Linguistica ONLINE
Issue Twenty Six
ISSN 1801-5336

Miscellanea

XV

July 2024 • Department of Linguistics and Baltic Languages, Masaryk University

Linguistica ONLINE

ISSN 1801-5336
electronic journal of the Department of Linguistics and Baltic Languages, Masaryk University, Czech Republic

home: **http://www.phil.muni.cz/linguistica/**
email: **linguistica@phil.muni.cz**

MISCELLANEA XV

http://www.phil.muni.cz/linguistica/art/issues/issue-026.pdf
published: July 26, 2024

CONTENTS

Issue Twenty Six

http://www.phil.muni.cz/linguistica/art/issues/issue-026.pdf

Sidney Martin Mota

Perceptual exploration of AO in diachronic AL > AU > AO > O
http://www.phil.muni.cz/linguistica/art/mota/mot-002.pdf
previously unpublished

Lidiia Melnyk

Understanding the dynamics of war-related Ukrainian tweets through BERTopic
http://www.phil.muni.cz/linguistica/art/melnyk/mel-001.pdf
previously unpublished

Linguistica ONLINE. Issue 26. Published: July 26, 2024
http://www.phil.muni.cz/linguistica/art/mota/mot-002.pdf
ISSN 1801-5336

PERCEPTUAL EXPLORATION OF AO IN DIACHRONIC AL > AU > AO > O[*]

Sidney Martin Mota (Escola oficial d'idiomes de Tarragona, EOI Tarragona, Spain, smart47@xtec.cat)

Abstract: The goal of the paper is to explore, by means of perceptual data, one of the possible intermediate steps, AO (Menéndez Pidal, 1968), in the diachronic sound change AL > AU > AO > O. AO is usually placed after the change AL > AU, as in Middle French (Bretos & Tejedor, 2015) and Middle Castilian (Menéndez Pidal, 1968). Ohala (1981) states that many sound change processes find their root in acoustic similarity, leading to signal misperception by the listener. Interestingly, AL and AU share similar acoustic information that can be misinterpreted (Recasens, 1996). In order to explore signal misperception in AL > AU > AO, and paying special attention to the transition AU and AO, ten subjects had to listen to the following stimuli AL, AO and AU uttered by a female Eastern Catalan speaker at two different speech rates, fast and slow. Subjects were forced to choose between AL, AU and AO in order to specifically explore AU categorization by the participants. Results show that AU can be perceived as AO. If AU is perceived as AO, then it may also be produced as such (Ohala, 1981), thus finding an opening for AO to emerge as another candidate in the sound change.

Keywords: diachronic, sound change, perception, acoustics

1 Introduction

The change AL > AU > O has historically been observed in the following Romance varieties: French (Vaissiere, 1996; Bretos & Tejedor, 2015), Castilian Spanish (Menendez Pidal, 1968), and Catalan (Recasens, 1996). Vulgar Latin already, albeit sporadically, presented cases of l-vocalization as in *cauculus* (Väänänen, 1963). In order to start the sound change, L should be velarized, which is acoustically similar to U in AU (Recasens, 1996). The sound change has another step, AU > O. Menendez Pidal (1968) makes a distinction between primary AU and secondary AU. The O which originated from a primary AU evolved from Latin AU > OU > O (*causa* > *cosa*, found in both Castilian and Catalan). On the other hand, the O which originated from a secondary AU is the result of Latin AL > AU > O *(altariu*>*otero*). Menendez Pidal (1968) also observed that the change from secondary AU > *O* yielded other candidates such as AO; however, O outlived the rest. According to Bretos and Tejedor (2015), French also yielded AO in the evolutionary path for l-vocalization: AL > AU > AO > O in words such as *alba* > *aube; cal(i)du* > *chaud; mal(i)fatius* > *mauvais* (Bretos &

[*] Previously unpublished. Peer-reviewed before publication. [Editor's note]

Tejedor, 2015). Both Menendez-Pidal (1968) and Bretos and Tejedor (2015) place the intermediate step AO after AU in AL > AU, but how did the change take place phonetically? I will briefly present two approaches that have tried to account for the change: the articulatory and the perceptual approaches. The present study will use concepts from both approaches to account for the historical sound change being explored.

As far as the articulatory approach is concerned, one can find Articulatory Phonology (AP) (Browman & Goldstein, 1992), which could, broadly speaking, account for this historical sound change by means of two articulatory mechanisms: *gestural overlap* and *gestural reduction*. For instance, l-vocalization, as in AL > AU, would be an example of consonant reduction, also known as *gestural reduction* in AP, where the tongue tip fails to complete the closure at the alveolar region of the palate. On the other hand, AU > O would be caused by *gestural overlap*, where two given articulatory configurations blend into one, thus yielding a new articulatory configuration which shares traits from the original ones. In addition, in this case, *gestural overlapping* would also be at work since labialization from the second element in AU, /w/, which is functioning at a different but simultaneous tier, would be exerting its own influence on the acoustic traits of the lingual configuration corresponding to the first element in AU, /a/, thus possibly reinforcing the percept of either AO or eventually O. Such an account is in line with what Penny (1993) describes as *reciprocal assimilation*, in which two phonemes blend into one intermediate pronunciation as in Latin *causa* > Spanish *cosa*.

However, the fact that the same articulatory configuration may be misinterpreted by the listener as one sound or another complicates matters further. This has already been proposed by Ohala (1981), which is the example of a perceptual approach to sound change. In our study, the ambiguous acoustic output of the lingual configuration for AU may be misinterpreted as AO. Ohala (1981) would add that coarticulation may be one of the reasons that distorts what the speaker intended to utter. For example, in the sequence /ut/ the acoustic output for /u/ is affected by /t/, thus yielding an acoustically more fronted back vowel than in isolated form. Speakers who are used to such coarticulatory effects are in a better position to discard such effects and thus interpret the acoustic information correctly. On the other hand, speakers who are not used to these effects may easily misinterpret it and fail to reconstruct the original intended sound.

Experimental research has been carried out to investigate the perceptual and production mechanisms of l-vocalization in various languages. For instance, Recasens (2012) states that there is some perceptual evidence that listeners hear a back rounded vowel when presented with a schwa+lateral sequence in Romance languages. In addition, Martin (2005) found perceptual evidence of listeners hearing AU when presented with AL in Romance languages as well. Both examples point at the fact that the historical sound change AL > AU has a perceptual component. Interestingly, perceptual evidence has also been found in Germanic languages such as English (Wong, 2013; Szalay et al., 2022) and Swiss German (Leemann et al., 2014), suggesting that the sound change may not be localized to only a specific family of languages and that there may be physical conditions on both speech production and perception triggering the sound change (Ohala, 1993). To the author's knowledge no previous research has been conducted focusing on the intermediate step "ao", which is why this present study will shed some light on this specific outcome in the historical sound change AL > AU > AO > O.

