UDK 81'366.542=161.1=162.2=172:81'322
Maksim Duszkin
Institute of Slavic Studies PAS (Inštitut za slavistiko Poljske akademije znanosti, Varšava)
maksim.duszkin@ispan.waw.pl

Danuta Roszko
University of Warsaw (Univerza v Varšavi)
d.roszko@uw.edu.pl

Roman Roszko
Institute of Slavic Studies PAS (Inštitut za slavistiko Poljske akademije znanosti, Varšava)
roman.roszko@ispan.waw.pl

MULTILINGUAL CORPORA IN CONTRASTIVE RESEARCH ON THE VOCATIVE IN RUSSIAN, POLISH AND LITHUANIAN

The aim of this article is to report on a contrastive analysis of the vocative forms in Russian, Polish and Lithuanian. It was to have been prefaced by a short introduction discussing the benefits of using non-commercial, multilingual corpora in such research. Unfortunately, the quality of available corpora was insufficient for the research. Consequently, our core goal could not be met. Ultimately, we focused on assessing these corpora and indicating the reasons for which effective use of these corpora in contrastive studies on the vocative is challenging. We indicate the issues related to the way in which the corpora we investigated are built—they are the source of many abnormalities in alignment and morphosyntactic annotation, as well as misinterpretation of material from one language or another.

Keywords: parallel corpora, vocative case, parallel corpora-illustrated approach

Cilj prispevka naj bi bila kontrastivna analiza uporabe vokativa v ruščini, poljščini in litovščini, ki ji sledi kratek uvod o prednostih uporabe nekomercialnih vzporednih korpusov pri tej vrsti raziskav. Presenetljivo za nas, se je kakovost dostopnih korpusov izkazala kot nezadostna za zanesljivo izvedbo načrtovanih raziskav. Zaradi tega nismo mogli doseči prvotnega cilja prispevka. Osredotočili smo se za ovrednotenje korpusov in obrazložitev, zakaj je učinkovita uporaba obravnavanih korpusov v kontrastivnih študijah vokativa zaenkrat precej vprašljiva. Našteli smo konstrukcijske napake korpusov, ki so vir številnih nepravilnosti pri vzporejanju in oblikoskladenjskem označevanju ter napačni interpretaciji jezikovnega gradiva.

Ključne besede: vzporedni korpusi, vokativ, vzporedni korpusno-ilustrirani pristop

1 Introduction

1.1 Short history of the vocative

The vocative is not a topic that is frequently analyzed in contrastive linguistics as an independent category, as one of the nominal cases, or as the main appellative form. The

fact that the vocative is included in the declension system of nouns derives from ancient Greek grammar, in which the vocative was considered a case. Later Latin grammars never challenged this approach. A similar interpretation of the vocative can be found in German grammar, which was based on its Latin counterpart. In the 19th century, as Indo–European contrastive historical linguistics was on the rise, the classification of the vocative as a case was routinely transferred to the many descriptions of languages that were being constructed at the time.

The architects of contrastive historical linguistics were strongly influenced by Pānini's grammar of Sanskrit. Pānini's grammar was particularly influential not only because of the striking similarities between Sanskrit and European languages, but also due to the methodology that was adopted for describing the grammar of a language. It is therefore somewhat odd for the vocative to be bound to the declension system in 19th century Indo–European grammar since, in the old Indian tradition, the vocative was completely independent of this system. Only in the 20th century was the peculiar nature of the vocative finally noticed by Bühler (1934) and Kuryłowicz (1949), who created a new category: the appellative/vocative function (Bühler referred to it as *Appellfunktion*, whereas Kuryłowicz as *appel*), which could be perceived as an independent utterance. Heinz (1988: 337–38) juxtaposes the subjective cases (that only include the vocative) and the objective cases (all the remaining cases).

1.2 Language selection

We chose three interrelated Indo–European languages to conduct our contrastive study on the use of the vocative. Polish and Russian represent two groups of the Slavic languages (eastern and western), whereas Lithuanian represents the eastern group of the Baltic languages. The choice of languages was not arbitrary. First, in Lithuanian nominal forms have retained many archaic features. The Lithuanian vocative form still carries a number of Proto-Indo–European (PIE) features and has also created new markers which are absent in PIE altogether. Second, a few centuries ago, Russian dropped the vocative, which it had inherited from PIE. However, at the present time, a new vocative category is rapidly being formed in Russian. Third, Polish has retained the vocative that it inherited from PIE, yet over the last century a gradual decline in its use has been observed. Concurrently, the nominative and vocative forms are used syncretically at an increasing rate.

1.3 Aims

The aims of this research based on a corpus-illustrated / corpus-informed approach were as follows:

a) to determine the actual frequency in the use of the vocative forms in Russian, Polish and Lithuanian, including in common, everyday language;
b) to examine the extent to which non-commercial parallel corpora can be used in contrastive studies on the use of the vocative.

We are fully aware that achieving aim (a) heavily depends on the results of aim (b).

