Research Article

Aleksej Tikhonov*

What is the authentic internet register before & after the Russian invasion in Ukraine? Polish and Czech YouTube comments from 2021–2023

https://doi.org/10.1515/lpp-2024-0026
Received August 12, 2024; accepted November 18, 2024; published online December 9, 2024

Abstract: Over one million tokens of comments were collected for the study using data mining methods. The videos under which the comments were dug out were not chosen arbitrarily but according to the current official national YouTube trends in Poland and the Czechia. The comments were collected under the most popular videos in ten categories: cars, comedy, fashion & lifestyle, gaming, music, non-political interview, politics, report, sports, and video blog. The data collection was carried in 2021–2022 and 2023 from under 40 videos per language and period, 160 videos in total. The corpus data should reveal more about the internet register through the part of speech (POS) frequencies, and the syntactical statistics. In addition, the comments are stylistically clustered in R to make dependencies in linguistic usage visible and better understandable. The study aims to expand the term register, using Polish and Czech, to include language on the internet and distinguish between authentic and non-authentic internet registers in comparison to other registers. An additional sociolinguistic aspect of the analysis is the influence of the Russian war against Ukraine on the linguistic behavior of YouTube users.

Keywords: Polish; Czech; corpus linguistics; YouTube comments; internet register

1 Introduction

> ‘*nie wiem na chuj youtuberzy pchają się w rozmowy z politykami*
> *jak nie potrafią zostawić swoich poglądów na czas wywiadu*’[1]

[1] “I don’t fucking know why YouTubers start talking to politicians when they can’t leave their views aside during the interview $_{\mathrm{PL}}$” Comment of a YouTube user on a non-political video interview on

***Corresponding author: Aleksej Tikhonov**, University of Zürich, Zurich, Switzerland,
E-mail: aleksej.tikhonov@uzh.ch. https://orcid.org/0000-0002-0772-3397

In the text sequence above, the Polish commentator used YouTube's online video platform in 2021 to articulate his viewpoint regarding a video interview. This spontaneously composed brief textual expression engenders a wide-ranging terrain for the research, encompassing fields such as media studies, media psychology, political science, and last but not most minor, linguistic analysis by employing the prepositional phrase "na chuj" – which, in this particular context, serves as a vulgar synonym for the interrogative "how" or "why" – the author conspicuously designates his communication as colloquial. What distinguishes this instance is that colloquial language is typically relegated to oral communication; however, it finds expression in written form in this case. An ancillary element at play here pertains to the semantic backdrop of Polish politics, wherein the author casts aspersions on the competence of the YouTubers as interviewees on political subjects. Notably, the video does not cater to mere entertainment but addresses politics. Consequently, YouTube serves as a platform where, in contrast to traditional television, viewers can proffer their responses to the watched content directly and spontaneously.

Every day, YouTube users upload about 720,000 h of video material. The communication around these videos does not occur in one direction: from the video creator to the audience. Registered users can comment on the videos. This opportunity is crucial here, mainly because the comments under the videos fulfill the referential function. Users comment publicly under the videos, giving linguistics a rich source of authentic language material from a digital discourse. These circumstances make YouTube a dynamic field for analyzing synchronic changes in language and a special language register – the internet register. The Internet register is the variety of written language that appears in authentic online contexts and situations. This includes, for example, private and public chats in messaging services, short text messages or posts on X and Threads, as well as comments on platforms such as Instagram or YouTube. This article focuses specifically on the YouTube comments. While academic registers or spoken language registers have already been researched relatively well, the internet register shows deficits. As the following cross-section of research on the Polish Internet shows, researchers have focused primarily on semantics and English-Polish language contact. Other languages have been of great importance just now.

In summation, it is imperative to underscore that the Internet and the language employed within its domain constitute salient social, economic, and cultural spheres worthy of scholarly inquiry, as evidenced by the aforementioned research. The particular utilization of language within the Internet necessitates a contextual interpretation, which can aptly be characterized as a register, as expounded upon by

YouTube Poland. The comment is part of a comments' collection anonymized for the analysis as "polish_np_interview_1.txt".

Biber and Conrad (2009). Within this theory, the present article embarks upon a comprehensive examination of the Internet register within the domains of Polish and Czech. The central research question guiding this investigation pertains specifically to the identification of national and supranational differentials and similarities in the usage of the focus languages and the multilingual variation in the YouTube's comment section in both countries, with an overarching objective of delineating the contours of an authentic Internet register.

Prior to delving into the empirical investigation, it is imperative to situate this study within the broader landscape of research devoted to the exploration of language usage on the Internet within the contexts of Poland and Czechia. This contextual foundation is essential for a nuanced understanding of the linguistic dynamics at play within the language use on the Internet.

2 Relevant research

The selection of Polish and Czech as the focus languages for comparison is not an arbitrary one. They share attributes as being West Slavic, being located in Central Europe as immediate geographical neighbors, which, nevertheless, does not guarantee parallel linguistic development. This perspective has been duly acknowledged by certain researchers in the field of synchronic linguistics. Of particular importance is the dynamic area of the Internet, which serves as a fertile ground for linguistic investigation. Barska and Śnihur (2015) explored the burgeoning trend wherein the Internet plays an increasingly integral role among Polish, Czech, and Slovak younger generations, permeating their daily lives. This phenomenon underscores the pivotal role of the Internet as a socio-cultural force, thereby warranting scholarly scrutiny. Notably, Tomczyk and Kopecký (2016), while not directly addressing language usage, delved into an aspect of paramount concern within this digital landscape – the security of the Internet in Poland and Czechia, particularly in the context of safe online space for minors. This focus underscores the multifaceted dimensions of the Internet as a contemporary socio-linguistic domain, intertwined with critical issues extending beyond linguistic usage. Moreover, the intricate nexus between language and the Internet has led to innovative endeavors, including attempts to construct inter-Slavic languages online – a phenomenon manifesting in Poland, Czechia, and other Slavic countries (Meyer 2016). This emerging trend demonstrates the profound interplay between linguistic expression and Internet usage.

Loewe (2006) has provided a comprehensive exposition of the Polish Internet as a valuable source of research data for linguistics. Loewe's discourse encompasses various domains within the realm of linguistic research on the Internet, including examining the oralization of the written register. Simultaneously a similar point of view had taken root within the context of Polish on the Internet. The emergence of not only a distinctive written manifestation of colloquial language replete with its unique vocabulary and syntax but also demonstrated that this form of communication on the Internet exerted a palpable influence on offline communication dynamics (Grzenia 2007). A seminal development in the linguistic exploration of the Polish Internet occurred with the introduction of the corpus linguistic tool Monco (http://monco.frazeo.pl/) in the early 2010s as a project of the Laboratorium Lingwistyki Kulturowej i Międzykulturowej (Laboratory of Cultural and Intercultural Linguistics).[2] Notably, this resource remains active today, affording accessibility to online language datasets available to the general public, including sources such as gazeta.pl and tvn24.pl. This corpus search engine merits special attention for its adept assistance to researchers in uncovering instances illuminating the utilization of lexical elements, idiomatic expressions, and lexical-grammatical structures in authentic, contemporary Polish used on the Internet.