Having seen some examples of the possible articulatory, acoustic and perceptual underlying mechanisms of the sound change, I would like to draw the attention to the fact that

Menendez Pidal (1968) seems to suggest that the change from one intermediate sound to another may logically have implied a period of time in which different possible outcomes in the chain may have coexisted until one of them became the predominant candidate. It is beyond the scope of this study to account for the many variables outside the domain of perception which may have certainly conditioned the final outcome of the historical sound change. Gubian et al. (2023) suggest that a sound change is stochastic and speaker-specific in nature, which complicates matters when exactly determining the evolutionary path of any diachronic sound change. Their agent-based model (Cronenberg et al., 2023) provides the tools for investigating sound change, taking into account other factors than the strictly phonetic ones. For instance, they propose the existence of a speaker-specific phonology based on exemplars that may be the source of sound change processes. The speaker-specific phonology seems to be based on the speakers' experience with the specific sounds and lexicon in their own language. The present study deals with what Stevens and Harrington (2022) call the fine-grained (synchronic) phonetic biases. More specifically, what speakers of a language hear and how the same signal can be misinterpreted, thus triggering a possible sound change.

Based on Menendez Pidal (1968)'s observations about possible outcomes coexisting and utilizing perceptual data (Ohala, 1981), thus remaining in the more phonetic level of a sound change (Stevens & Harrington, 2022; Gubian et al., 2023; Cronenberg et al., 2023), I would like to explore the perceptual miscategorization of AU as AO in the AL > AU > AO > O sound change (Menendez Pidal, 1968; Bretos & Tejedor, 2015).

2 Method

2.1 Linguistic material

A perception forced-choice test was designed in which each stimulus had to be categorized as AL, AU or AO. The stimuli were obtained from a female speaker of Eastern Catalan in her 30's, saying the following words: *pal, pao* and *pau*, the phonetic transcriptions of which are: /pal/, /pao/ and /paw/. The reason why an Eastern Catalan speaker was chosen is because of the dark /l/ in this variety. Dark /l/s is a condition for l-vocalization to occur since dark /l/ and /w/ have similar acoustic patterns; that is, low F1 and F2 (Recasens, 1996). Ten repetitions were obtained for each token, which were inserted in the carrier phrase *Digues* _____ (Say _____). The speaker was instructed to read each sentence ten times at a self-chosen slow speech rate and then at a faster one. Faster rate, in this study, represents casual speech. The sentences were presented via PowerPoint. The subject was recorded with a Behringer XM2000 microphone connected to the mobile preamp audio interface M-Audio. EMU (Cassidy & Harrington, 2001) and Praat 6.3.10 were employed for signal processing. The sentences were sampled at 11,025Hz. The average duration of each group of stimuli (al-fast speech rate; ao-fast speech rate; au-fast speech rate; al-slow speech rate; ao-slow speech rate; au-slow speech rate) was calculated and then a representative of each group was selected for the perception test, which was the closest individual stimulus to the average duration in each group. The /p/ of the stimuli selected for the perception test was then manually removed in order to produce the following tokens: fast and slow *al*, fast and slow *ao*, fast and slow *au*.

2.2 Subjects

Ten subjects were recruited for the forced-choice test, all of whom reported they were bilingual speakers of both Spanish and Catalan. None of the subjects reported hearing problems. Six were women and 4 were men. Two subjects were in their 40's, four in their 50's and four in their 60's.

2.3 Experiment

Each stimulus was randomly presented five times with a total sum of 30 stimuli (6 stimuli x 5 repetitions). The ten subjects heard each stimulus through the same pair of headphones (SONY, MDR-ZX) connected to a laptop's internal soundcard (ASUS SonicMaster). Subjects had to complete a forced-choice test while the experimenter played each stimulus. The three options were AL, AO and AU. Each subject had 30 seconds to respond. No repetition was allowed. Miscategorization was analyzed using the average percentage of incorrect stimuli identification.

Chi-square test results were obtained from <https://www.socscistatistics.com/>.

4 Results

4.1 Phonetic context

The chi-square test indicates that the two variables (the phonetic context in the stimuli (AL stimulus, AU stimulus and AO stimulus) and the categories chosen by the listeners in the test (AL response, AU response and AO response) are associated with each other (X^2 (4, N = 300) = 231.7486, $p < .01$). The phonetic context has an effect on what listeners hear and it does so differently depending on the phonetic context. A summary of the results for each phonetic context can be found in table 1 (observed cell totals along with expected cell totals in parentheses). Miscategorization was present in AL, AU and AO. AL was miscategorized 5% of the times, AU 34% and AO 45%. AU was interpreted as AO (25%) and AO as AU (32%). In both cases the expected cell totals are very similar to the observed cell totals, indicating that the two phonetic contexts are very likely to be confused.

	AL response	**AU response**	**AO response**
AL stimulus	95 (39)	4 (34)	1 (27)
AU stimulus	9 (39)	66 (34)	25 (27)
AO stimulus	13 (39)	32 (34)	55 (27)

Table 1: contingency table for the three phonetic contexts: AL stimuli, AU stimuli and AO stimuli. The categories subjects chose when hearing one of the stimuli are in the top row (AL response, AU response and AO response). Observed cell totals along with expected cell totals in parentheses.

4.2 Speech rate

Generally, fast rate, representing a more casual speech in this study, yields higher miscategorizations (see table 2): slow (25.3%) vs fast (32.8%). More specifically, in two out of the three phonetic contexts: fast AL (8%) > slow AL (2%); fast AO (56%) > slow AO (34%). However, speech rate does not affect each phonetic context equally since the opposite pattern is observed in fast AU (28%) < slow AU (40%).

Stimuli	AL response	AO response	AU response
AL-slow	49	0	1
AL-fast	46	1	3
AO-slow	6	33	11
AO-fast	7	22	21
AU-slow	7	13	30
AU-fast	2	12	36

Table 2: contingency table in which the first column is the stimulus (AL slow/fast, AO slow/fast, AI slow/fast). The categories subjects chose when hearing one of the stimuli are in the top row (AL response, AO response and AU response).