2 State of research

2.1 The Indo–European (IE) context

The vocative is a category that the IE languages are familiar with; PIE singular masculine and feminine nouns and adjectives had vocative forms (Beekes 2011), which were then inherited by the languages that were formed as a result of the breakdown of their protoplast. However, the old vocative has not been retained in many modern languages: it was preserved in its full diversity in the Baltic languages (Lithuanian and Latvian); it also occurs in Slavic languages (Belorussian, Bulgarian, Czech, Croatian, Macedonian, Serbian, Ukrainian); Celtic languages (Scottish Gaelic, Irish, Manx); as well as in Kurdish, Hindi–Urdu, and Greek. The Germanic languages have all dropped the old vocative, as have the Romance languages. However, in Romanian, the vocative was re-formed (Anstatt 2008). The morphology of the vocative forms differs from typical case construction (word stem + inflection). It is assumed that in PIE, the vocative form agreed with the stem of the lexeme and was not inflected (Beekes 2011: 186).

2.2 Russian language

In modern Russian, two genetic phenomena exist that are independent of each other and can be understood as the vocative. First, there are historical forms of the vocative from old Church Slavonic e.g., *Боже*$_{(voc)}$ < *Бог*$_{(nom)}$ (Черных 1962: 173). The fact that the vocative was being dropped and replaced by the nominative case had already been evident in the Ostromir Gospels from the 11th century (Кузнецов 1953: 122). The old Russian vocative was used in singular forms of some nouns until the 14th or 15th century (Иванов 1990: 273) and was then fully replaced by the nominative form. In the later periods, the vocative was used as a stylistic figure.

In the period from the 17th until the 19th century, the vast majority of descriptive Russian grammars labelled the vocative as one of the cases in the Russian language. According to them, the vocative was the case whose form was identical to the nominative form (with the exception of the words that either come from Church Slavonic or imitate it), both in the singular and in the plural. This view was shared by Heinrich W. Ludolf (Ludolf 1696), Vasilii Adodurov (Adodurov 1731), Mikhail Lomonosov (Ломоносов 1755: 64–75), Nikolai Grech (Греч 1827: 53–54), and Aleksandr Vostokov (Востоков 1831: 22–23). However, the notion that Russian had a vocative whose form was identical to the nominative was firmly rejected by Gerasim Pavskii (Павский 1842: 275). He was perhaps the first author of a Russian grammar who did not include the vocative in the inflection patterns for nouns and removed the vocative from the list of the Russian cases.

The second phenomenon is the so-called new Russian vocative, present in the colloquial register: *Нин*$_{(voc)}$ < *Нина*$_{(nom)}$, *ребят*$_{(voc)}$ < *ребята*$_{(nom.pl)}$. It is formed for the nouns whose nominative inflection is either *–а* or *–я*; however, the new vocative has not fully replaced the nominative in direct, second-person forms of address. It is also possible to use the noun in the nominative case in all the instances that permit the new vocative.

The new Russian vocative is often discussed by authors of historical grammars from the 1950s and 1960s: they emphasise that it is a Russian-specific phenomenon that is genetically unrelated to the old vocative (e.g., Кузнецов 1953: 123).[1] The new vocative is hence understood as a special form of the noun which does not have the status of a separate case (see, e.g., Зализняк 1967). Daniel indicates that, in the written Russian texts, the new vocative was already present in the second half of the 19th century; this form was then strongly marked as rustic. The new vocative then began to occur in the intellectual fiction in the 1960s (Даниэль 2009: 243): this form was then perceived as colloquial, rather than rustic.

In the *Russian National Corpus* (Добровольский et al. 2005), the tag *voc* is used to indicate the vocative and encompasses the use of the old and new vocative forms. It is worth mentioning that the authors of the corpora approach the vocative as one of the Russian cases (Ляшевская et al. 2005: 124).

2.3 Polish language

The historical change of the vocative in Polish has been analysed by, *inter alia*, Rachwał (1992) and Anstatt (2005). Apart from the analysis and documentation of historical change in Polish academic texts concerned with this subject, the vocative was also considered as a part of the category of addressative forms, and more broadly as a form belonging to the lexicogrammatical category of the honorifics (see: Huszcza 1996, Przybylska 2001, and others). Many researchers claim that the use of the vocative forms in Polish is in decline; the language also permits for the nominative to be used instead of the vocative (Łuczyński 2007). The vocative form has been retained in the traditional addressative expressions and is a stylistic marker of formality in a conversation; it can also be used in a non-formal conversation to emphasise the emotional meaning of utterances. Anstatt (2005) has conducted a comprehensive analysis of the vocative in which she confirms the dissonance in the use of its forms. She highlights that the use of the vocative is compulsory with distant forms of address and it is optional and somewhat rare with non-distant forms of address. Łuczyński (2007) takes a different position, claiming that there are no reasons to argue that the use of the Polish vocative is indeed dwindling. Anstatt (2005) does also suggest that the vocative should be understood as a derivational category rather than as a case. A similar suggestion for categorising the vocative in Polish is made by Przybylska (2001); she recommends that a separate category of addressatives be created, in which the degree of politeness in utterances is considered key. She also poses the question as to whether the category should be considered lexical or grammatical.

