hostka* has recently appeared in Czech. It is currently non-normative and may be unacceptable to some users.

ginning of the 21st century has witnessed a revival of feminitives in Ukraine. Many of them are not yet standardized, and different word-forming models often compete, reflecting the evolving nature of the language: Ukr. *колега — колегиня*, *колежанка* ‘colleague — a female colleague’; *психіатр — п*сихіатриня, психіатерка, психіаторка ‘psychiatrist — a female psychiatrist’ (cf. also Плачинда 2018; Синчак 2022).

However, despite the significant word-formation potential and tradition of using feminitives in Czech, the frequency list shows that only two feminitives ředitelka ‘directress’ and *herečka* ‘actress’ are among the 2000 most frequent Czech lemmas, and none in Ukrainian. Male names of professions, on the other hand, have high frequency in both languages, including Cz. *policista* ‘policeman’, *řidič* ‘driver’, *prezident* ‘president’; Ukr. *депутат* ‘deputy’, *міністр* ‘minister’, *політик* ‘politician’, etc. The low frequency of Czech feminitives primarily reflects the influence of public discourse in journalistic texts, which make up a substantial part of the corpus, and the use of masculine forms as a generic form for both genders[10], especially in plural and ambiguous situations where the gender of the occupant is unclear. Additionally, in Czech words with grammatical feminine gender, such as *osoba* ‘person’, *osobnost* ‘personality’, *postava* ‘character’, *oběť* ‘victim’, are frequently used to refer to both genders, with no masculine analogues for these words.

Abbreviations

Finally, the registers of lemmas include abbreviations frequently used in journalistic texts. Knowledge of abbreviations is an important prerequisite for comprehending public discourse, not only for understanding texts in mass media but often for local orientation and understanding administrative requirements (forms, etc.). For this reason, we consider it appropriate to include this material in textbooks for learners of a foreign language.

High-frequency abbreviations include units of weights and measures, which are internationally defined, they do not differ in general: Ukr./Cz. *м/m* — metre*, км/km* — kilometre*, г/g* — gram. On the other hand, students should be made aware of the slight differences in some of the abbreviations: Ukr. *млн.*/Cz. *mil. — million,* Ukr. *га*/Cz. *ha — hectare,* Cz. *h* (hodin) — hour. Additionally, there are abbreviations used similarly in both languages; however, their meanings may not be obvious, often belonging to administrative terms. The Ukrainian list of lemmas has a significantly higher number of abbreviations than the Czech one: Ukr. *м.* (місто) — city, *с.* (село) — village, *грн.* (гривня) — UAH hryvnia; *ім. (імені)* — named in honour of smb., *р. (рік)* — year*, рр. (роки)* — years*;* Cz. *sv.* (svatý) — saint, *n.* (nad) — ‘on the river’ (both often in proper names), *č.* (číslo) — number, *Kč* (koruna česká) — CZK Czech crown, etc.

[10] E.g., *Diváci se mohou těšit na milostné písně od sólových zpěváků i kapel.* ‘Viewers can enjoy love songs performed by solo singers and bands.’

Abbreviations of states and international institutions are marked with a high frequency: Ukr. / Cz. *США/USA* — United States of America*, ЄС/EU* — European Union. Here, we can also find some differences between two languages, for instance, some abbreviations with higher frequency are found in Ukrainian: *СРСР* — Union of Soviet Socialist Republics, *НАТО* — North Atlantic Treaty Organization, *РФ* — Russian Federation. In addition, a number of specific abbreviations are used in Ukrainian, especially in the journalistic and administrative texts; many of them are also included in the list of the most frequent lemmas. These are mostly cases where the use of abbreviations is not common in Czech. Knowledge of these abbreviations is essential for understanding news in Ukrainian mass media: *ВЗ* (військовий збір) — military tax and the name of the newspaper "Wysoki Zamok", *СБУ* (Служба безпеки України) — Security Service of Ukraine, *ДТП* (дорожньо-транспортна пригода) — road traffic accident, *УПА* (Українська повстанська армія) — Ukrainian Insurgent Army, *Кабмін* (Кабінет Міністрів України) — Cabinet of Ministers of Ukraine, *МВС* (Міністерство внутрішніх справ України) — Ministry of Internal Affairs of Ukraine, *МЗС* (Міністерство закордонних справ України) — Ministry of Foreign Affairs of Ukraine.