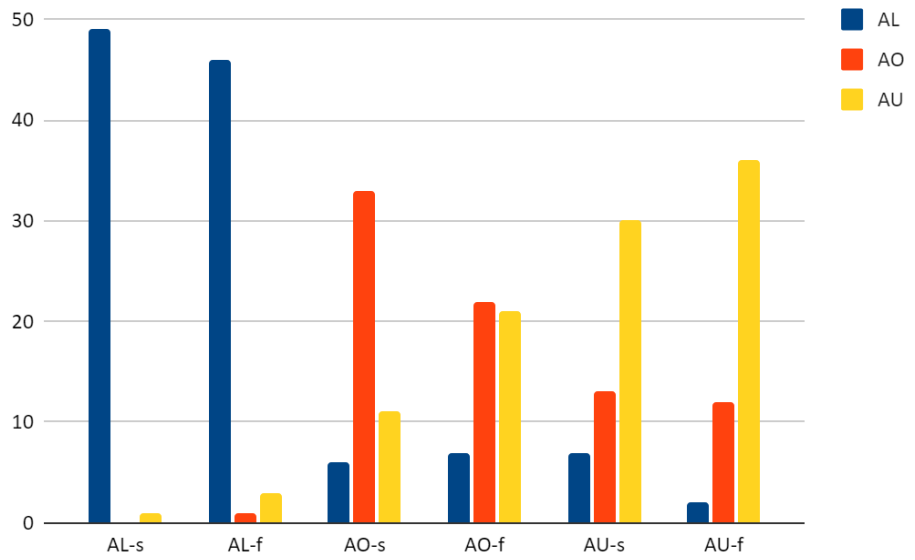

Figure 1: bar graph showing the results of the perception test by 10 subjects; y-axis shows number of times that the listeners chose AL (blue), AO (orange) and AU (yellow) for each stimulus; along the x-axis we find the labels of the stimuli used in the study (AL, AO and AU at different speech rates: fast and slow).

5 Discussion and Conclusions

The data points at the possibility of AU being perceived as AO. One cannot but wonder whether AU and AO coexisted, at least in the phonetic level but as time progressed, one of

the candidates would stand out more than the other depending on factors other than the purely phonetic ones. If we take a look at Figure 1 again, we can see that overlapping exists between the three candidates but in different degrees. AL is heard as AU and AO but showing very low percentages of miscategorization, whereas AU and AO yield higher percentages of miscategorization. Also, a faster speech rate, representing more casual speech, yielded higher miscategorizations, which may hint at the possibility that it is harder for listeners to reconstruct the intended sound (Ohala, 1981; Recasens, 1996). Interestingly, AO can be interpreted as AU and vice versa, which hints at the possibility of a period of confusion while transitioning from one candidate to another.

Of note, and to complicate matters further, both phonetic context and speech rate interact. For instance, AU slow was more miscategorized than AU fast, which is the opposite pattern observed in the other two phonetic contexts, AL slow vs AL fast and AO slow vs AO fast. Such differences are probably due to the acoustic properties of the F2 transition, suggested by Recasens (1996). It seems that the combination of F2 transition and slow rate has increased miscategorization in this specific phonetic context, AU, which leaves an open door for further investigation into the effect of speech rate and phonetic context on the emergence of candidates in a sound change.

It is also interesting to see that the same stimulus can be perceived differently, which could be an example of what Gubian et al. (2023) call in their model the speaker-specific phonology, one of the sources of sound change. A preliminary analysis of the results for each participant revealed that categorization was by and large performed differently. If the sound is interpreted differently, then it could consequently be articulated differently as well, thus contributing to the sound change. If speaker-specific phonology were inexistent, then all the speakers in the perception test would perceive the same sound, but this is not so in the study.

One must bear in mind that we are drawing conclusions based on synchronic perceptual data, which is only an approximation how speakers at different periods of time perceptually behaved, influenced at the same time by many variables. The assumption that speakers today behave like speakers in the past should be taken with precaution. Agent-based models like the one proposed by Gubian et al. (2023) are trying to account for and represent sound change in a more realistic manner by incorporating into their model not only phonetic details from acoustic, articulatory and perceptual studies but also the concept of probability in order to explain sound change influenced by many other variables.

It is also critical to take into account the lexicon of a language in which the sound change takes place. For instance, results from phonetic analyses may indicate that a specific phonetic context may be more prone to sound change but such a change is not observed in the diachronic data due to its low frequency in the lexicon (Martin, 2005). Therefore, accounts on sound change based solely on phonetic analyses should be considered with precaution, since sound changes are affected by many other variables (Gubian et al., 2023).

One of the limitations of the present study, apart from accounting for diachronic sound changes using synchronic data, is having few subjects in the perception test. More subjects are required in order to obtain more robust results. In addition, an acoustic analysis of the different phonetic contexts at the two speech rates is needed in order to investigate how the two factors interact. Correlating the perception and acoustic information will undoubtedly help us understand the underlying mechanism of this specific sound change at the phonetic

level, which would only be part of the explanation of the change as a whole, albeit key to understanding it. Further analysis of the responses of each of the participants in the study will be carried out in order to explore what Gubian et al. (2023) call in their model the speaker-specific phonology and the stochastic nature of a sound change.

References

BRETOS, Jesús – TEJEDOR, Didier. 2015. *Cahiers de phonétique diachronique de la langue française*. Madrid: UAM.

CRONENBERG, Johanna – GUBIAN, Michele – HARRINGTON, Jonathan – REICHEL, Vanessa. 2023. "Investigating sound change through computational agent-based modelling: An R package". In Skarnitzl, Radek – Volín, Jan (eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*. Prague, 3046–3050.

GUBIAN, Michele – CRONENBERG, Johanna – HARRINGTON, Jonathan. 2023. "Phonetic and phonological changes in an agent-based model". *Speech Communication* 147, 93–115.

LEEMAN, Adrian – KOLLY, Marie-José – BRITAIN, David – WERLEN, Iwar – STUDER-JOHO, Dieter. 2014. "The diffusion of l-vocalization in Swiss German". *Language Variation and Change* 26/2, 191–218.

OHALA, John. 1981. "The listener as a source of sound change". In Masek, Carrie – Hendrik, Roberta – Miller, Mary Francis (eds.), *Papers from the Parasession on Language and Behaviour*. Chicago, 178–203.

———. 1993. "Sound change as nature's speech perception experiment". *Speech Communication* 13, 155–161.

MARTIN, Sidney. 2005. "On the evolutionary path of l-vocalization in the Occitan spoken in Val d'Aran". *Linguistica occitana* 4, 42–59.

VAISSSIERE, Jaqueline. 1996. "From Latin to Modern French: On diachronic changes and synchronic variations". *AIPUK, Arbetisberitche, Institut für Phonetik und digitale Sprachverarbeitung* 31, 61–74.

RECASENS, Daniel. 1996. "An articulatory-perceptual account of vocalization and elision of dark /l/ in the Romance languages". *Language and Speech* 39/1, 63–89.

———. 2012. "Coarticulation in Catalan dark [l] and the alveolar trill: General implications for sound change". *Language and Speech* 56/1, 45–68.

STEVENS, Mary – HARRINGTON, Jonathan. 2022. "Individual variation and the coarticulatory path to sound change: agent-based modeling of /str/ in English and Italian" *Glossa: A Journal of General Linguistics* 7/1, 1–34.

SZALAY, Tünde – BENDERS, Titia – COX, Felicity – RROCTOR, Michael. 2022. "Reconsidering lateral vocalization: Evidence from perception and production of Australian English /l/". *Journal of Acoustical Society of America* 152, 2106–2116.