2.4 Lithuanian language

In the research on Lithuanian grammar, no debate has taken place on the nature of the vocative. Many authors accept (and do not question) that the vocative is simply one

[1] The first text that analysed the new vocative was most likely the paper *Die Form des Vokativs im Russischen* (Obnorskij 1924).

of the seven cases of Lithuanian (e.g., Ulvydas 1965; Šukys 1984). However, Laigonaitė (in Jakaitienė et al. 1976) only mentions six cases and does not include the vocative. This view of the declension system in Lithuanian reoccurs in later editions of the academic grammar (see: Амбразас 1985: 90; Ambrazas 2005: 68). Importantly, even though these grammars state that the declension system of the noun only encompasses six elements, the declension paradigms they provide include seven cases. In her contrastive study of Russian and Lithuanian, Lichačiova (1985) also does not include the vocative in the declension system of the Lithuanian noun. Valeckienė, in turn, introduces the category of honorifics, in which the vocative is the key component (Valeckienė 1998: 211). She believes that the vocative, despite not meeting the criteria to be classified as a case, should be considered in the declension paradigm of the noun (1998: 251–52). In her follow-up study (Valeckienė 2000: 100–01), she provides context-based examples in which it is possible to use the nominative and the vocative interchangeably. These examples include the pronoun *tu* [you$_{(sg)}$], which correlates with the nominative form, e.g. Lt *Kur* ***tu***$_{(nom)}$, ***vaikas***$_{(nom)}$, *eini?* [Where are **you** going, **child**?].[2] Valeckienė also notes that some of the vocative forms overlap with the nominative forms. She concludes that all this speaks in favour of categorising the vocative as one of the cases. I. A. Seržant (2015: 187–88) rejects the syncretism of the nominative and the vocative, pointing to the strictly defined functions of the noun, which significantly deviate from the norms of standard European languages, as can be exemplified by the title of Nikolai Chernyshevskii's novel *Kas*$_{(nom)}$ *daryti*$_{(inf)}$*?* [*What is to be done?*].

The nature of the vocative is also discussed by A. Paulauskienė (2008: 8), who claims that if an element ceases to be a part of one structure, it becomes a part of another. She claims that like all the other cases, the vocative either has both the stem and the inflections or does not take any inflection (e.g. *sesut-*$_{(voc)}$ [sister]). Paulauskienė further indicates that, unlike all the other cases, the vocative does not create structures with other forms in the sentence, yet this does not mean that it is not a part of the sentence. Supposedly, this assumption about the vocative is confirmed by the fact that its use requires other additional, specific forms to be used e.g., the optative or the second-person imperative. Moreover, the vocative influences the intonation contour of the sentence. She also points to the fact that both the nominative and the vocative are used to name a person or an object, yet only the use of the vocative constitutes a direct form of address towards that object or person. The nominative, being deprived of this feature, is considered to be the unmarked element in this opposition. Paulauskienė claims that this feature supposedly means that it is possible to neutralise the opposition by substituting the vocative form with the nominative form that has the vocative's function e.g., Lt *Tai kur* ***kūmas***$_{(nom)}$ *eini?* [So where are you going, **godfather**?]. This substitution does, however, require a change in the sentence intonation. The author concludes that the existence of this neutralisation means that the vocative cannot be removed from the declension paradigm of the noun.

[2] The survey we conducted did not confirm her findings. Its participants dismissed the examples of the syncretism of the nominative and the vocative that she uses, categorising them as Russianisms.

3 Methodology

Our initial goal was to conduct a study that would use corpus resources. We selected exclusively the free-access multilingual corpora for our analysis.

We did not endeavour to define what the vocative is, or what status it has in the grammars of the languages we chose to examine. Any form that has been assigned the *voc* property in the corpus is therefore the subject of our analysis. The tagsets of all the corpus resources we have analysed (Russian, Polish and Lithuanian) discern the vocative as part of the declension system of a noun.

Prior to the analyses, we had been aware that the major weakness of the multilingual corpora is the fact that their representation of the spoken register is almost non-existent; we were also aware (see 2. above) that this is where the vocative forms are far more likely to be frequently represented. Hence, we intend to place emphasis on the multilingual resources that contain subtitles, which should imitate the spoken register.

4 Quality of search results for the vocative tag in the available non-commercial multilingual corpora

It came to our attention that not all of the multilingual corpora were valid candidates for our planned analyses; their invalidity goes beyond simply not representing at least one pair of the languages that we intended to analyse. For instance, the manually aligned *PELCRA Polish–Russian Parallel Corpus* (Pęzik et al. 2011) is only available for download. The lack of a dedicated search tool can be a real obstacle to many linguists. The resources not being tagged is another serious issue that effectively makes it impossible to search for particular lexeme classes. Therefore, we directed our attention towards the tagged corpora.

ParaSol (von Waldenfels 2012) is a manually annotated corpus that matches our criteria and contains all the language pairs that are of our interest. Unfortunately, conducting searches in the corpus is a major obstacle, as it requires its user to know the CQP query language as well as the desired tag forms. Simply searching for a particular form results in the following message being displayed "(0 0 0 hit.) 0 hits overall." Moreover, we found that the alignment itself contains serious and numerous mistakes, which means that the corpus is not usable for our intended research; for instance, a single Polish sentence (95212) *Jaka ja dla ciebie Klaudyna* has been assigned 42 Lithuanian sentences. We also question the degree of representativeness of the corpus for particular language pairs (pl–lt and ru–lt pairs contain only three texts each).

OPUS, the largest available corpus (Tiedemann 2016) much akin to *ParaSol*, comes with an unintuitive user interface, yet we noted that it contains far fewer alignment issues. It does, however, contain more language errors, as shown below:

> Ru — *Hamau…* [Natasha$_{(voc)}$] = Pl *Musimy wylaczyc alarm.* (Correct spelling: *Musimy* ***wyłączyć*** *alarm.*) [We must turn off the alarm.]

In order to find the subcorpus for a particular set of languages in the OPUS corpus, one has to resort to a blind search since it is not possible to browse the entirety of the resources at once. The values of tags used are not described, while the CQL manual only discusses the tagsets for English and German.