Zabawa meticulously examined lexical borrowings of English origin within the Polish internet message boards. The cross-pollination of languages in this digital space serves as a rich field for linguistic exploration (Zabawa 2010). Moving forward, Czerski et al. (2016) and Roziewski et al. (2016) engaged in an ambitious attempt by constructing an N-gram collection based on a vast corpus encompassing the entirety of Polish internet content, generously provided by The Common Crawl Foundation project. Following rigorous lexical processing, the resultant dataset has been harnessed for extracting flat N-gram compilations, subsequently applied with marked success in machine learning within the context of natural language processing. To enhance search capabilities, an innovative contribution was made by developing NEKST, a semantic search engine situated at the Institute of Computer Science, Polish Academy of Sciences (2016). This resource represents a notable advancement in the field of semantic search technology.

Furthermore, Derecka (2019) conducted a comparative investigation into online linguistic discrimination targeting trans individuals, both in English and Polish contexts. This study sheds light on the linguistic aspects of discrimination and provides valuable insights into linguistic disparities in these two languages. Jarynowski et al. (2020) delved into the connotations of the COVID-19 pandemic within the Polish internet landscape. Of particular note is their pioneering use of internet data in conducting a corpus and sociolinguistic study, reflecting the prevailing sentiments of Polish society amidst the unprecedented challenges posed by the pandemic. Finally, Król and Zdonek (2023) conducted a comparative analysis, examining the interplay between language usage and references to cultural heritage in Polish and English contexts. Their findings suggest that semantic references to cultural heritage are,

[2] http://lalikumi.uw.edu.pl/informacje-o-korpusach/ Last Access: 19 September 2023.

albeit slightly less frequent, a notable presence on the Polish Internet, with the disparity potentially influenced by the respective numbers of language speakers. This study underscores the intricate relationship between language, culture, and digital spaces.

The pioneering investigations into the Czech language's manifestation on the internet embarked upon a multifaceted exploration, recognizing the interplay between its written form and its concurrently colloquial nature, tinged with the influence of slang, as early as the year 2000 (Vitochová and Bozděchová 2000). Notably, the early 2000s witnessed one of the initial research initiatives to cast Czech on the internet as a subject of philological relevance, with a focus on a concrete case study. This study delved into the mentions of tornadoes in the Czech territories, spanning from the 12th century to the contemporary internet register (Setvák et al. 2002).

Jandrová et al. (2006) marked a pivotal juncture by examining Czech on the internet not merely as a modification of written or colloquial language but as an autonomous linguistic phenomenon, particularly within the domain of online chats. This innovative perspective laid a firm foundation for further analytical investigations, acknowledging that linguistic materials on the internet are influenced not only by linguistic factors but also by broader social, psychological, cultural, and stylistic considerations, thus transcending a purely structuralist description and analysis (Pilátová 2008).

Kozáková (2010) made noteworthy observations regarding the intensification of lexical borrowing from English and the escalating internationalization of Czech on the internet. These phenomena initially germinated within the digital realm, subsequently infiltrating the offline sphere, particularly among individuals working in the IT sector, and eventually permeating colloquial Czech. The anthology "Interaktion von Internet und Stilistik," published by Wonisch & Tošović in 2016, further contributed to the discourse by examining Czech and Polish within the context of the hybridization of written and colloquial registers.

Crucially, the progress in this domain saw the emergence of two seminal corpora. The NET corpus, inaugurated as the first iteration of a synchronous corpus of semi-official internet communication in the Czech language, was introduced by Jeziorský in 2019. Concurrently, the first generation of the ONLINE1 corpus, encompassing the dynamic content of the Czech internet landscape, including online journalism, discussions, and social networks, spanning the years 2017–2021, was compiled by Cvrček and Procházka in 2020.

A pivotal milestone in the research on Czech internet language was reached with the advent of Multidimensional Analysis (MDA) of register variability. This innovative approach, guided by linguistic statistical considerations (Cvrček et al. 2018), prompted the study of two Czech corpora in 2020: a "traditional" corpus replete with extensive metadata and an internet-sourced corpus characterized by an opportunistic composition, representative of the “searchable” web. The findings underscored that while certain traditional text categories, such as journalism or non-fiction, exhibit linguistic features amply covered by data gleaned from internet crawling, the spectrum of variations in traditional corpora proves more comprehensive due to their inclusion of texts for which no commensurate substitutes are available through general web crawling techniques. Traditional corpora encompass an array of textual genres, including fiction and user-generated content such as comments on Facebook and forums (Cvrček et al. 2020).

Moreover, this innovative approach was further refined by Marklová et al., who supplemented it with reception tests conducted on native speakers. Notably, these tests revealed marked disparities in ratings between spoken and written modes, suggesting that while somewhat unprepared written situations, like chatting on WhatsApp or commenting on a Facebook post, do occur, the spoken modalities epitomize greater spontaneity and dynamism. Conversely, prepared and static situations may also manifest in spoken modalities, exemplified, for instance, by presentations at corporate meetings. Nonetheless, the hallmark of prototypicality remains closely associated with written situations (Marklová et al. 2023).

Building upon the methodological frameworks established by Loewe (2006) in the case of Polish, and Jandrová et al. (2006) and Cvrček and Procházka (2020) for Czech, the present study embarks upon an uncharted domain of linguistic analysis within the Internet register – the realm of YouTube comments, a specific facet of internet discourse that has hitherto remained unexplored.

3 The research design

In both Poland and Czechia, YouTube stands as one of the preeminent online platforms, offering users an active space for commentary and interaction. In Poland, the YouTube user base comprises approximately 24.6 million accounts, which translates to around 65 % of the population (Tur 2020). Meanwhile, in Czechia, approximately 5.9 million users are registered on YouTube, encompassing approximately 55 % of the population (Mouchová 2023). It is crucial to interpret these statistics with a degree of caution, recognizing that they do not represent absolute values. Factors such as the existence of fake accounts created for SPAM-related activities, the presence of inactive accounts, and instances where a single individual maintains multiple accounts are not mentioned in these statistics. Nevertheless, these statistics serve as a substantial indicator of the platform’s significance within the digital landscape. Furthermore, the active engagement observed on the platform, as elucidated in the forthcoming sections of this article, provides compelling evidence attesting to its pertinence for linguistic analysis.

The methodology employed for data collection adhered to the following predefined principles. Within each of the represented countries on YouTube, national trends are discernible, as evidenced by the TOP 200 most-viewed videos specific to that country. Among these highly viewed videos, a classification into ten distinct content categories was established by the author of this article, which is contingent upon the thematic orientation of the videos: 1) cars, 2) comedy, 3) fashion and lifestyle, 4) gaming, 5) music, 6) non-political interviews, 7) politics, 8) reports, 9) sports, and 10) video blogs. From each of these categories, four most popular videos were selected for the purpose of this study. It should be noted that comments were extracted exclusively from videos where the comment function remained enabled, with data mining techniques employed to gather these comments. However, it is imperative to clarify that only the comments themselves were collected, while responses to these comments were not included within the scope of this study to maintain focus on the primary analysis and not to extend the study with the conversational aspect of the linguistic data.

The data collection process transpired during three distinct time intervals: firstly in September 2021, secondly in March 2022, and lastly in February 2023. The initial survey was conducted precisely six months before the expansion of the Russian war against Ukraine, with the subsequent survey taking place during the initial month following the commencement of the full-scale invasion. The third and final survey occurred one year subsequent to the invasion. Consequently, this study assumes an additional dimension as a sociolinguistic and political linguistic investigation, unveiling evolving patterns in the behaviour of YouTube users in Poland and Czechia before, immediately after, and one year following the invasion, thus yielding insightful findings.