VÄÄNANËN, Veikko. 1963. *Introduction au latin vulgaire*. Paris: Librairie C. Klincksieck

WONG, Nicole. 2013. "The perception of /l/ vocalization by listeners with speech therapy as children". *Studies in the Linguistic Sciences: Illinois Working Papers* 38, 184–196.

Linguistica ONLINE. Issue 26. Published: July 26, 2024
http://www.phil.muni.cz/linguistica/art/melnyk/mel-001.pdf
ISSN 1801-5336

UNDERSTANDING THE DYNAMICS OF WAR-RELATED UKRAINIAN TWEETS THROUGH BERTOPIC[*]

Lidiia Melnyk (Friedrich Schiller University, Jena, lidiia.melnyk@uni-jena.de)

Abstract: The ongoing Russia-Ukraine war has amplified the significance of the digital sphere as a vital platform for information exchange and discourse. Using BERTopic's hierarchical and dynamic topic modeling, this study examines geolocated tweets from Ukraine in Russian, Ukrainian, and English. We propose three primary hypotheses: 1) Russian-speaking Ukrainians are adopting Ukrainian on social media, reaffirming their national identity; 2) Language choice among Ukrainians is context-dependent, reflecting pragmatic goals and targets of communication; 3) The dynamics of topics within the corpora reflects the media timeline and main information occasions of the development of military actions. Covering developments from February 2022 to June 16, 2023, this study provides insights into evolving linguistic dynamics and pragmatic choices amid the invasion, contributing to our understanding of multilingual discourse during challenging times.

Keywords: BERTopic, hierarchical topic modeling, dynamic topic modeling

1. Introduction

The language dynamics in Ukraine, especially in the context of the Russo-Ukrainian conflict, are a complex and evolving phenomenon, and various studies have highlighted the intricate relationship between language preference, national identity, and the impact of the conflict on linguistic choices (Barrington 2022; Chayinska et al. 2022; Pavlenko & Blackledge 2001).

The language landscape in Ukraine is diverse. Hentschel & Palinska (2022) report that in the Black Sea area, Ukrainian survey respondents exhibit diverse language preferences. Roughly 40 percent claim a single native language, with 30 percent identifying as Ukrainian speakers, 10 percent as Russian speakers, and less than 1 percent as Suržyk speakers. Surzhyk is a mixed language blending Ukrainian and Russian, prevalent in bilingual communities, particularly in central, eastern, and southern Ukraine. It combines elements from both languages, often characterized by its creative play with words and everyday use in private spheres, reflecting linguistic diversity and challenging societal divisions (Kostiučenko 2023). Another 40 percent state proficiency in two native languages, with various combinations of Ukrainian, Russian, and Suržyk. Approximately 17 percent consider all three languages as native. Moreover, Chayinska et al. (2021) emphasise that linguistic lines in Ukraine are not rigid, as many citizens in south-eastern regions use both Ukrainian and Russian interchangeably.

[*] Previously unpublished. Peer-reviewed before publication. [Editor's note]

The language issue is deeply intertwined with national identity, and it has become a focal point since the Russian invasion of Ukraine (Barrington 2022). The language one speaks is often seen as an integral part of national identity. Petriv (2022) argues that the conflict has led to a shift in attitudes towards the Russian language, fostering a broader adoption of Ukrainian and internal transformations within the language. This shift, known as "soft ukrainisation", is further accelerated by media adoption of Ukrainian and prominent Ukrainians publicly switching to it.

Given these social processes, we hypothesise a significant decrease in Russian content generated by Ukrainian users on Twitter. Furthermore, we anticipate differences in the most frequently discussed topics based on the language used. We assume that the topic dynamics within the corpora are representative of the main information occasions reported in the media.

In the context of the Russo-Ukrainian war, information warfare and disinformation campaigns have been prevalent, especially on social media platforms like Twitter (Chen & Ferrara 2023). Russian disinformation campaigns have been widely documented, aiming to influence public opinion domestically and abroad (Badawy, Ferrara & Lerman 2018). However, Ukrainians have also actively used social media to counter these narratives, seeking international support and promoting their perspective on the conflict (Cohen 2022; Garner 2022).

In response to these challenges, social media platforms have taken measures to combat misinformation (Cohen 2022). These efforts aim to limit the spread of false information and ensure the dissemination of accurate and reliable content. True and false information about the war plays a significant role in shaping public opinion, and this dynamic extends to social media (Tao & Peng 2023).

Researchers have leveraged large datasets from platforms like Twitter and Reddit (Zhu et al. 2022) to analyse user behaviour regarding the Russo-Ukrainian conflict (Chen & Ferarra, 2023). By employing topic-modeling techniques, we aim to dissect language usage nuances, content preferences, and audience targeting strategies. This approach allows us to identify linguistic choices tailored to specific audiences, providing insights into how different languages are utilised to address distinct user groups and their motivations in discussing the event.

2. Methodology

2.1 Data preprocessing

The corpus utilised in this study stems from the open-source corpus uploaded to Kaggle.1 We have filtered it to only include tweets with a unique tweet id geolocated in Ukraine, obtaining a total of 539,097 tweets. After removing retweets, a refined multilingual corpus of 435,135 tweets was obtained. The data collection period spanned from March 23, 2022, to May 16, 2023, in accordance with the API requirements of Twitter.

1 BwandoWando. (2023). (Sunset) UA Ukraine Conflict Twitter Dataset[Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/5934908

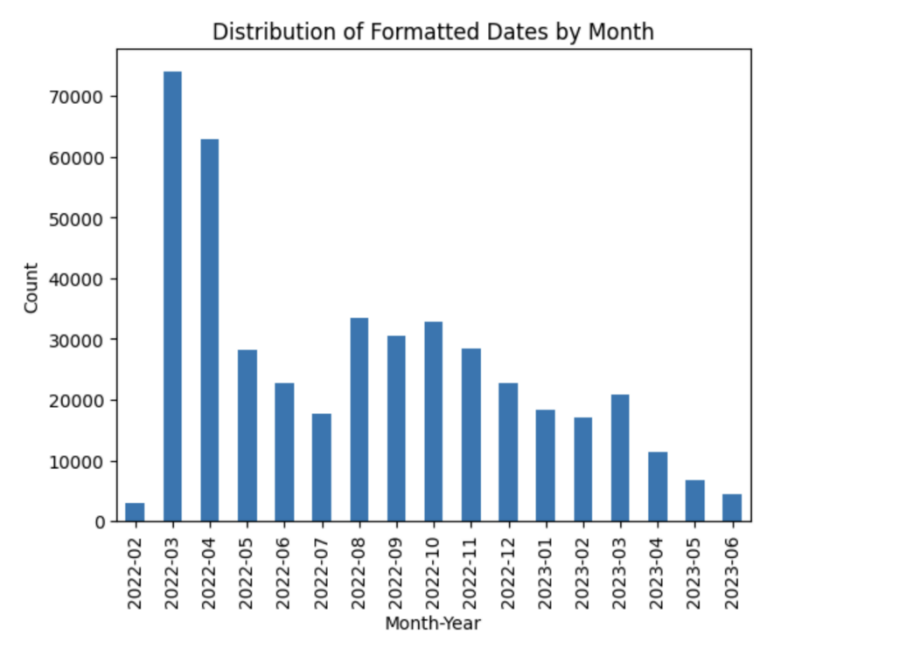

Figure 1: Distribution of tweets per month

The distribution of tweets (Figure 1) exhibits variations across different months, with the absolute peak observed in March 2023. Throughout the research period, the monthly tweet count consistently exceeded 20,000, indicating a substantial volume of data available for analysis.