Korpus Polsko-Rosyjski Uniwersytetu Warszawskiego (Łaziński et al. 2015) is a relatively large corpus. It has been automatically tagged and aligned. The number of alignment errors is not high overall, but the number of tagging errors is very high in the area of our interest. The number of the results of the vocative form query amounted to 24,483 results for Polish and 1,497 for Russian; these specific forms are not highlighted in the search results. We endeavoured to analyse the first 200 results that were provided, expecting to find 200 instances in which the vocative was used; yet, 142 examples did not contain any forms of the vocative whatsoever. For instance: Pl *Takich elementów jest wiele.* [There are many such elements.]. The remaining 58 examples contained either vocative forms (e.g., *przyjacielu* [friend$_{voc}$] or the forms derived from the vocative e.g., *gówniarzu* [little shit$_{voc}$], *gnido* [dirtball$_{voc}$], *Jezus Mario* [Jesus Christ$_{voc}$], *matko jedyna* [Mother of God$_{voc}$] (see Przybylska 2001). When the search is calibrated to look for the vocative forms in both languages, the results state that the same number of the vocative forms as in Polish has been found *i.e.*, 24,483.

In a similar test done for Russian, 41 out of the 100 results were in fact exclamations that originated from the old vocative forms e.g., *боже мой* [My goodness]; 33 of the results were the vocative forms from the Gospels e.g., *Господи неба и земли* [Lord of heaven and earth] and the remaining 26 results contained no vocative forms whatsoever. We limited the search to the texts written after 1945, hoping to find more examples for the new vocative there. We obtained 84 sentences as a result, 77 of which contained the exclamatory form *боже*. The remaining results did not include any vocative forms, e.g. *Ну, ка-акже!* [Well, of course!]. We concluded that the annotation errors render this corpus unusable in the research on the vocative.

The Russian National Corpus [Добровольский et al. 2005) also contains subcorpora of the texts for the pairs ru↔pl, ru↔lt, and a multilingual subcorpus. All the languages of interest are tagged. The practicality of its application comes with severe limitations, however. First, there is no way to limit the search to a language in which a particular phrase, tag or form would be searched for. The search results will always contain the examples in all of the languages; in addition, statistics are calculated for an entire parallel corpus, not for one specific language included in it. To identify the specific examples from a particular language, the researcher has to manually inspect all the results and perform the computations.

Second, disambiguation is severely lacking. Only partial disambiguation has been performed for the Polish texts and the tags for the word forms in all the other languages represent all the possible grammatical readings. For instance, out of the 170 uses of the vocative in the lt–ru pair for the single text by Icchokas Meras' *Dingęs be žinios (1972–2005)* | *Без вести пропавший* [Missing in action], **only two of the examples**

can be conditionally categorised as vocative forms of the noun (the context indicates that the vocative is indeed present in the Russian word *Господи*):

> Lt ***Viešpatie****, juk įeis jie tuoj, galvas nuleidę, ir nuraminti norės ir gerą žodį pasakyti norės* […] = Ru *Господи, ведь войдут же они сейчас, головы опустив, и успокоить захотят* […]
>
> [Lord$_{(voc)}$, will they come in now with their heads down and want to reassure […]]

The remaining search results consist of a variety of forms that have been mislabelled as the vocative; for instance, the nominative plural masculine pronoun *kitas* [another] in the following sentence:

> Lt *Ir tik kitą kartą, daug vėliau, ateis* ***kiti***$_{(nom.pl)}$ *ir perduos kažką* […] = Ru *И только в следующий раз, намного позже, придут другие, и передадут что-нибудь* […]
>
> [And only the next time, much later, others will come and give something]

The number of mislabelled examples in the results for lt↔ru language pairs is very high, which invalidates the automatically calculated statistics and forces the researcher to resort to tedious manual analysis of the results. There are many mistakes in *voc* tag use in ru↔pl part annotation too: e.g., misannotated forms *popełnienie* ([…] *sąd uwzględnia* […] ***popełnienie*** *przestępstwa wspólnie z nieletnim* […] — accusative), *księżycu* (*Będą znaki na słońcu,* ***księżycu*** *i gwiazdach* — it is locative case, in the corpus anotated as vocative), *lewo* (*Ulryk von Biberstein spojrzał w* ***lewo*** […] — the form *lewo* here is a part of the adverbial phrase *w lewo*, it is not a noun in the vocative case).

For the multilingual subcorpus, this issue is further exacerbated by mislabelled results from the other languages. The fact that the number of results displayed is influenced by all the other languages stems from the premise on which the RNC was created *i.e.*, that the multilingual corpus is to display the results for all the 21 languages provided in the corpus. In terms of identifying the vocative forms, we also found that there was a multitude of incorrect interpretations for both Czech and Ukrainian. The very first result page for the search for the vocative forms in Mikhail Bulgakov's novel *The Master and Margarita*, is enough to realise that it contains no vocative forms whatsoever,[3] e.g.

> Uk — *РОЗДІЛ 11* ***РОЗДВОЄННЯ*** *ІВАНА Бір на протилежному березі* ***річки****, ще годину* ***тому*** *освітлений травневим сонцем, потьмарився, розпливсь і розчинився.* = Cs — 11. ***Ivanovo rozdvojení*** *Les na protějším* ***břehu řeky****, ještě před hodinou* ***projasněný*** *májovým sluncem, zešedl, rozmazal se a nakonec docela zmizel*.
>
> [En (RNC version) — CHAPTER 11. Ivan Splits in Two The woods on the opposite bank of the river, still lit up by the May sun an hour earlier, turned dull, smeary, and dissolved.]