The comments corresponding to a given video were anonymized and systematically archived in chronological order, each under a distinct file with nomenclature following the pattern "politics_PL1.txt." In this nomenclature, the file name conveys the thematic category of the video (politics), the country (PL for Poland or CZ for Czechia), and the popularity ranking of the video within its respective category (with "1" designating the most popular video). Subsequently, these text files were collated into a corpus format within the SketchEngine (Kilgarriff et al. 2014) platform, facilitating a comprehensive and detailed examination of the data.

The subsequent sections of this article delve into the composition of the corpus and the languages featured within the TOP 200 videos. Following this, an in-depth analysis of the TOP 100 lemmas is conducted, with particular attention to the stylometric interrelationships among the various thematic blocks spanning the three-year timeframe.

4 The corpus design

Comments on the videos were systematically mined within the framework of the ten thematic video categories mentioned earlier. For YouTube Poland, in total, a corpus comprising 1.076.195 tokens was built and subsequently annotated. The distribution of the tokens across the ten categories and the three-year temporal scope under consideration is elucidated as follows in Figure 1.

In the case of YouTube Czechia, a sum total of 260,779 tokens were collected over the identical time span and in strict accordance with the same methodological framework (Figure 2).

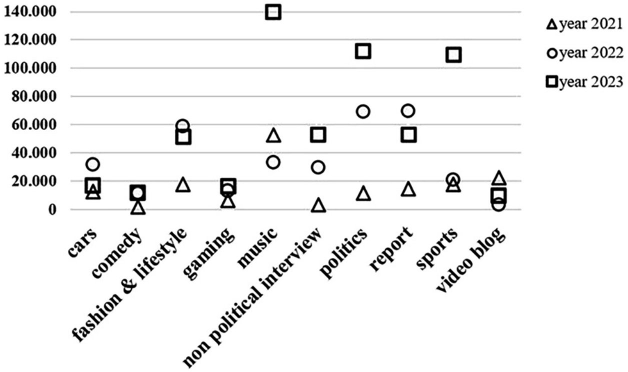

Figure 1: Nu. of tokens (y-) in the respective categories (x-axis) for YouTube Poland 2021–2023.

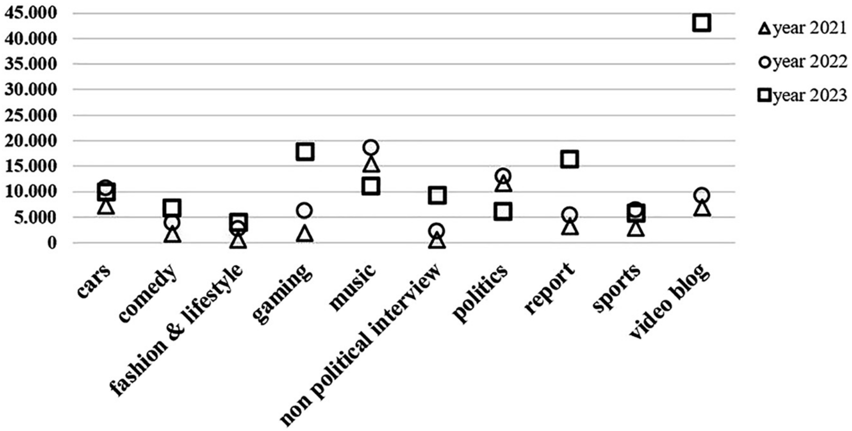

Figure 2: Nu. of tokens (y-) in the respective categories (x-axis) for YouTube Czechia 2021–2023.

When considering the absolute numerical data, it might erroneously appear that users in Poland exhibit more comments on YouTube, thereby generating a larger pool of linguistically relevant data in comparison to the users in the neighboring country. However, this initial observation fails to encapsulate the complete picture. As previously elucidated, Poland, being the larger of the two countries, boasts a larger user base on YouTube than Czechia, in absolute terms. To be precise, Poland registers approximately 24,6 million YouTube accounts, while Czechia accounts for 5,9 million, signifying that Poland possesses 4.17 times as many YouTube users as Czechia. By applying this factor, derived from the quotient of 4,17, to the 260.779 tokens collected in Czechia, we can extrapolate that the equivalent token count for Poland would be approximately 1.087.448 tokens. Comparing to Poland with 1.076.195 tokens, this adjustment ensures that both countries are placed on an equal footing in terms of their productivity when it comes to commenting on YouTube videos. In total, the corpus of this study amounts to 1.336.974 tokens.

5 Analysis

The analysis section of this article conducts a comprehensive view at the 100 most frequent lemmas in the comments. Subsequently, a stylometric examination will be performed, involving the statistical clustering of the data. An additional facet of the analysis pertains to the languages employed within the videos, thereby extending the scope of investigation and enhancing the overall comprehensiveness of the dataset. This step contributes to a more holistic understanding of the data landscape and ultimately contributes to the formulation of the study's conclusive findings. Following these three analytical steps, the study culminates in the formulation of its concluding remarks and findings.

5.1 Most frequent lemmas & POS

The subsequent section of this study delves into an analysis of the frequencies of lemmas that directly pertain to the contextualization of the Russian war in Ukraine and its repercussions. Furthermore, an examination of the most frequently used POS is conducted to facilitate a comparative evaluation of language usage trends within the Internet register across both countries, thereby deepening the level of analysis.

5.1.1 Frequent lemmas from the context of Ukraine and the war

Among the most frequently employed lemmas a marked semantic divergence becomes apparent between YouTube comments in Poland and Czechia. During the period spanning 2021–2022, comments originating from Poland exhibit a more pronounced political orientation when contrasted with those from Czechia. This trend further intensifies in 2023, as political topics increasingly dominate Polish commentary. Notably, the prevailing political discourse centers prominently on Russia's war against Ukraine.

In the years 2021–2022, the usage frequency of specific lemmas in Polish commentary underscores the political topics. Notably, *Ukraina* (Ukraine) was ranked 90th, *Putin* occupied the 95th position, and *wojna* (war) the 103rd position among the most frequently lemmas in Polish commentary. By 2023, this trend crystallizes further, with *Ukraina* (Ukraine) ascending from the 90th place to the 62nd, and *wojna* (war) surging from the 103rd to the 61st position. This evolution in language usage mirrors the growing salience of political discourse in the context of the Russia-Ukraine conflict in Poland. Consequently, an examination using the search query [lemma = "wojna"][]{1,10}[lemma = "Ukraina|ukraina"] provides insights into the various contexts within which political debates unfold, thereby shedding light on the evolving opinions within these YouTube comments:

> *Wojna na Ukrainie rozpoczęła się w 2014 kiedy USA obaliły rząd Kijowski i posadziły pod żyrandolem przychylne sobie władze.*[3] (polish_politics_1.txt)

> *Mówią, że gdyby nie ta wojna na Ukrainie, to USA by zbankrutowało??*[4] (polish_politics_1.txt)

The selection of these two examples may give the impression that discussions about the war are limited to political videos. However, a comprehensive examination of comments across various other video categories can provide additional insights into this phenomenon:

> *Putin niewie co robi jest imbecylem, Rosja ma 144mln ludności z tego połowa to Wojsko, ma takie rezerwy Złota że myślę że można ich porównać do Arabii sałdyjskiej i liczne kontrakty na całym świecie są dostawcami gazu i ropy z czego wiele krajów nie zrezygnuje, ta wojna będzie trwała aż Rosja nie zaimie Ukrainy , chodźby i lata, dziwi mnie taka pomoc Ukrainie ze strony Polaków szybko umiemy wybaczać jeszcze naście lat temu mordowali Polaków, fakt czasy się zmieniły ale czy napewno się nie cofamy.*[5] (polish_report_3.txt)

[3] "The war in Ukraine began in 2014 when the US overthrew the Kiev government and placed sympathetic authorities under the chandelier. $_{\mathrm{PL}}$".