We have decided to focus only on tweets in English, Russian and Ukrainian, resulting in a final dataset size of 289,806 tweets tagged as being in one of the research languages. Given the absence of language tags in some tweets, a language check was performed using the cld3-based (compact language detector v3) approach. This language detector, developed by Google, leverages neural networks and supports a wide array of languages, providing a novel and comprehensive approach to language identification (Foong 2021). Adding the language detection step, we were able to increase our corpus by 41.095 tweets. In the multilingual corpus, English tweets constituted the largest portion, with a total of 256,468 instances. Ukrainian tweets were the second most prevalent group, comprising 53,932 instances. Russian tweets were less common in the corpus, accounting for only 20,501 instances.

As we can see (Figure 2, next page), tweets in the Russian language are not that common in the corpus, ranking as only the fifth biggest group. The setting of the geolocation of the tweets to Ukraine was the only method available to us to filter the corpus so that it includes only users based in Ukraine. Users do not need to make their geolocation available by default, but can opt in to do so.

In the first stage of our research, we did not carry out any preprocessing on purpose in order to observe whether it has any impact on topic modeling. We assumed that proceeding without preprocessing might positively improve keywords and topic selection as the model might fit its predictions to the hashtags or use the number within the tweets as a keyword itself as the numbers might be used to report casualties as well as to trace the dates of specific operations. This resulted in hashtags and numbers being the most common keywords.

Therefore, we decided to further fine-tune our trained model after the post-processing of the tweets. For each of our three languages we removed digits, Twitter handles, URLs, and

special characters. We also removed stop words. We removed the non-Cyrillic words as long as they are not a part of a hashtag. Even though we have seen some disruption in the topic keyword due to the dominant presence of the hashtags, we considered that hashtags might be pivotal to the understanding and clustering of the tweets and can sometimes themselves be part of a narrative, as will be demonstrated in subsequent sections (see sections 3.2; 3.3; 3.5). In fact, Alash and Al-Sultany (2020) suggest that keeping hashtags in a corpus might actually improve the quality of topic modeling.

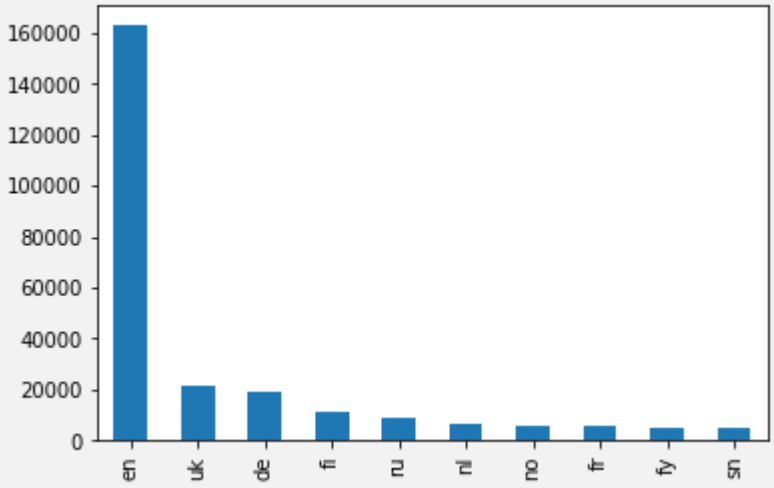

Figure 2: Distribution of tweets per language

2.2 Topic modeling

Topic modeling, a computational technique introduced by Blei, Ng, and Jordan (2003), is utilised to unveil latent thematic structures within text collections. This method identifies dominant themes and discursive patterns, revealing the rhetorical strategies and framing techniques employed in a given discourse (Blei et al. 2003). It automates the extraction of latent topics, providing insights into shared themes across texts and enabling the study of meaning construction, power dynamics, and discursive practices (Blei et al. 2003).

Topic modeling is frequently applied to identify the key narratives and ideas in the case of crisis or war. There is already extensive research built on topic modeling application around the COVID-19 pandemic (Mathayomchan et al. 2023; Wicke & Bolognesi 2020; Qin & Rochieri 2020). Karpina & Chen (2022) applied topic modeling to the tweets of the selected users to mine the narratives of British politicians around the war in Ukraine, while Shultz (2023) utilised it to identify the impact of the Russian's government tweets on the Russo-Ukrainian war. We are applying hierarchical topic modeling with BERTopic to the robust Twitter corpus without preselecting/filtering users to be representative of a certain group. Therefore, we believe that with our data and suggested topic modeling approach we will be able to conduct a comprehensive analysis of the prevalent topics and narratives in Ukrainian society within the research timeframe.

BERTopic, as described by Grootendorst (n.d.), leverages transformer-based language models and c-TF-IDF to create dense clusters, ensuring interpretable topics while retaining

essential words in topic descriptions. It involves three stages: obtaining document embeddings, grouping embeddings into semantically similar clusters, and creating topic representations for these clusters (González-Pizarro & Alavi n.d.).

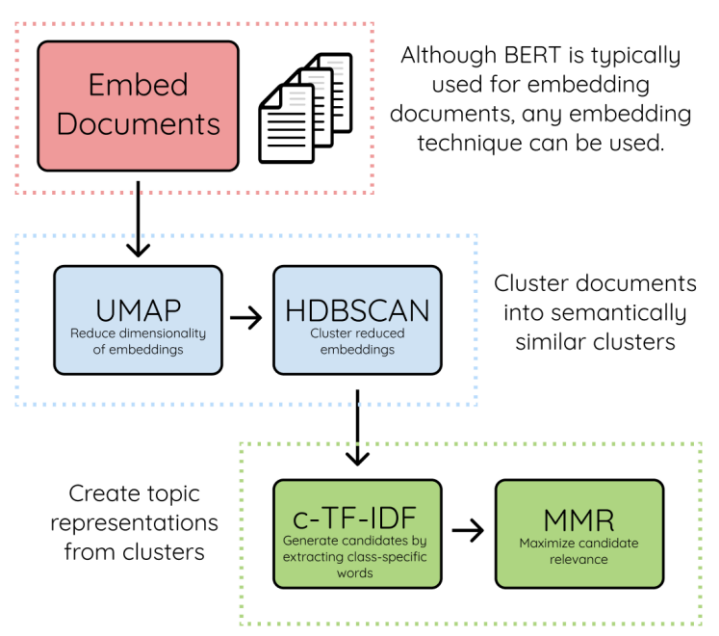

Figure 3: Structure of BERTopic

Figure 3 describes the structure of BERTopic, where at a first-stage BERT models are being applied for embedding. Then, UMAP is applied for dimensionality reduction and HDBSCAN is used to carry out the clustering. 'In the third stage, BERTopic exploits from c-TF-IDF to generate representing topic keyword candidates' (González-Pizarro & Alavi n.d.). Within this stage Maximal Candidate Relevance (MMR) is calculated to improve the selection of candidate keywords and documents.