The incorrect annotation of the forms in Ukrainian and Czech means that the number of results that have to be sifted through in the multilingual corpus is much greater than that in the bilingual corpora. Every single annotation error in this corpus, therefore, influences the overall count of the results displayed.

[3] In all the language versions of the text that have been included in the corpus, a startling 11,518 occurrences of the *voc* tag are identified.

In our research, we also considered using the *InterCorp* (Čermák et al. 2012) corpus. In terms of the languages of interest, it is significantly smaller in comparison to the *OPUS* corpus described herein, yet it is by no means a small corpus in its own right. However, this corpus also contained alignment errors e.g.:

> Pl — ***A oto pani skarb, Małgorzato Nikołajewna.***
>
> [En (Intercorp version) — And here is your property, Margarita Nikolayevna.]
>
> = Ru — ***А вот и ваше имущество, Маргарита Николаевна****, — и он подал Маргарите тетрадь с обгоревшими краями, засохшую розу, фотографию и, с особой бережливостью, сберегательную книжку, — десять тысяч, как вы изволили внести, Маргарита Николаевна.*
>
> [En (Intercorp version) — And here is your property, Margarita Nikolayevna.' Koroviev handed Margarita a manuscript-book with burnt edges, a dried rose, a photograph and, with special care, a savings-bank book: 'The ten thousand that you deposited, Margarita Nikolayevna.]

Once the results have been narrowed down to the Core subcorpus, we found no further alignment issues. We did, however, find errors in how the vocative forms were identified in Polish e.g., the word form *Julio* is interpreted as the vocative form of the lemma *Julia*, while it is in fact a nominative form of the lexeme/name *Julio*; this also holds true for the form *Pawle* or the initialism *ALU* (which is misinterpreted as the vocative form of the name *Ala*):

> Pl — ***Julio*** *idzie w naszym kierunku.*
>
> [En (Intercorp version) — Julio is coming towards us.]
>
> Pl — *Gospodarz* ***Pawle*** *już przywykł słuchać tych słów spokojnie, jakby nie dotyczyły jego.*
>
> [En (Intercorp version) — Pavle had got used to listening to these words calmly as if they did not refer to him.]
>
> Pl — ***ALU****? — zdziwił się Artur.*
>
> [En (Intercorp version) — "GPP feature?" said Arthur.]

We also found that all the exclamations that originated from the vocative forms were also identified as vocatives.

The resources for Lithuanian found in the corpus have not been tagged, which makes it impossible to search for the vocative forms for this particular language.

Finally, we have also noted a particularly low count of the Russian word forms tagged as the vocative in this particular corpus. For the Lithuanian–Russian resources, there were only 557 such word forms that represented merely six lexemes (*бог* [god], *господь* [lord], *отец* [father], *мама* [mother], *папа* [father], *Иисус* [Jesus]). For the Russian–Polish resources, this goes up to 3,961 word forms that, in theory, encompass 41 lexemes. However, a share of these lexemes is mislabelled e.g., the lexeme *смяташ* does appear in the results while it is not a Russian lexeme at all. This occurs in a Bulgarian text which has been mistakenly treated as a Russian one and included in the

Russian language resources. Other misinterpreted vocative forms involve the word *Лен* (considered to be a new Russian vocative for the feminine name *Лена*, tag: Npfsvy). The analysis of all the uses of the word *Лен* in the Russian–Polish pair indicates that every single use of the form *Лен* is misinterpreted in the corpus as a vocative, while it is, in fact, a completely different lexeme *Лен*, which should have been assigned the tag Npmsny e.g.:

> Ru — O, ***Лен***, они пытались разлучить меня с тобой. = Pl — *Oh,* ***Len****, Próbowali nas rozłączyć.*
>
> [Oh, Len, they tried to separate us.]

The final resources analysed within the scope of this paper come from *The CLARIN-PL parallel corpora* (Duszkin et al. 2021). For the particular language pairs of our interest (pl–lt, pl–ru, ru–lt), these corpora are smaller than their competitors in *OPUS* and *InterCorp*. However, unlike all the corpora enumerated above, the resources included in *The CLARIN-PL parallel corpora* have been aligned manually, which means that the quality of the results obtained from them is higher. Moreover, the multilingual resources of *CLARIN-PL* have been tagged automatically with the use of the state-of-the-art tagger versions, which makes it possible to conduct comparative research on the use of the vocative forms. The limitations of these corpora should be mentioned, however. These stem from the internal balance in the representation of particular language pairs. As it turns out, the resources shared for all the languages that are within the scope of our research (ru, pl, lt) are restricted to just a few European Union regulations, in which we do not expect to find the vocative. The majority of the Polish–Lithuanian content consists of the source texts — legal, academic, and specialised — and their translations. The content of the vocative forms in such texts is also minuscule. The number of subtitles and fiction for this pair is unfortunately extremely low, and these text types are the most common sources of the vocative. The texts representing the Polish–Russian pair are far more balanced as these resources contain a significant number of fiction texts and subtitles. However, they mainly contain texts translated from English and other languages. As a result, the representation of the vocative forms is far from uniform for Polish and Russian, which is the result of the different approaches to the adaptation of foreign proper names for Polish and Russian. For Polish, first names tend to be adapted to a greater degree:

> Pl *Janie, przestań!* = Ru *Джон, хватит.*
>
> [John, stop it!]
>
> Pl *Nie wstydzisz się, Tomku?* = Ru *Как тебе не стыдно, Том.*
>
> [Are you not ashamed, Tom?]