[4] "Some people say that if it weren't for the war in Ukraine, the USA would have gone bankrupt?? $_{\mathrm{PL}}$".

[5] "Putin doesn't know what he's doing, he's an imbecile, Russia has 144 million people, half of them are military, it has such gold reserves that I think they can be compared to Saudi Arabia and numerous contracts around the world are suppliers of gas and oil, which many countries will not give up, this war will last until Russia takes over Ukraine, it will last for years, I am surprised about such Polish help for Ukraine, we can forgive fast, even if a dozen years ago they murdered Poles, it is true that times have changed but are we sure we are not going backwards? $_{\mathrm{PL}}$".

> *Do wszystkich putinowskich trollów opłacanych za szmatławe ruble: 2 wojna światowa i wydarzenia na Wołyniu skończyły się 80 lat temu. Teraz mamy wojnę na Ukrainie, najbliższym nam genetycznie narodzie, który przez 200 lat żył z Polakami w jednym państwie.*[6] (polish_sports_3.txt)

The comments above serve to underscore that political themes significantly influence content related to reporting and sports. Additionally, it is worth noting that, alongside Polish, comments in Ukrainian and Russian languages also make appearances beneath the comment section of those videos in Poland (Figure 3).

Among the 75 comments in an East Slavic language, the distribution of languages is dominated by Russian: 20 comments are composed in Ukrainian, while 25 (+30 comments published by trolls) are in Russian. However, a noteworthy observation pertains to the remaining 30 comments written by Russian trolls, all of which share identical content – a sentence accompanied by a link to a video of Russian propaganda:

> *В доме у нацистов на Украине:*[7] [YouTube-link to a Russian propaganda video] (polish_report_3.txt)

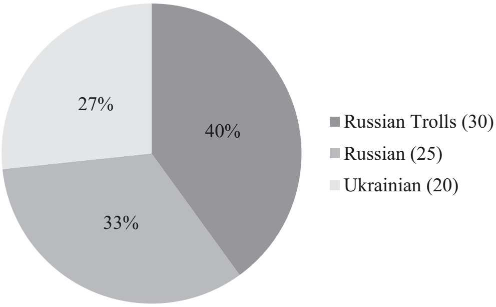

Figure 3: Distribution of Ukrainian and Russian comments about Russia's war against Ukraine on YouTube Poland.

[6] "To all Putin trolls paid with rag rubles: World War 2 and the events in Volhynia ended 80 years ago. Now we are at war in Ukraine, our genetically closest nation, which has lived with Poles in one country for 200 years. $_{\mathrm{PL}}$".

[7] "In a Nazi house in Ukraine RUS" – The author of the comments in this section is likely a Russian troll, who consistently disseminates the Putin narrative portraying Ukraine as a state influenced by (Neo) Nazis. Beyond the semantic implications, the author's choice of the preposition *на* (on) instead of *в* (in) within the prepositional phrase is notable. This preposition choice can predominantly be ascribed to the contextual framing of Ukraine from the Russian point of view as the superior power in contrast to Ukraine.

The comment was repetitively spread by a Russian troll[8] and therefore cannot be assessed as material that would be relevant to this study. However, the following results can serve as pertinent examples:

> *Україна переможе! З нами Бог та правда!* ☝[9] (polish_politics_1.txt)
>
> *Дякуєм за підтримку України!*[10] (polish_politics_3.txt)
>
> *Bedą bluzy wlepy mogły by być jakiś fajny z logiem зерна українського нема Super gratulacje!!!*[11] (polish_report_4.txt)
>
> *Поляки поздравляем Вас братья! Спасибо от народа Украины за все, что вы для нас делаете!* ❤❤❤❤❤❤[12] (polish_sports_3.txt)

An intriguing observation lies in the consistent expressions of gratitude directed towards the Poles in Ukrainian and Russian, acknowledging their support for Ukraine. However, there exists an exception within the commentary, revolving around the discourse on economic policies, specifically the Polish embargo on Ukrainian wheat. In this context, a user, despite primarily utilizing Polish for his or her comment, resorts to using Ukrainian to articulate the *wheat*. This shift to the Ukrainian language serves for distinguishing Polish wheat from its Ukrainian counterpart within the discourse in a more obvious way.

In contrast, the topic of war remains largely peripheral on YouTube Czechia. Over the entire span from 2021 to 2023, there are no references to terms such as *war*, *Ukraine*, *Russia*, or other relevant keywords among the 100 most frequently used tokens. Notably, *Rusko* (Russia) is ranked 390th in 2023, *Ukrajina* (Ukraine) the 513th position, and *válka* (war) stands at the 818th spot.

Despite the detected distinctions evident in the comments from Poland and Czechia, users from both countries exhibit certain commonalities in their linguistic patterns too. As delineated by Figures 1 and 2, it is apparent that music videos fall

[8] Also known as *Russian web brigades*, established, among others, by Yevgeny Prigozhin, with the primary tasks of conducting disinformation campaigns and disseminating pro-Russian hate speech across the internet (Tanchak 2017, Aro 2022). It is important to note, however, that not all individuals engaging in such activities are acting under direct commission from the Russian government; some do so voluntarily.

[9] "Ukraine will win! God and truth are with us! ☝ $_{\text{UA}}$".

[10] "Thank you for supporting Ukraine! $_{\text{UA}}$".

[11] "There will be sweatshirts with stickers they could be some nice ones with the $_{\text{PL}}$ No Ukrainian Wheat $_{\text{UA}}$ logo Great congratulations!!! $_{\text{PL}}$".

[12] "Poles, congratulations, brothers! Thank you from the people of Ukraine for everything you do for us! ❤❤❤❤❤❤ $_{\text{RUS}}$" – The author of the comment congratulates Poland on a victory in a sporting event in Russian.

within a category characterized by a moderate volume of comments, ranging between approximately 11.018 and 18.627 tokens in Czechia and 33.160 and 52.875[13] tokens in Poland.

5.1.2 Parts of speech (POS) and semantic trends

Beyond the contextualization of the war, additional semantic trends emerge within the comments in both countries. In the years 2021–2022, a numeral was a common feature in both countries' comments. Specifically, *jeden* (meaning *one* in Polish and Czech) ranked as the 71st most frequently used lemma in Poland and the 76th in Czechia. This reflects a global trend observed in the national contexts of Poland and Czechia. After a video is published, users frequently begin to comment with the numeral *one* or *first* (*pierwszy* in Polish and *první* in Czech), which can subsequently be traced back to the lemma *one* in the annotation. This behavior appears to be driven by a competitive spirit to be the first to comment under the video, with no particular semantic relevance or connection to the content of the video. In Poland, this trend intensifies in 2023, with the numeral rising to the 49th position. In Czechia, it remains relatively consistent, being on the 78th position.

Another notable trend observed in both countries pertains to semantically nearly identical commentary found under music videos. Expressions like *good song* or *nice song* are frequently used here. In Poland during 2021–2022, *piękny* (beautiful, nice) ranked 86th, followed by *piosenka* (song) in the 87th position. In Czechia, the rankings were 35th for *písnička* (song) and 44th for *dobrý* (good), similarly following each other. Positions 36 to 43 were occupied by function words excluded from the study. In 2023, this trend intensifies in Poland, with *dobry* (good) reaching the 18th position and *piosenka* (song) occupying the 30th position. In Czechia, there is a slight decline in this trend, with *dobrý* (good) at the 35th position and *písnička* (song) at the 77th position.