In contrast to Ukrainian, Czech political parties are often abbreviated: *ODS* (Občanská demokratická strana) — Civil democratic Party, *ČSSD* (Česká strana sociálně demokratická) — Czech Social Democratic Party. Furthermore, sports organizations in Czech also frequently use abbreviations: *TJ* (Tělovýchovná jednota) — sports unity, *FC* (fotbalový klub) — football club, *SK*[11] (sportovní klub) — sports club.

Word formation types

The list of the most frequent lemmas can also be used to provide information about the frequently used word-formation types[12], and it also provides the corresponding frequently used examples. For example, we can compare the word-formation types of agent nouns in Czech and Ukrainian. In Czech, the suffix *-el* and *-ík/-ník* prevails (*obyvatel* 'inhabitant', *ředitel* 'director'; *návštěvník* 'visitor', *útočník* 'invader', etc.), less frequent are suffixes *-or (autor* 'author'*, primátor* 'mayor'), *-ec* (*herec* 'actor', *poslanec* 'member of the parliament'), *-ář* (*čtenář* 'reader'*, novinář* 'journalist'). Feminitives are formed most often with the sufix *-ka* (*manželka* 'wife'*, ředitelka* 'directress'). The productive suffixes in Ukrainian include: *-ник* (*робітник* 'worker', *письменник* 'writer')*, -ик* (*політик* 'politician', історик 'historian'), *-тель* (*вчитель* 'teacher', *житель* 'resident'), *-ець (виборець* 'voter', *фахівець* 'specialist'), *-ар (лікар* 'doctor', *секретар* 'secretary'), *-ист/-іст* (*фінансист* 'finance specialist', *журналіст* 'journalist'), etc.

[11] This abbreviation can also stand for Slovakia.

[12] For Ukrainian, see Карпіловська 2000, Полюга 2001; for Czech, SAUČ 2018.

Specifical grammatical differences

The conducted analysis of the parts of speech ratio in the Czech and Ukrainian registers has not revealed any significant differences. In particular, the number of verbs in both lists is almost identical (378 Czech and 355 Ukrainian lemmas). Both languages have the verb *být* / *бути* 'to be' and the verb *mít* / *мати* 'to have'. However, in the Ukrainian, there are two constructions for expressing 'to have': *я маю — у мене є* 'I have got'.

It should be noted that Czech verbs have higher frequencies, and *být* 'to be' is the absolute leader in frequency. In contrast, the first 30 positions in the Ukrainian register are occupied by synsemantic words *і* 'and', *на* 'on', *не* 'not', *в* 'in', *що* 'what' among others. The higher frequency of the Czech verb *být* 'to be' is due to the systemic difference between Czech and Ukrainian — in Czech the verb 'to be' is used as an auxiliary verb in the past tense, conditional mood and passive voice (cf. past tense, Cz. *četl jsi*, Ukr. ти читав 'you read'; subjunctive mood, Cz. četl *bych*, Ukr. *(я)* читав би, passive voice, Cz. *je pochválen*, Ukr. *(він є) похвалений*). Furthermore, the higher frequency of the Ukrainian personal pronouns (він 'he', *я* 'I', вона 'she') is structurally stipulated, it shows different grammatical status in both languages (Cz. *píšu* — Ukr. *(я)* пишу).

In addition, the comparison of the lemmas lists demonstrates some systematic spelling differences. For example, the high frequency of the negative *не* 'not' in Ukrainian shows an important, though formal, difference between the languages — the negation is written separately from the verbs. The high frequency of Czech *se* (reflexive and passive element) is caused by the same reason (cf. *nestaví se — не будується*).