After creating a topic model for our corpora, we employed hierarchical topic modeling to reduce dimensionality and determine the optimal number of topics automatically. BERTopic's clustering mechanism not only identifies the most suitable number of topics but also enables us to recognise topic similarities, facilitating the grouping or joint analysis of related topics (Grootendorst n.d.). The hierarchical model was generated using Scipy's (Scipy n.d.) linkage function, which employs the Nearest Point Algorithm method to calculate the distance (d(s,t)) between two clusters (s and t). This approach streamlines the process, as described by the formula:

$$d(s,t) = \min(\text{dist}(s[i], t[i])$$

BERTopic offers an option to automatically label topics based on the three most frequent words within each topic. However, these labels may not always provide a clear understanding of the topic's content. To enhance topic labelling, we applied a Bart-based zero-shot classifier after reducing the topic dimensionality. This classifier helped in identifying the most suitable labels from the options we suggested after reviewing the topic modeling results.

Additionally, we employed dynamic topic modeling in BERTopic, allowing us to track topic development over time. This feature calculates topic representations for each selected

period without the need to rerun the model multiple times, as outlined by Grootendorst (n.d.). The nature of our corpus enables us to identify distinct time spans for each topic, facilitating the application of dynamic topic modeling.

3. Results and discussion

3.1 Language distribution

A steady decline in the number of Russian tweets originating from geolocations within Ukraine was expected as more Ukrainians were anticipated to shift towards using the Ukrainian language for their external communication, as has been outlined by Racek et al. (2023).

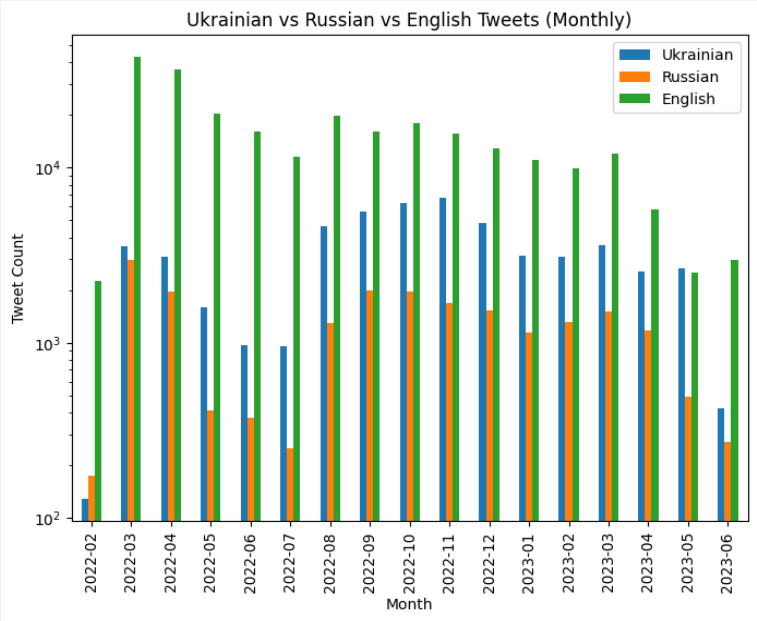

Figure 4: Language distribution

In Figure 4, a logarithmic scale plot illustrates the distribution of tweets in three primary languages. Initially, Russian tweets predominated over Ukrainian ones, maintaining this pattern until April 2023. However, there was a noticeable drop in Russian tweet frequency from May to July 2022. Subsequently, starting in August 2022, Russian tweet dynamics closely mirrored Ukrainian ones. This indicates a temporary decline in Russian tweets during the summer months of 2022, but no significant abandonment of the Russian language in external communication on the topic of war.

These findings imply that despite the ongoing war, the use of Russian in outward communication did not substantially decrease. This could be attributed to a pre-existing shift towards Ukrainian even before the invasion (Racek et al. 2023). It is worth noting that our research is limited to the topic of war, and language usage patterns may differ in other contexts. These observations warrant further investigation into the factors influencing language preferences and the relationship between language use and socio-political dynamics during the study period.

3.2 Topic modeling of the Russian corpus

We integrated our Russian corpus into the multilingual BERTopic model using embeddings from the DeepPavlov/distilrubert-base-cased-conversational model, a Russian-language model with 6 layers, 768 hidden units, and 12 attention heads, trained on Conversational RuBert (DeepPavlov/distilrubert-base-cased-conversational, n.d.). Our BERTopic configuration included bi- and tri-grams, with a minimum topic size of 10, influenced by the corpus' modest scale. We also introduced lemmatisation during preprocessing to standardise the corpus and enhance topic identification.

Figure 5: Distribution of the topics within tweets in Russian

The initial BERTopic run detected 106 topics, with 4,107 tweets classified as uncategorised and excluded from further analysis. Subsequently, through hierarchical topic modeling, we merged some small, detailed topics, reducing the number of topics in the Russian corpus to 70. We chose to analyse the top 15 topics due to their small sizes, making them less suitable for analysis.

As we can see from the distribution of the topics in Figure 5, only one topic, 'Ukrainian advances', is focused on the successes of the Ukrainian military itself, whereas most of the topics seem to focus on the losses and impact of it on the Russian military such as 'Threats',

‘Anti-Russian Sentiment’ and ‘No losses’. We spotted some derogatory slurs such as *shameless jackals (позорные шакалы)* as well as swear words. We can also observe the use of dehumanisation when referring to the Russian soldiers as in using the term ‘Cargo 200’ (груз [200]). The term is used in the Ukrainian social media to hint at the massive losses of the Russian military, hence ‘Cargo 200’ as the way of transporting the deceased soldiers.

Figure 5 contains some of the topics focused on the casualties and attacks in specific regions predominantly in the east and south of Ukraine including Kherson, Bakhmut and Mariupol, where the use of Russian to talk and report on the war might be due to its regional dominance in daily use. We assume that the tweets are coming either from the local media or locals living there. Due to the scope of the research, we were not able to focus on the individual authors.