Regarding specific POS-categories, both Poland and Czechia exhibit some similarities, but they also manifest significant differences (Figure 4).

Within the 100 most frequent lemmas, each language comprises two adjectives. The use of names, adverbs, and numerals remains consistently low, each occurring in less than three instances per 100 most used lemmas. The fundamental difference

[13] The year 2023 was omitted from the discussion due to an extraordinary surge in the volume of comments, reaching 139.690 tokens in Poland. This notable increase may be attributed to the prevalence of automatically generated comments, a practice within the music industry that operates within a legal gray area.

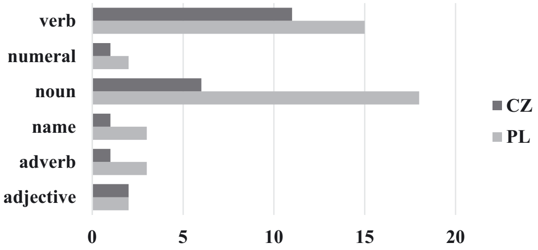

Figure 4: POS-distribution in Poland and Czechia in 2021–2023.

lies in the overall statistical profile of these categories when comparing the two countries. After excluding function words and emojis, semantically loaded POS make up 43 % of Poland's 100 most frequent lemmas. Conversely, in Czechia, this figure stands at 22 %. The comparison of verb and noun usage is noteworthy in this context, with Polish users demonstrating a notably higher frequency of verbs, and an employment of nouns three times more often than in Czechia. A possible explanation for the phenomenon could be that users in Czechia use emojis more often and replace verbs, nouns or entire sentences with them.

5.2 Clustering the comments

The *stylo* package, developed by Eder and colleagues (Eder et al. 2016) for the R statistics software, plays a pivotal role in the analysis conducted in this section. The package enables the assessment of various distance measures, shedding light on the stylistic relationships between texts and investigating whether style, topic, or other factors contribute to finding out more about the relationship between texts. To facilitate this analysis, comments from the same video are compiled into a single txt-file, resulting in a data setup comprising four files for each category and country. In total, there are 40 txt-files for the years 2021–2022 and 40 txt-files for 2023. The clustering process involves creating dendrograms for 2021–2022 and 2023. The dendrograms represent the result of measuring 2-g of the tokens using Eder's Delta distance for YouTube Poland in 2021–2022 as follows in Figure 5.

The observation made in Section 5.1.1 regarding frequent lemmas related to Ukraine and the war, which indicated similarities between the categories of politics, sport, and report, is corroborated by the dendrogram analysis. Within the upper branch of the overall statistical tree, the subcluster is bounded by "music_PL_1" and "sports_PL3." The core of this branch, comprising the closest ends of the final branches, primarily consists of comments related to sports, non-political interviews, fashion and lifestyle, supplemented by report and cars, and further complemented by comments under political videos. Consequently, this sub-cluster can be defined by

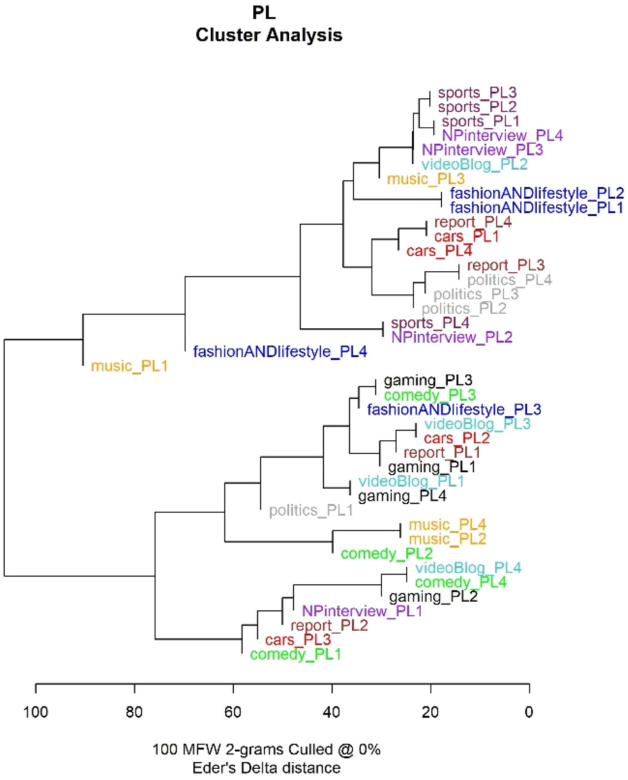

Figure 5: Clustering of the comments from YouTube Poland 2021–2022.

its discussions on sports, fashion and lifestyle, cars, and politics. These topics are discussed in a similar manner and as indicated in Section 5.1.1, tend to be the most politicized and serious topics. In contrast, the lower sub-cluster can generally be characterized as a hybrid entertainment cluster, where no specific topics dominate. Instead, gaming, comedy, music, and video blogs are discussed in a similar manner within this sub-cluster. In 2023, the trend with comments on politics and non-political interviews consistently expressed in a similar manner solidifies further. This alignment is evident in the lowest sub-cluster, as depicted in the dendrogram (Figure 6).

The political or political influenced comments have largely isolated themselves from other comments and form a semantically consistent sub-register. This sub-register can be characterized by the fact that even non-political interviews are influenced by politics, as well as to some extent, reports and sports. Compared to 2021–2022, reports and sports have shifted somewhat towards the entertainment

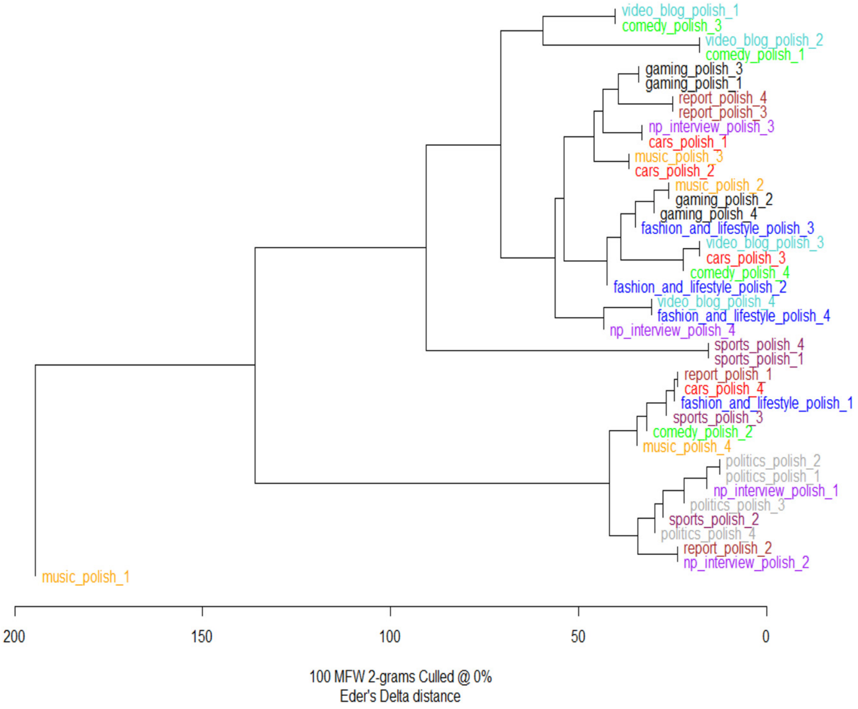

Figure 6: Clustering of the comments from YouTube Poland 2023.

cluster. It is also worth emphasizing that in most cases, comments related to video blogs, comedy, and gaming are situated at the opposite end of the overall cluster. This positioning underscores the stylistic differentiation between these spheres in contrast to political topics.