Modal expressions

Significant differences are revealed among modal predicates. There are only eight modal verbs, three modal predicative adverbs and three modal adjectives among the most frequent lemmas in Czech (*třeba* 'perhaps', *lze* 'possible'; *možný* 'possible', *nutný* 'necessary', etc.). On the contrary, Ukrainian is much more abundant in modal verbs, modal adjectives and adverbs. In addition, the Ukrainian language is distinguished by the presence of aspect pairs of modal verbs. However, there is only one aspect pair of verbs among the most frequent lemmas: могти – змогти 'can'. Other perfect counterparts have lower frequency. The different status of the modal verbs in Czech (compared to Ukrainian) is likewise reflected in their higher frequency: Cz. *moci* 'can', *chtít* 'want', *muset* 'have to' occur among the first 100 lemmas in Czech. Nevertheless, the lower frequency of the Ukrainian modal verbs is caused by higher number of them[13] and the occurrence of aspectual pairs of a modal verb. Another significant difference is the greater number and high frequency of the Ukrainian predicative adjectives:

[13] Cf.: 'can': *могти, змогти*, 'want': *хотіти, хтіти* (colloquial), *бажати.*

повинний ‘obliged’, потрібний ‘needed’, необхідний ‘necessary’, etc.; and modal predicative adverbs: e.g. можна ‘allowedly’, треба ‘required’, потрібно ‘necessarily’, etc.

Conclusion

Language corpora can be used in various ways in foreign language teaching. Corpus data provides important and engaging material for comprehending and modifying the lexical minimum compiled by expert linguists and teachers. This research explores the possibilities of using corpus for the teaching and learning languages, in particular, description of typical combinations of words, dynamics of language changes, variability, regional attachment, etc.

In particular, it can be useful for a better understanding of the public discourse. The frequency list of lemmas can be used not only as an aid in teaching vocabulary and constructing the lexical structure of textbooks. In addition, it reflects a number of important structural differences of the languages. It can serve as an alternative resource for specifically targeted teaching, as well as allowing you to create materials aimed at the specific combination of source and target language.

An overview of potential ways of using the list of lemmas for the purposes of foreign language teaching is provided. In particular, lemmas with high frequency provide relevant frequent examples for certain semantic groups. In addition, analyzing lists of the most frequent lemmas can reveal significant **differences in the semantic structure** of the two languages (e. g., non-existent equivalents — доба in Ukr., meaning shifts *tam* in Cz. — *там, туди* in Ukr., etc.). The frequency lists not only provide often used examples of **proper names and surnames**, but also of **geographical** names that regularly appear in public discourse. Moreover, the analysis uncovered disparities in the frequency of certain semantic groups. For instance, Ukrainian had a high frequency of words related to **military and religious** topics, while Czech had a high frequency of **sports-related** words. In addition, the analysis showed a low frequency of **feminitives** in general discourse, not only in Ukrainian, where it could be expected, but also in Czech.

Corpora can ensure an important source of **abbreviations**, which are necessary for understanding messages in the mass media in particular. Here, significant differences between Czech and Ukrainian discourse emerge as well.

In addition to aforementioned areas, lemma lists provide basic information on **frequently used word-formation types**, supply suitable and frequent examples. Furthermore, the conducted analysis reveals some structural and grammatical differences between the languages, such as the high frequency of the Czech verb ‘to be’, which reflects its wider use in Czech, or the high number of modal devices (modal verbs, modal adjectives and adverbs) in Ukrainian.

REFERENCES

Anderš Josef, Cholodová Uljana. *Ukrajinština vážně i vesele*. Olomouc: Univerzita Palackého v Olomouci, 2022.

CEFR: *Common European Framework of Reference for Languages: Learning, teaching, assessment* — Companion volume, Council of Europe Publishing, Strasbourg, available at. 2020. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4> 01.03.2023.

Cvrček Václav, Richterová Olga. (eds.). SYN, version 9. Příručka ČNK. 9.06.2022. <https://wiki.korpus.cz/doku.php/en:cnk:syn:verze9.> 10.03.2023.