An additional aspect pertains to the perceived focus of the corpus on Russian readers and soldiers stationed within Ukrainian territory. This is exemplified by topics such as ‘Nazism’ and ‘No losses’, which shed light on the dissemination of misinformation and disinformation prevalent within the Russian media landscape. Notably, the ‘Nazism’ topic actively employs prominent Russian hashtags that aim to encourage public support for Russia’s military actions. This topic frequently employs hashtags like #nazism and #fascism in the middle of the Russian patriotic hashtag list, setting an ironic sub-tone to the tweets.

Additionally, the inclusion of the hashtag ‘RuZZia’, featuring a Z symbol used as a marking on Russian military vehicles, further strengthens the dissonance between the meaning of the Russian state media agenda and the perceived consequences of it. This term is associated with ‘Nazi-inspired’ attributes (Staalesen 2022). The ‘RuZZia’ hashtag opposes the following hashtags:

- 'своих не бросаем' meaning 'we don’t leave ours [behind]'. This is one of the well-known Russian propagandistic slogans, used to justify the invasion as protection of the Russian speaking people ('our' people). The effectiveness of it can be traced down to the annexation of Crimea (Yagodkina 2020).
- 'МнеНеСтыдно' (#IAmNotAshamed). This is another pro-war hashtag, aimed to increase Russian group unity and express support for the Russian actions in Ukraine (*How Russia attempts to widen its arsenal of pro-war propaganda* 2022).

However, within the topic we discover sentences such as:

- '😢Оккупанты снова обстреляли Харьков.К сожалению, погибли 5 человек, среди них – 9-летний мальчик.#UkraineRussiaWar #war #WARINUKRAINE #война #ZaРоссию #zапобеду #своихнебросаем #zамир #украина #Харків https://t.co/DxpO8BKH9g '/😢Occupiers shelled Kharkiv again. Unfortunately, 5 people died, including a 9-year-old boy.#UkraineRussiaWar #war #WARINUKRAINE #war #ForRussia #zavictory #we don’t leave our own #zamir #Ukraine #Kharkiv https://t.co/DxpO8BKH9g'
- #ИркутскRU не дошел до КиеваUA... Эстафета передается следующим 200-ым#ukraine #russia #war #UkraineWar #UkraineUnderAttack #UkraineRussianWar #RussiaUkraineWar #RussiaUkraine #Москва #Украина #Россия #иркутскаяобласть #Москва #Сибирь #z #v #своихнебросаем #мыроссией #мывместе

> https://t.co/feKxpbFo1O/ #IrkutskRU did not reach KievUA... The baton is passed to the next 200#ukraine #russia #war #UkraineWar #UkraineUnderAttack #UkraineRussianWar #RussiaUkraineWar #RussiaUkraine #Moscow #Ukraine #Russia #Irkutsk region #Moscow #Siberia #z #v

The first tweet conveys distressing news about renewed shelling in Kharkiv, attributing it to the occupiers. The emotional tone is accentuated by mentioning casualties, including a 9-year-old boy. The sarcastic use of the hashtags utilised by the Russian media contrasts with the events described in the tweet to show the consequences of the actions of the military on the civilian population. The second tweet above mocks the advances of the Russian military emphasising its losses. It can also be interpreted as a threat to be seen by the Russian soldiers due to the reference to '200' as in 'Cargo 200' explained above.

A deeper look at the topic might be necessary to determine whether its content aims to ridicule and expose disinformation and propaganda or whether it includes solely propagandistic tweets, published from a Ukrainian geolocation, aimed at the Russian-speaking population of eastern Ukraine. The topic 'Russian invasion', on the other hand, does not seem to be propaganda-dominated, as it does not include any specific vocabulary used and promoted by the Russian media such as, for example, 'special military operation'. It includes phrases such as 'Russian invasion' and 'War with Ukraine' instead.

The tweets within the corpus are also directly targeted at Russian readers, aiming to confront them with the other sources of information. On this basis, we were able to identify the topic we named 'True information'. Within this topic, the tweets pledge to provide access to the 'true' information:

- 🔴 LIVE: Телеканал UATV информирует граждан России о реальной ситуации в Украине. Подключайтесь к эфиру государственного украинского вещателя на всех платформах
  https://t.co/Bcl7dUPI6S
  #stoprussia #StopRussianAgression/
- 🔴 LIVE: UATV channel informs Russian citizens about the real situation in Ukraine. Connect to the air of the state Ukrainian broadcaster on all platforms.
  https://t.co/Bcl7dUPI6S
  #stoprussia #StopRussianAgression

We can, therefore, assume that the tweets within this topic are primarily targeted at a Russian audience, encouraging them to consult other sources of information, but also hinting at the misinformation in the Russian media.

In general, we can see from Figure 5 that the corpus of the Russian tweets coming from the Ukrainian geolocation is focused on reporting the situation in the country with an emphasis on the eastern and southern regions. Some of the topics pursue the goal of being read by Russian readers, either to convince them of the information being hidden from them, ridicule the Russian media, or impose threats to discourage them from participation in the military action on Ukrainian territory. Notably, one of the frequent topics, 'Russian news', does not appear to follow the aforementioned goal, but rather contains news reports on the different events happening in Russia, which are not necessarily war-related.

3.2 Dynamic topic modeling of the Russian corpus

As with the English corpus, the development of topics over time is defined by the media reporting on the topic or the timeline of the events in Ukraine. The development of the topics is displayed in figures 6 and 7.

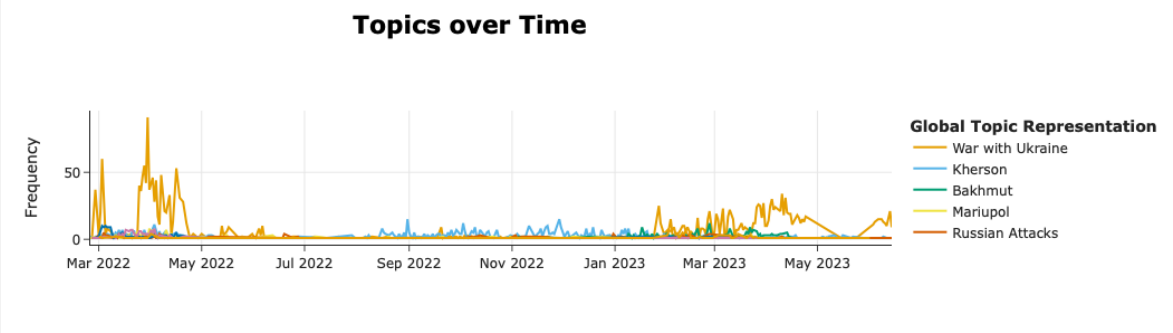

Figure 6: Development of topics in the Russian corpus over time. Part 1

During the period late February to early March, a notable surge in tweet activity around the 'War with Ukraine' topic is evident, followed by a subsequent decline until winter 2023. Additionally, minor fluctuations are observable in the evolution of the 'Kherson' topic, corresponding to the swift advances of the Ukrainian military in the southern region, culminating in the reclaiming of Kherson.