Another notable observation within the overall cluster is the separation of the most popular music video, labelled with the number 1 at the end of the file's name, in both figures. This separation, especially in the case of 2023, raises the possibility that these comments may be artificially generated to boost the video's popularity, as previously discussed in footnote 13. The stylistic-statistical disparity between these comments and all other comments, including those from other music videos, is substantial enough to suggest a clear distinction between human-generated and machine-generated texts.

Similar to Poland, a distinct sub-cluster emerged in Czechia in 2021–2022 characterized by political discussions, including sports and partially the topic of cars, which can be found in the centre of the dendrogram (Figure 7).

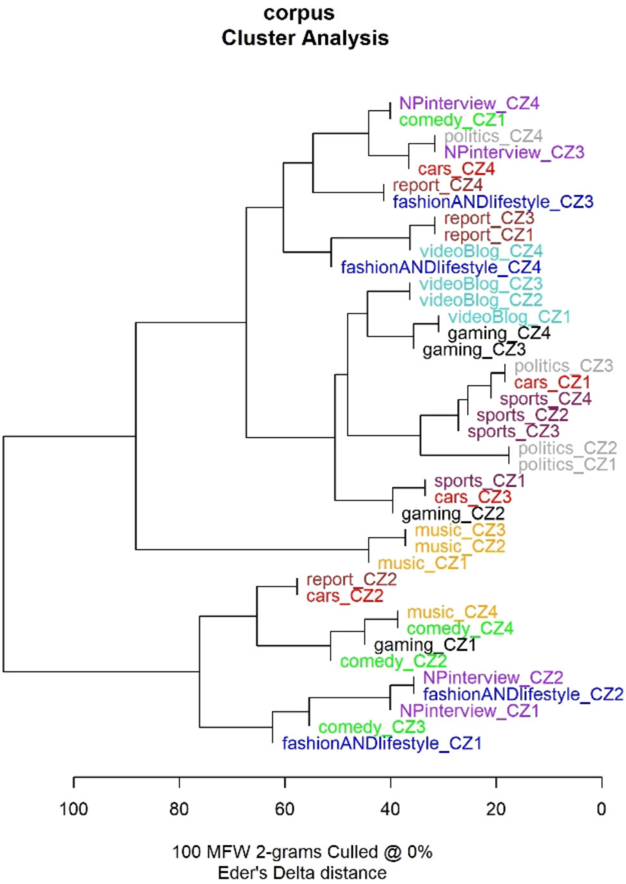

Figure 7: Clustering of the comments from YouTube Czechia 2021–2022.

A subtle difference is that political and sporting themes in YouTube Czechia tend to converge even more closely than in Poland. In neighbouring sub-clusters, comments related to music, gaming, and video blogs are also present and are positioned relatively close to the political discourse. In contrast to Poland, the entertainment cluster constitutes a smaller portion of the whole data in Czechia and is defined by comedy, non-political interviews, as well as fashion and lifestyle.

While 2023 represents a precise crystallization of the Internet register into semantic areas for Polish comments, it brings an almost complete hybridization for Czech comments. In 2023, the sub-cluster about politics and sports has completely dissolved and is spread across the entire cluster. However, there are subtle tendencies towards unity among video blogs, music comments, and comedy on YouTube

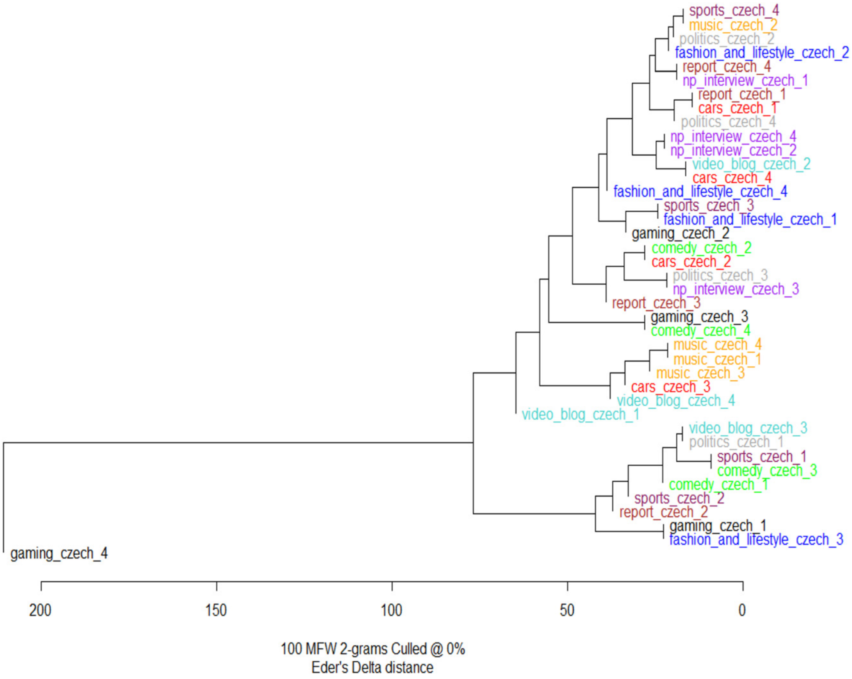

Figure 8: Clustering of the comments from YouTube Czechia 2023.

Czechia, but they cannot be unequivocally identified as the central semantic areas. An interesting observation, similar to the one about the most popular music videos in Poland, is that in 2023, for the first time, a collection of comments separates itself distinctly from all other comments – specifically, "gaming_czech_4," located at the very bottom left of the cluster (Figure 8).

In summary, it appears that Polish users are somewhat divided between those engaging in political discourse and those seeking entertainment, whereas Czech users are gradually forming a shared Internet register that is primarily characterized by entertainment topics.

5.3 Languages of the videos

In the context of examining language usage within the comments, it is pertinent to consider the languages employed in the videos themselves. In Poland, videos encompassed five languages (sorted by frequency in descending order): Polish, English, Russian, Ukrainian, and Turkish. These six languages are distributed among the 200 videos as follows in Figure 9.

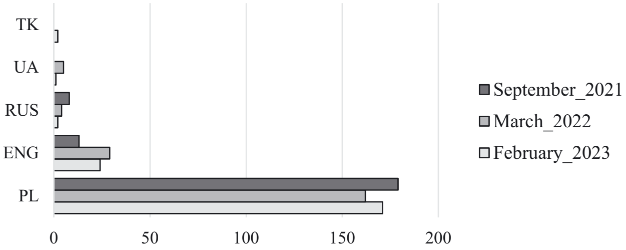

Figure 9: Languages used in the top 200 videos on YouTube Poland.