Especially given the actual use of the vocative, we would like to point to the fact that the foreign female first names in the Russian texts are often not inflected at all:

> Pl *Co myślisz o Helenie?* = Ru *Что вы думаете о Элен?*
>
> [What do you think about Helene?]

It is particularly important as the marked feature of the Russian feminine vocative forms is that they have no inflection, e.g. *Лена*$_{(nom)}$ vs *Лен*$_{(voc)}$.

5 Conclusions

Our initial research strategy was to use the non-commercial multilingual resources to conduct a contrastive corpus study on the use of the vocative in Russian, Polish and Lithuanian; we were unable to carry out this task in full. Unfortunately, almost all of the corpora available in free access were not suitable for conducting the research on the vocative;[4] this is the outcome of a number of causes and conditions. Foremost, some of the corpora fail to provide effective tools that would facilitate the searches for the vocative forms based on the morphosyntactic features thereof (e.g. *PELCRA Polish–Russian Parallel Corpus*). Certain corpora do not provide enough support to their users and come with poor user interfaces. For instance, the GUIs of *ParaSol* and *OPUS* are particularly unintuitive. As a result, the user is forced to blindly choose the resources for the searches and faces serious issues when formulating queries as there is no information provided regarding the tagsets used by the corpus; in some cases (e.g., for *OPUS*), there is no indication as to which resources have even been tagged.

Many of the corpora have also been automatically aligned, which constitutes a significant hindrance during search result analysis. *ParaSol* and *OPUS* contained the largest number of alignment errors. *CLARIN-PL corpora,* a small *Core* subcorpus of *InterCorp* and the *PELCRA Polish–Russian Parallel Corpus* were the only manually aligned corpora.

Many of the resources have not been pre-processed either, which means that some of the corpora contain not only spelling errors but also assign tags to forms that do not exist in certain languages; this further contaminates the search results. Spelling errors aside, we have also noticed that the search results contain formatting characters and other markers that should have been removed in pre-processing.

OPUS and *InterCorp* also misidentified languages. In the resources of *OPUS*, languages that use the Cyrillic script are not properly distinguished; for instance, many of the texts written in Russian are identified as Ukrainian. In *InterCorp*, some Bulgarian texts are categorised as Russian and even tagged with a tool appropriate for the Russian language.

However, the most serious flaw of the corpora, which makes them of little use for contrastive research on the vocative case, is the poor quality of tagging itself. Even if we ignore the fact that some of the resources are not tagged at all (e.g., the Lithuanian subcorpus in *InterCorp*), the automatic tagging has also proven unreliable. The taggers

[4] We also checked the tagging quality of some resources available in the commercial SketchEngine corpus (https://www.sketchengine.eu/). During the trial period, we gained access to the Polish Web 2012 corpus (plTenTen12, RFTagger). We used the query [tag = "subst:sg:voc.*"]. Among the first hundred examples only 26 contained vocative form, other 74 were mistakenly tagged as vocative.

used to process the resources in Polish and Russian misinterpret the exclamations derived from the vocative forms as the vocative itself. Depending on the register, the overall number of incorrect morphosyntactic interpretations can even exceed 50% of all search results. The taggers used for the Polish corpora will frequently assign the vocative tag to the forms whose spelling is identical to that of the vocative forms (e.g., np. Pl *synu*$_{\text{loc}}$). The Polish taggers also struggle with categorising the names whose endings resemble the vocative inflection—e.g., the title of the TV programme *Galileo* is interpreted as a feminine vocative form of the noun *Galilea*; similarly, the country *Togo* [Togo] is also misinterpreted as the vocative form of the noun *toga* [a toga]. We have also encountered the instances in which the vocative tag is assigned to the misspelt forms—e.g., *Dojdziemy do celu* ***tyko*** *wówczas* [...] [We will get there [...]]. Instead of the word *tyko*$_{\text{(voc)}}$ [pole], the word *tylko* [only] should have been used. *InterCorp* alone contained ca. 135 such errors (with this misspelt form). The Lithuanian taggers also yield unsatisfactory results. They struggle to identify proper names (and especially foreign first names) and to differentiate between the vocative forms and participles. In the RNC corpus, no Russian and Lithuanian resources have been disambiguated; consequently, the search results for the vocative query contained a large number of incorrect examples. In some instances, forms marked as vocative are in fact never vocative.

The fact that the overall percentage of the resources shared for Russian, Polish and Lithuanian was fairly low constituted yet another limitation for the research on the vocative forms in the corpora described herein.

References

Vassily E. Adodurov, 1731: Anfangs-Gründe der Rußischen Sprache. *Deutsch-Lateinisch- und Rußisches Lexicon, samt denen Anfangs-Gründen der Rußischen Sprache. Zu allgemeinen Nützen bey der kayserl. Academie der Wissenschaften zum Druck befördert*. St. Petersburg: Gedr. in der Kayserl. Acad. der Wissenschafften Buchdruckerey. [Page numbering of the part is independent: 1–48].

Vytautas Ambrazas (ed.), 2005: *Dabartinės lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Henning Andersen, 2012: The new Russian vocative: Synchrony, diachrony, typology. *Scando-Slavica* 58/1. 122–67.