Čadská Milada et all. *Čeština jako cizí jazyk*. Úroveň A2. MŠMT: Tauris, 2005.

Dürlich Luise, François Thomas. "EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language". *Proceedings of LREC*. 2018: 873–879.

François Thomas, Gala Nùria, Watrin Patrick, Fairon Cédrick. "FLELex: a graded lexical resource for French foreign learners". *Proceedings of LREC*. 2014: 3766–3773.

Hádková Marie et all. *Čeština jako cizí jazyk*. Úroveň A1. Praha: MŠMT, 2005.

Holá Lída, Bořilová Pavla. *Čeština expres 1* (úroveň A1/1). Praha: Akropolis, 2011.

Holá Lída, Bořilová Pavla. *Čeština expres 3* (úroveň A2/1). Praha: Akropolis, 2014.

Holub Jan. *Čeština jako cizí jazyk*. Úroveň A2. MŠMT: Tauris, 2005.

Milton James. "The development of vocabulary breadth across the CEFR levels". I. Vedder — I. Bartning — M. Martin (eds.): *Communicative proficiency and linguistic development: intersections between SLA and language testing research. Second Language Acquisition and Testing in Europe* Monograph Series 1. 2010: 211–232.

Pintard Alice, François Thomas. "Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words". *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, Marseille, France. European Language Resources Association. 2020: 85–92.

SAUČ: *Slovník afixů užívaných v češtině*. Josef Šimandl (ed.). <http://www.slovnikafixu.cz/jak_slovnik_vyuzivat.> 15.03.2023.

Sohsah Gihad N., Ünal Muhammed Esad, Güzey Onur. "Classification of word levels". *Br J Educ Technol* 46 (2015): 1097–1101. <https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12338.>. 01.03.2023.

Вимоги: *Вимоги до рівнів володіння державною мовою*. Затверджено Рішенням Національної комісії зі стандартів державної мови 24.06.2021 року № 31. <https://zakon.rada.gov.ua/laws/show/z0925-21#Text.> 01.03.2023.

Grac: ГРАК. *Генеральний регіонально анотований корпус української мови (2017–2022)*. М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко, Т. Ніколаєнко та ін. Київ, Львів, Єна. <uacorpus.org.> 10.03.2023.

Дорошенко Д., Station Станіславський М., Страшкевич В. *Російсько-український словник ділової мови*. Київ — Харків, 1930. <https://archive.org/details/dilovoimovy.> 15.03.2023.

Європейське мовне портфоліо: Методичне видання / Уклад. О. Карп'юк. Тернопіль: Лібра Терра, 2008.

Карпіловська Є. А. *Суфіксальна підсистема сучасної української літературної мови: будова та реалізація:* Дисертація д-ра філол. наук. Київ, 2000.

Плачинда Галина. *Словничок фемінітивів для прес-офіцерів та прес-офіцерок територіальних управлінь Державної служби України з надзвичайних ситуацій*. Київ, 2018.

Полюга Лев. *Словник українських морфем*: близько 40 000 слів. Львів: Світ, 2001.

Прокопчук Надія, Гузар Олена. *Проєкт оцінювання рівня української мови*. 2020. <https://pcuh.stmcollege.ca/wp-content/uploads/2021/03/UKR-CEFR-LP-Teacher-Guide_2021-Final.pdf>. 01.03.2023.

РУМ: *Російсько-український словник*. Гол. ред. А. Кримський. Т. I–IV. Харків: Червоний шлях, 1924–1933.

Синчак Олена. *Вебсловник жіночих назв української мови*. Львів, 2022. <https://r2u.org.ua/html/femin_details.html/>. 21.03.2023.

Стандартизовані: *Стандартизовані вимоги до рівнів володіння українською мовою як іноземною А1–С2*. Укладачі: Мазурик Данута, Антонів Олександра, Синчак Олена, Бойко Галина. Схвалено рішенням колегії Міністерства освіти і науки України протокол від 22.05.2018 № 5/1–7. Львів, 2018.