The use of Polish within videos demonstrates a consistent high presence, accounting for 89 % in 2021, 81 % in 2022, and 85 % in 2023. Notably, English-language videos in Poland hold the second position, constituting 7 % in 2021, 14 % in 2022, and 12 % in 2023. This pattern may suggest a reciprocal relationship: as the viewership of Polish-language videos increases, the consumption of English-language videos diminishes, and vice versa. More evident trends emerge with regards to the remaining two Slavic languages. In 2021, Russian-language videos constituted 4 % of the total most-viewed video content on YouTube Poland. In 2022, this share halved to 2 %, a trend that persisted in 2023, amounting to approximately 1 %. In 2022, Ukrainian-language videos made their debut and immediately accounted for 3 % of the content, although in 2023, they were on par with Russian videos, each comprising around 1 %. Additionally, Turkish-language videos made their inaugural appearance in the study in Poland and Czechia, capturing 1 % of the content in 2023. It is noteworthy that the presence of Turkish language is a global phenomenon, particularly in the realm of viral entertainment videos from Turkey and is not unique to Poland or Czechia.

The observation made on YouTube Czechia regarding the interrelation between the country's national language and the foreign language English supports the notice made in this matter in Poland. Indeed, in Czechia, there is a discernible falling trend in the prevalence of Czech videos among the 200 most-viewed videos (Figure 10).

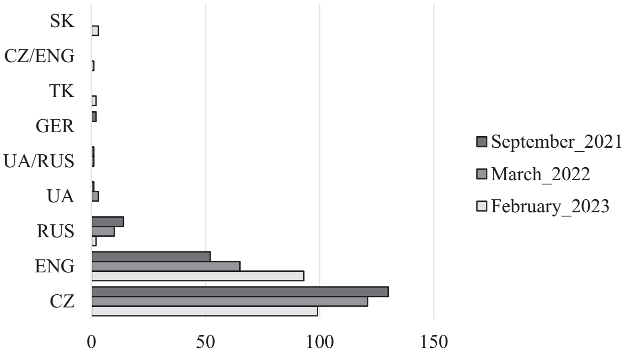

Figure 10: Languages used in the top 200 videos on YouTube Czechia.

In 2021, Czech videos constituted approximately 65 % of the most-viewed content. This figure declined to 60 % in 2022 and further decreased to 49 % in 2023. Conversely, English-language videos exhibited a contrasting ascending trend, accounting for 26 % in 2021, 32 % in 2022, and 46 % in 2023. At the same time, in contrast to Poland, Czechia exhibits a greater diversity of foreign language videos on YouTube. A total of nine languages and language combinations were observed in the videos: Czech, English, Russian, Ukrainian, bilingual Ukrainian-Russian, German, Turkish, bilingual Czech-English, and Slovak. In tandem with the English, 2023 marked the inclusion of the first bilingual Czech-English video within the top 200.

Nevertheless, Czechia's trends mirror those observed in Poland in certain respects. Russian-language videos exhibit a decline in viewership, and Ukrainian videos experienced an increase in 2022 followed by a notable drop in 2023. When conducting a holistic comparison between Poland and Czechia, it is evident that the popularity of Russian-language content is on the decline. Ukrainian videos experienced a surge in popularity in 2022, which, however, subsided within the same year. Multiple factors can underlie this phenomenon, including migration from Ukraine to Poland, migration (or return) from Poland to Russia, and the overarching political context of the ongoing war. In this conflict, Russia is the aggressor, while Ukraine is a nation defending itself. Poland and Czechia both demonstrate solidarity with Ukraine, which can contribute to shifts in language preference and content consumption patterns. However, Czechia distinguishes itself from Poland by also watching bilingual Ukrainian-Russian videos. The remaining languages in the statistical data exhibit unique trends, each with a share of approximately 1 %. This observation applies not only for Turkish but also to German and Slovak.

6 Conclusions

Returning to the primary research question of this study – to identify national and cross-national differences and similarities in language use within YouTube comments in Poland and Czechia, with the overarching goal of characterizing an authentic Internet register – it is first necessary to provide a succinct overview of the analysis.

One of the findings was the marked divergence in language usage between Poland and Czechia, particularly in the realm of political discourse. Polish comments exhibited a strong political orientation, with a growing emphasis on Russia's war against Ukraine, whereby the observations showed as well critical discourses within the question on Ukrainian wheat in the Polish market. Over the years 2021–2023, Polish commenters increasingly used terms like *Ukraina* (Ukraine), *Putin*, and *wojna* (war), reflecting a heightened focus on the ongoing geopolitical tensions. This trend was particularly pronounced in 2023, indicating the growing salience of political discourse within Polish YouTube comments.

In contrast, Czech comments appeared to be less politically charged, with no references to terms such as *war*, *Ukraine*, *Putin*, or *Russia* among the 100 most frequently used tokens over the same period. While political and sporting themes converged more closely in Czechia than in Poland, the overall linguistic digital landscape in Czech comments exhibited a greater convergence around entertainment topics. This suggests that Czech users were gradually forming a common Internet register primarily characterized by entertainment, in contrast to Poland's divided landscape between political discourse and entertainment.

However, both countries shared commonalities in their linguistic patterns, particularly in commenting on music videos. Expressions like *good song* or *nice song* in Polish and Czech were frequently used in both countries, reflecting a universal appreciation for music transcending national boundaries. Another supranational tendency was to comment under the videos *first* or *one* trying to be the first person commenting on the video.

A crucial aspect of the analysis was the consideration of the languages used in the videos themselves. This revealed intriguing insights into language preference among viewers. In Poland, over 80 % of the videos were in Polish and between 6,5 and 14,5 % in English. Notably, as the viewership of Polish-language videos decreased, the consumption of English-language videos grew, suggesting a reciprocal relationship between the two languages. In Czechia, a different trend was observed, with Czech videos declining in prevalence (49 % in 2023) while English-language videos gained ground (46 % in 2023). Interestingly, Czechia exhibited a greater diversity of foreign language videos, including Russian, Ukrainian, German, Turkish, and bilingual combinations as Ukrainian-Russian and Czech-English. This diversity hinted at the influence of global factors and international content on Czech YouTube viewers. In both countries, a clear decrease of Russian was observed.

In general, the numerous findings indicated that the examined YouTube comments from Poland and Czechia within various thematic categories from 2021 to 2023 are a profitable source for research on the current Internet register. The approach can be applied to other languages. The key findings include a notable divergence in the use of political language in Polish comments, reflecting the Russia's war against Ukraine, and incorporating comments on sports and non-political interviews into this context. In contrast, Czech comments were less politically oriented. Nevertheless, both countries showed similarities in commenting on music videos, but linguistic patterns differed in other categories, as for example the more frequent use of nouns and verbs by the Polish users and replacing the same POS through emojis in Czechia. A stylometric examination revealed in a clearer way the two major clusters of comments around (i) politics and (ii) entertainment in Polish. Language usage trends were also influenced by the languages used in the videos themselves, with Polish videos mainly in Polish, and the falling trend of videos in Czech and the increase of English videos as well as bilingual Czech-English videos.

In conclusion, this analysis underscores the dynamic and context-dependent nature of Internet language usage. It demonstrates how national and supranational influences, including political events, cultural factors, and video language, shape the linguistic landscape of YouTube comments. The formation of an authentic Internet register is a complex process influenced by a multitude of factors, and this research provides valuable insights into this evolving digital discourse. Understanding these patterns and variations is essential for grasping the nuances of online communication in an increasingly interconnected world, in which the users of one country tend to use their own language and stay in the national context, and the users of other countries tend to become a part of the global Internet register and use YouTube for entertaining purposes.