Tanja Anstatt, 2005: Der polnische Vokativ: Aussterbende Kasusform oder produktiv verwendetes Wortbildungsmittel. *Zeitschrift für Slawistik* 50/3. 328–47.

Tanja Anstatt, 2008: Der slavische Vokativ im europäischen Kontext. *Linguistische Beiträge zur Slavistik* XV. Ed. Ljudmila Geist, Grit Mehlhorn. München: München Sagner. 9–26.

Karl Bühler, 1934: *Sprachtheorie*. Jena: Gustav Fischer Verlag.

František Čermák, Alexandr Rosen, 2012: The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17/3. 411–27.

Maksim Duszkin, Danuta Roszko, Roman Roszko, 2021: New Parallel Corpora of Baltic and Slavic Languages — Assumptions of Corpus Construction. *Lecture Notes in Artificial Intelligence* 12848. 172–83.

Adam Heinz, 1965: System przypadkowy języka polskiego. *Język i językoznawstwo*. Ed. Adam Heinz, Warszawa. 312–419.

Romuald Huszcza, 1996: *Honoryfikatywność: Gramatyka — pragmatyka — typologia*. Warszawa: Wydawnictwo Akademickie „Dialog".

Evalda Jakaitienė, Adelė Laigonaitė, Aldona Paulauskienė, 1976: *Lietuvių kalbos morfologija*. Vilnius: Mokslas.

Jerzy Kuryłowicz, 1949: Le problème du classement des cas. *Biuletyn Polskiego Towarzystwa Językoznawczego*, 1949/9. 20–43.

Marek Łaziński, Magdalena Kuratczyk, 2015: Korpus polsko-rosyjski Uniwersytetu Warszawskiego. *Polskojęzyczne korpusy równoległe.* Ed. Ewa Gruszczyńska, Agnieszka Leńko-Szymańska. Warszawa: Instytut Lingwistyki Stosowanej UW. 83–95.

Ala Lichačiova, 1985: Ar galima laikyti šauksmininko formą linksniu. *Mūsų kalba* 1985/2. 29–32.

Edward Łuczyński, 2007: Wołacz we współczesnej polszczyźnie. *Język Polski 87*/2. 149–56.

Heinrich Wilhem Ludolf, 1696: *Henrici Wilhelmi Ludolfi Grammatica Russica que continet non tantum praecipua fundamenta Russicae Linguae, verum etiam Manuductionem quandam ad Grammaticam Slavonicam.* Oxford [Oxonium]: [E Theatro Sheldoniano].

Sergey Obnorskij, 1924: Die Form des Vokativs im Russischen. *Zeitschrift fur Slavische Philologie* 1/1–2. 102–16.

Aldona Paulauskienė, 2008: Opozicijos ir jų neutralizacija gramatinių kategorijų paradigmose. *Kalbų studijos* 13. 5–14.

Piotr Pęzik, Maciej Ogrodniczuk, Adam Przepiórkowski, 2011: Parallel and spoken corpora in an open repository of Polish language resources. *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics* Ed. Zygmunt Vetulani. Poznań, Wydawnictwo Poznańskie i Fundacja Uniwersytetu im. A. Mickiewicza. 511–15.

Renata Przybylska, 2001: Czy w języku polskim istnieje osobna kategoria adresatywów? *Język w komunikacji*. T. 1. Ed. Grażyna Habrajska. Łódź: Wydawnictwo Wyższej Szkoły Humanistyczno-Ekonomicznej w Łodzi. 180–86.

Maria Rachwał, 1992: O przyczynach zmian systemu adresatywnego języka polskiego w XIX wieku. *Język a kultura: Polska etykieta językowa*. Tom 6. Ed. Janusz Anusiewicz, Małgorzata Marcjanik. Wrocław: Wiedza o Kulturze. 41–49.

Ilja A. Seržant, 2015: The nominative case in Baltic in a typological perspective. *Argument Realization in Baltic*. Ed. Axel Holvoet, Nicole Nau. 137–98.

Jonas Šukys, 1984: *Linksnių ir prielinksnių vartojimas*. Kaunas: Šviesa.

Jörg Tiedemann, 2016: OPUS — parallel corpora for everyone. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT): Projects/Products.* Riga: Baltic Journal of Modern Computing. 384.

Kazys Ulvydas (ed.), 1965: *Lietuvių kalbos gramatika 1*. Vilnius: Mintis.

Adelė Valeckienė, 1998: *Funkcinė lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Adelė Valeckienė, 2000: Linksnis ir jo funkcijos lietuvių kalboje. *Acta linguistica Lithuanica* 42. 66–104.

Ruprecht von Waldenfels, 2012: ParaSol: introduction to a Slavic parallel corpus. *Prace Filologiczne* LXIII. 293–301.

Витаутас Амбразас (ред.), 1985: Г*рамматика литовского языка*. Вильнюс: Mokslas.

[Vitautas Ambrazas (red.), 1985: *Grammatika litovskogo jazyka*. Vil'njus: Mokslas.]

Александр Х. Востоков, 1831: *Русская грамматика Александра Востокова, по начертанию его же сокращенной грамматики полнее изложенная*. Санкт-Петербург: Тип. И. Глазунова.

[Aleksandr H. Vostokov, 1831: *Russkaja grammatika Aleksandra Vostokova, po načertaniju ego že sokra*ščennoj *grammatiki polnee izloženn*aja. Sankt-Peterburg: Tip. I. Glazunova.]