Сулима Микола. *Українська фраза*. Харків, 1928.

СУМ: *Словарь української мови:* в 4-х т./За ред. Б. Грінченка. Київ. 1907–1909.

Уманець М., Спілка А. *Словарь росийсько-український.* Львів: НТШ. 1893–1898. <https://archive.org/details/slovnikII/page/n1/mode/1up>. 21.03.2023.

Федонюк Валентина. *2000 найкорисніших чеських слів і висловів*. Київ: Арій. 2022.

Шведова Марія. «Застосування корпусу у викладанні української мови як іноземної». 斯拉夫语言、文化、翻译与教学：现状与前景 *[Слов'янські мови, культури, переклад і викладання: сучасний стан і перспективи].*长春 [Чанчунь], 2022: 227–232.

REFERENCES

Anderš Josef, Cholodová Uljana. *Ukrajinština vážně i vesele*. Olomouc: Univerzita Palackého v Olomouci, 2022.

CEFR: *Common European Framework of Reference for Languages: Learning, teaching, assessment* — Companion volume, Council of Europe Publishing, Strasbourg, available at. 2020. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4> 01.03.2023.

Cvrček Václav, Richterová Olga. (eds.). SYN, version 9. Příručka ČNK. 9.06.2022. <https://wiki.korpus.cz/doku.php/en:cnk:syn:verze9.> 10.03.2023.

Čadská Milada et all. *Čeština jako cizí jazyk*. Úroveň A2. MŠMT: Tauris, 2005.

Doroshenko D., Stanislavskyi M., Strashkevych V. *Rosiisko-ukrainskyi slovnyk dilovoi movy.* Kyiv — Kharkiv, 1930. <https://archive.org/details/dilovoimovy.> 15.03.2023.

Dürlich Luise, François Thomas. "EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language". *Proceedings of LREC* (2018): 873–879.

Fedoniuk Valentyna. *2000 naikorysnishykh cheskykh sliv i vysloviv*. Kyiv: Arii, 2022.

François Thomas, Gala Nùria, Watrin Patrick, Fairon Cédrick. "FLELex: a graded lexical resource for French foreign learners". *Proceedings of LREC* (2014): 3766–3773.

Grac: *Heneralnyi rehionalno anotovanyi korpus ukrainskoyi movy (2017–2023).* M. Shvedova, R. Von Waldenfels, S. Yaryhin, A. Rysin, V. Starko, T. Nikolaienko, ta in. Kyiv, Lviv, Jena. <uacorpus.org.> 10.03.2023.

Hádková Marie et all. *Čeština jako cizí jazyk*. Úroveň A1. Praha: MŠMT, 2005.

Holá Lída, Bořilová Pavla. *Čeština expres 1* (úroveň A1/1). Praha: Akropolis, 2011.

Holá Lída, Bořilová Pavla. *Čeština expres 3* (úroveň A2/1). Praha: Akropolis, 2014.

Holub Jan. *Čeština jako cizí jazyk*. Úroveň A2. MŠMT: Tauris, 2005.

Karpilovska Ye. A. *Sufiksalna pidsystema suchasnoyi ukrayinskoyi literaturnoyi movy: budova ta realizaciya:* Dysertaciya doktora filol. nauk. Kyiv, 2000.

Milton James. "The development of vocabulary breadth across the CEFR levels". I. Vedder — I. Bartning — M. Martin (eds.): *Communicative proficiency and linguistic development: intersections between SLA and language testing research. Second Language Acquisition and Testing in Europe* Monograph Series 1 2010: 211–232.

Pintard Alice, François Thomas. "Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words". *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READing DIfficulties (READI)*, Marseille, France. European Language Resources Association (2020): 85–92.

Plachynda Halyna. *Slovnychok feminityviv dlia pres-ofitseriv ta pres-ofitserok terytorialnykh upravlin Derzhavnoi sluzhby Ukrainy z nadzvychainykh sytuatsii*. Kyiv, 2018.