An aspect that was not initially addressed by the research question but emerged prominently in the data is the presence of automatically generated content and disinformation campaigns within the Internet register. This unintended focus led to the exploration of two key areas: (I) Comments generated by bots, primarily utilized to boost the popularity of videos, particularly in the case of music and gaming content, as evidenced by the stylometric analysis using *stylo* (Eder et al. 2016). (II) Disinformation disseminated by trolls, specifically Russian trolls, with recurring syntactic patterns evident in the KWIC analysis. Both these areas underscore the challenges posed by the Internet register and highlight pertinent research domains spanning multiple disciplines, from linguistics and political science to media studies and law & economics within the music industry.

Research funding: This work was funded by Deutsche Forschungsgemeinschaft (no.: SFB/CRC 1412).

References

Aro, Jessikka. 2022. *Putin's trolls: on the frontlines of Russia's information war against the world*. New York, NY: Ig Publishing.

Barska, Anetta & Janusz Śnihur. 2015. The role of internet in marketing activities aimed at Gen Y consumers, based on selected regions of Poland, Czech Republic and Slovakia. *Economic and Environmental Studies* 15(2 (34)). 189–209.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style/Douglas Biber, Susan Conrad*. Cambridge Textbooks in Linguistics. Cambridge: University Press.

Cvrček, Václav, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian Jan Zasina. 2018. From extra- to intratextual characteristics: charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory* 17(2). 351–382.

Cvrček, Václav, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková, Adrian Jan Zasina & Vladimír Benko. 2020. Comparing web-crawled and traditional corpora. *Language Resources and Evaluation* 54(3). 713–745.

Cvrček, Václav & Pavel Procházka. 2020. *ONLINE1: Monitorovací Korpus Internetové Češtiny*. Praha: Ústav Českého národního korpusu FF UK. www.korpus.cz.

Czerski, Dariusz, Krzysztof Ciesielski, Michał Dramiński, Mieczysław Kłopotek, Paweł Łoziński & Sławomir Wierzchoń. 2016. What NEKST? – Semantic search engine for polish internet | SpringerLink. *Challenging Problems and Solutions in Intelligent Systems*. 335–347.

Derecka, Magdalena. 2019. Manifestations of transphobia in computer-mediated communication. A case study of language discrimination in English and polish internet-mediated discourse. *Studies in Polish Linguistics* 14(3). 101–123.

Eder, Maciej, Jan Rybicki & Mike Kestemont. 2016. Stylometry with R: A package for computational text analysis. *The R Journal* 8(1). 107.

Grzenia, Jan. 2007. *Komunikacja Językowa w Internecie*. Warszawa: Wydawnictwo Naukowe PWN.

Jandová, Eva, Jaroslav David & Diana Diana Svobodová. 2006. *Čeština na WWW chatu*, Vyd. 1. Ostrava: Ostravská univerzita v Ostravě, Filozofická fakulta.

Jarynowski, Andrzej, Monika Wójta-Kempa & Vitaly Belik. 2020. Trends in perception of COVID-19 in polish internet. *medRxiv*. https://doi.org/10.1101/2020.05.04.20090993.

Jeziorský, Tomáš. 2019. *NET v1: Korpus Poloofíciální Internetové Komunikace*. Praha: Ústav Českého národního korpusu FF UK. www.korpus.cz.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1). 7–36.

Kozáková, Věra. 2010. Some observations about the language of the internet and the Czech language. *Revista de Administratie Publica Si Politici Sociale* 1. 31–39.

Król, Karol & Dariusz Zdonek. 2023. Cultural heritage topics in online queries: A comparison between English- and polish-speaking internet users. *Sustainability* 15(6). 5119.

Loewe, Christopher. 2006. The internet and its resources in polish linguistic research: Overview. *Biuletyn Polskiego Towarzystwa Językoznawczego*(62). 93–104.

Marklová, Anna, Olga Buchmüller, Christoph Demian, Roland Meyer & Luka Szucsich. 2023. Register in Czech: Designing an MDA-based experimental study. *Linguistics Beyond and Within* 9. 60–75. in print.

Meyer, Anna-Maria. 2016. Slavic constructed languages in the internet age. *Language Problems and Language Planning* 40(3). 287–315.

Mouchová, Klára. 2023. Sociální sítě v číslech. *Forendors*. https://www.forendors.cz/p/e14c5c86d69c52f664b2be31b1b257a0.

Pilátová, Jindřiška. 2008. Eva Jandová a Kol.: Čeština Na Www Chatu. Ostrava 2006, 262 s. ISBN 80-7368-253-2. *ROSSICA OLOMUCENSIA* XLVII. 81–83.

Roziewski, Szymon, Wojciech Stokowiec & Antoni Sobkowicz. 2016. N-gram collection from a large-scale corpus of polish internet | Request PDF. *Machine Intelligence and Big Data in Industry*. 22–34.

Setvák, Martin, Milan Šálek & Jan Munzar. 2003. Tornadoes within the Czech Republic: From early medieval chronicles to the “internet society”. *Atmospheric Research* 67–68. 589–605.

Tanchak, Peter. 2017. The invisible front: Russia, trolls, and the information war against Ukraine. In Olga Bertelsen (ed.), *Revolution and war in contemporary Ukraine: The challenge of change*, 253–282. Alberta: University of Alberta.

Tomczyk, Łukasz & Kamil Kopecký. 2016. Children and youth safety on the internet: Experiences from Czech Republic and Poland. *Telematics and Informatics* 33(3). 822–833.

Tur, Henryk. 2020. *Poznalismy pierwsze, wyłącznie polskie dane dotyczące YouTube*. Telepolis. https://www.telepolis.pl/tech/rozrywka/poznalismy-pierwsze-wylacznie-polskie-dane-dotyczace-youtube.

Vitochová, Kateřina & Ivana Bozděchová. 2000. *Čeština na internetu*. Praha: Univerzita Karlova, Ústav českého jazyka a teorie komunikace.

Wonisch, Arno & Branko Tošović (eds.), 2016. *Interaktion von Internet und Stilistik, Internet und Stil. Slawische Sprachkorrelationen*. Graz: Institut für Slawistik der Karl-Franzens-Universität Graz.

Zabawa, Marcin. 2010. The influence of English on the Polish language of Internet message boards: Investigating the role of individual differences and the context. *Linguistica Silesiana* 31. 219–233.

Bionote

Aleksej Tikhonov
University of Zürich, Zurich, Switzerland
aleksej.tikhonov@uzh.ch
https://orcid.org/0000-0002-0772-3397

Aleksej Tikhonov is a linguist of the Department of Slavic Studies at the University of Zürich. In 2020, he defended his doctoral thesis at the Humboldt University of Berlin on the linguistic author identification of handwritten historical documents of a Czech refugee community in the 18th century' Prussia. Subsequently, he held a position in the collaborative research centre 1412 "Register" at the Humboldt University of Berlin and a Postdoctoral position as member of the collaborative research project on "The History of Pronominal Subjects in the Languages of Northern Europe," established between the Humboldt University of Berlin and the University of Oxford. Since 2020, Aleksej Tikhonov has been actively engaged in the development of transcription algorithms for the Multilingual Handwritten Text Recognition Project at the University of Freiburg. His expertise and research efforts are primarily dedicated to Czech, Russian, German, Polish, Ukrainian, and Yiddish. His research interests are: German-Slavic language contact, corpus linguistics, digital humanities, multilingualism in the post-migrant society, and (semi)automatic text recognition. In Zürich he is working on his habilitation on the Slavic languages in German rap and their identity constructing function.