Николай Греч, 1827: *Практическая русская грамматика, изданная Николаем Гречем*. Санкт-Петербург: Тип. Имп. Санкт-Петербургского воспитательного дома.

[Nikolaj Greč, 1827: *Praktičeskaja russkaja grammatika, izdannaja Nikolaem Grečem*. Sankt-Peterburg: Tip. Imp. Sankt-Peterburgskogo vospitatel'nogo doma.]

Михаил А. Даниэль, 2009: Новый русский вокатив: история формы усеченного обращения сквозь призму корпуса письменных текстов. *Корпусные исследования по русской грамматике*. Под ред. К. Л. Киселевой, В. А. Плунгяна, Е. В. Рахилиной, С. Г. Татевосова. Москва: Пробел-2000. 224–44.

[Mihail A. Daniėl', 2009: Novyj russkij vokativ: istorija formy usečennogo obraščenija skvoz' prizmu korpusa pis'mennyh tekstov. *Korpusnye issledovanija po russkoj grammatike*. Pod red. K. L. Kiselevoj, V. A. Plungjana, E. V. Rahilinoj, S. G. Tatevosova. Moskva: Probel-2000. 224–44.]

Дмитрий О. Добровольский, Алексей А. Кретов, Сергей Шаров, 2005: Корпус параллельных текстов: архитектура и возможности использования. *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*. Москва: Индрик. 263–96.

[Dmitrij O. Dobrovol'skij, Aleksej A. Kretov, Sergej Šarov, 2005: Korpus parallel'nyh tekstov: arhitektura i vozmožnosti ispol'zovanija. *Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy.* Moskva: Indrik. 263–96.]

Андрей А. Зализняк 1967: *Русское именное словоизменение*. Москва: Наука.

[Andrej A. Zaliznjak 1967: *Russkoe imennoe slovoizmenenie*. Moskva: Nauka.]

Валерий В. Иванов, 1990: *Историческая грамматика русского языка*. Москва: Просвещение.

[Valerij V. Ivanov, 1990: *Istoričeskaja grammatika russkogo jazyka*. Moskva: Prosveščenie.]

Пётр С. Кузнецов, 1953: *Историческая грамматика русского языка. Морфология*. Москва: Изд-во Московского ун-та.

[Petr S. Kuznecov, 1953*: Istoričeskaja grammatika russkogo jazyka. Morfologija*. Moskva: Izd-vo Moskovskogo un-ta.]

Михайло В. Ломоносов, 1755: *Российская грамматика Михайла Ломоносова*: Санкт-Петербург: Имп. Акад. наук.

[Mihajlo V. Lomonosov, 1755: *Rossijskaja grammatika Mihajla Lomonosova*: Sankt-Peterburg: Imp. Akad. nauk.]

Ольга Н. Ляшевская, Владимир А. Плунгян, Дмитрий В. Сичинава, 2005: О морфологическом стандарте Национального корпуса русского языка. *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*. Москва: Индрик. 111–35.

[Ol'ga N. Ljaševskaja, Vladimir A. Plungjan, Dmitrij V. Sičinava, 2005: O morfologičeskom standarte Nacional'nogo korpusa russkogo jazyka. *Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy.* Moskva: Indrik. 111–35.]

Герасим П. Павский, 1842: *Филологические наблюдения протоиерея Г. Павского над составом русского языка, Второе рассуждение. Об именах существительных*. Санкт-Петербург: Тип. Имп. Акад. Наук.

[Gerasim P. Pavskij, 1842: *Filologičeskie nabljudenija protoiereja G. Pavskogo nad sostavom russkogo jazyka, Vtoroe rassuždenie. Ob imenah suščestvitel'nyh*. Sankt-Peterburg: Tip. Imp. Akad. Nauk.]

Павел Я. Черных, 1962: *Историческая грамматика русского языка*. Москва: Учпедгиз.

[Pavel Ja. Černyh, 1962: *Istoričeskaja grammatika russkogo jazyka*. Moskva: Učpedgiz.]

Povzetek

Članek poskuša ovrednotiti uporabnost nekomercialnih večjezičnih korpusov za kontrastivne raziskave vokativa v ruščini, poljščini in litovščini. Naša analiza je pokazala, da so nekomercialni večjezični korpusi, ki so uporabnikom trenutno dostopni, za takšne raziskave večinoma neprimerni. Rezultati korpusnega iskanja za vokativ vsebujejo veliko neustreznih oblik, kar pomeni, da lahko nepravilno identificirane vokativne oblike ne le otežujejo, temveč celo onemogočajo izvedbo jezikoslovnih analiz. Trenutno stanje je posledica načina gradnje obravnavanih korpusov, ki je vključevala samodejno vzporejanje in označevanje. Poleg tega smo ugotovili, da označevanje nekaterih korpusov ne vključuje večznačnosti. Odsotnost oblikoskladenjskih oznak je velika ovira za učinkovito raziskovanje slovničnih kategorij.

Ob tem avtorji poudarjamo, da je vokativ posebna slovnična kategorija, značilna le za nekatere jezikovne registre. Denimo v pravnem, tehničnem ali strokovnem jeziku vokativa praktično ni, zaradi česar pride pri uporabi korpusov ta nesorazmerna pojavnost toliko bolj do izraza, saj večina korpusnih gradiv temelji ravno na besedilih naštetih registrov.