Poliuha Lev. *Slovnyk ukrainskykh morfem*: blyzko 40 000 sliv. Lviv: Svit, 2001.

Prokopchuk Nadiia, Huzar Olena. *Proekt otsiniuvannia rivnia ukrayinskoyi movy*, 2020. <https://pcuh.stmcollege.ca/wp-content/uploads/2021/03/UKR-CEFR-LP-Teacher-Guide_2021-Final.pdf>. 01.03.2023.

RUS: *Rosiisko-ukrayinskyi slovnyk*. Hol. red. A. Krymskyi. T. I–IV. Kharkiv: Chervonyi shliakh, 1924–1933.

SAUČ: *Slovník afixů užívaných v češtině*. Josef Šimandl (ed.). <http://www.slovnikafixu.cz/jak_slovnik_vyuzivat.> 15.03.2023.

Shvedova Maria. "Zastosuvannia korpusu u vykladanni ukrainskoyi movy yak inozemnoyi [Application of the corpus in teaching Ukrainian as a foreign language]". 斯拉夫语言、文化、翻译与教学：现状与前景 [Slovianski movy, kultury, pereklad i vykladannia: suchasnyi stan i perspektyvy]. 长春 [Chanchun], 2022: 227–232.

Sohsah Gihad N., Ünal Muhammed Esad, Güzey Onur. "Classification of word levels". *Br J Educ Technol* 46 (2015): 1097–1101. <https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12338.>. 01.03.2023.

SRU: Umanets M., Spilka A. *Slovar rosyisko-ukrainskyi*. Lviv: NTSH, 1893–1898. <https://archive.org/details/slovnikII/page/n1/mode/1up>. 21.03.2023.

Standartyzovani vymohy do rivniv volodinnia ukrainskoyu movoyu yak inozemnoyu A1-C2. Ukladachi: Mazuryk Danuta, Antoniv Oleksandra, Synchak Olena, Boiko Halyna. Skhvaleno rishenniam kolehii Ministerstva osvity i nauky Ukrainy protokol vid 22.05.2018 # 5/1–7. Lviv, 2018.

Sulyma Mykola. *Ukrainska fraza*. Kharkiv, 1928.

SUM: *Slovar ukrainskoi movy*: v 4-kh t./Za red. Borysa Hrinchenka. Kyiv. 1907–1909.

Synchak Olena. *Vebslovnyk zhinochykh nazv ukrainskoyi movy.* Lviv, 2022: <https://r2u.org.ua/html/femin_details.html/>. 21.03.2023.

Vymohy do rivniv volodinnia derzhavnoyu movoyu. Zatverdzheno Rishenniam Natsionalnoi komisii zi standartiv derzhavnoyi movy 24.06.2021 roku No. 31. <https://zakon.rada.gov.ua/laws/show/z0925-21#Text.> 01.03.2023.

Yevropeiske movne portfolio: Metodychne vydannia/Uklad. O. Karpiuk. Ternopil: Libra Terra, 2008.

Јана Коцкова, Хана Ситар

НАЈЧЕШЋЕ ЛЕМЕ У УКРАЈИНСКОМ И ЧЕШКОМ КОРПУСУ КАО ИЗВОР ЗА УСВАЈАЊЕ И НАСТАВУ СТРАНОГ ЈЕЗИКА

Резиме

Рад испитује потенцијал језичких корпуса у настави страних језика, посебно чешког и украјинског. Анализом најчешће леме у оба језика истражујемо њихов потенцијал за допуњавање и осмишљавање наставних материјала. Корпусна анализа података може се користити за додавање културно и друштвено условљених семантичких група, увођење продуктивних модела творбе речи и истицање разлика у граматичким конструкцијама. Најчешће леме такође пружају одговарајуће примере различитих појава у усвајању језика, као што су често коришћене скраћенице и властита имена.

Кључне речи: језички корпус, лема, усвајање Л2, украјински, чешки.