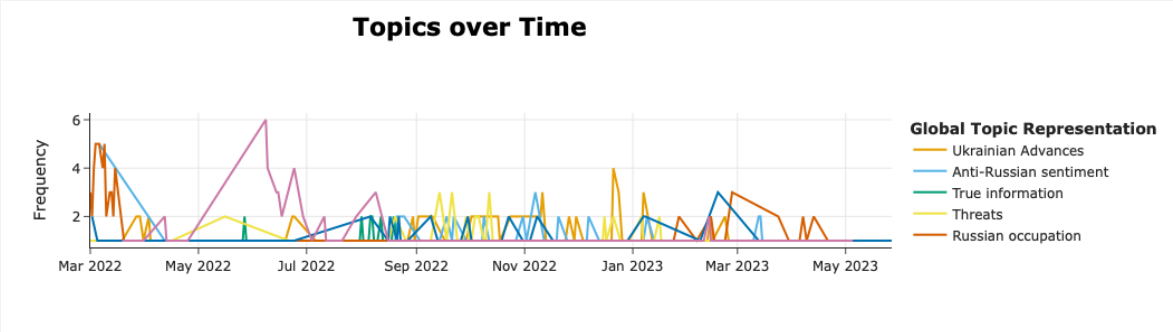

Figure 7: Development of topics in the Russian corpus over time. Part 2

Particularly noteworthy are the surges in tweets related to 'True information' topics, aligning with achievements of the Ukrainian military, including the liberation of sections of the Kharkiv region in June 2022 and progress in the southern territories during autumn 2022. Notably, the patterns for 'Anti-Russian sentiment' and 'Russian occupation reporting' lack steady continuity, instead emerging in response to other contextual events. The topics on Figure 7are rather scattered, which might also be due to the relatively small sizes of these topics, resulting in some semantically connected tweets posted over a short period of time being misattributed as thematically connected.

3.3 Topic modeling of the Ukrainian corpus

For our research paper, we utilised a pre-trained Ukrainian language model, ukr-models/xlm-roberta-base-uk, a transformer-based language model specifically designed for the

Ukrainian language. It is based on the XLM-RoBERTa architecture, which leverages bidirectional encoders to produce contextualised word embeddings. By utilising this language model as the embedding model for BERTopic, our research aimed to benefit from its language-specific capabilities to obtain high-quality embeddings and achieve accurate topic clustering for the analysis of Ukrainian social media discourse during the conflict.

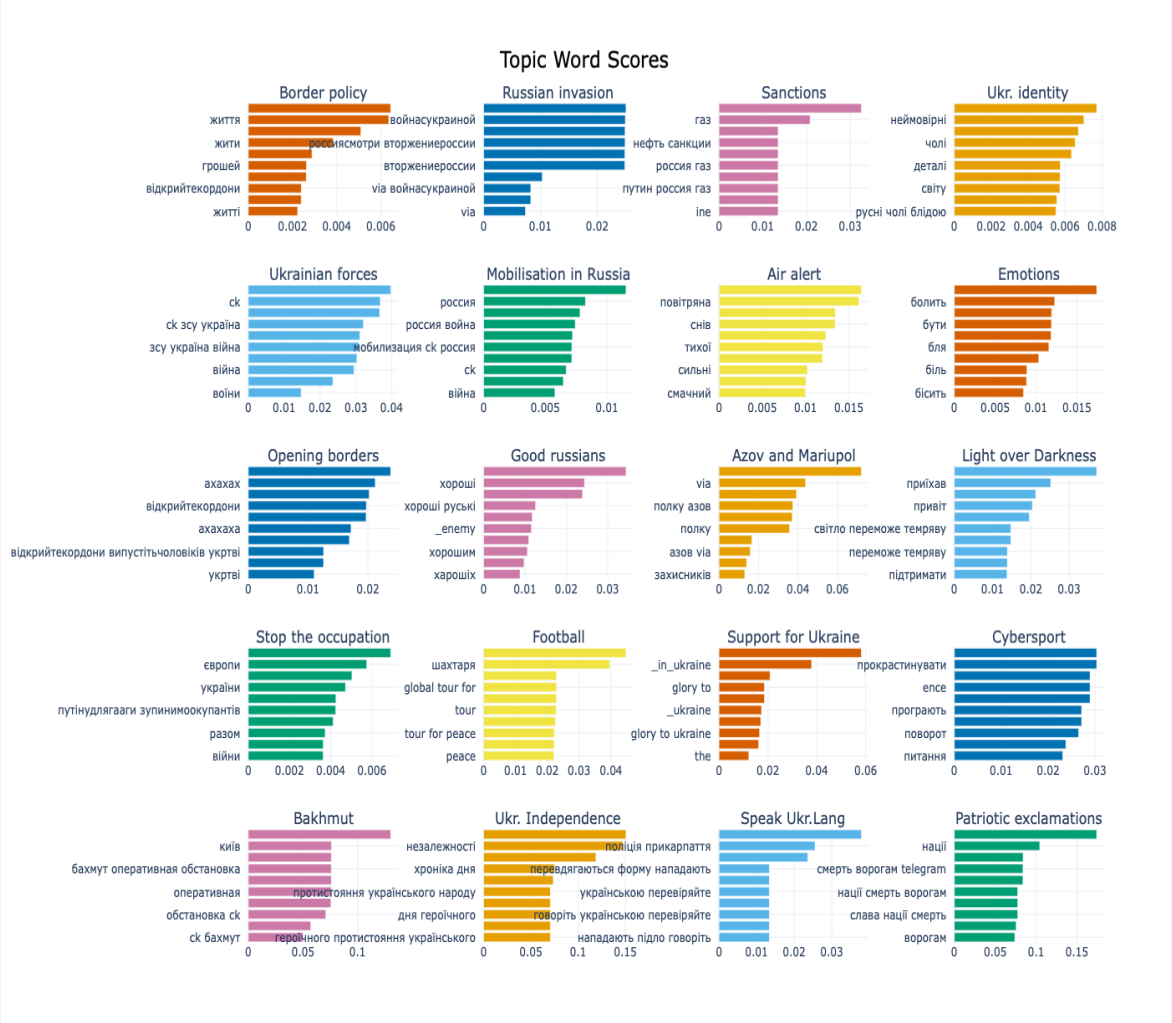

Figure 8. Distribution of the topics within tweets in Ukrainian

For the BERTopic model, the n_gram_range was set between mono- and tri-grams. Minimal topic size was at 25 to ensure more robust topic identification. First, we have identified 320 topics, as with the Russian topic model, the 'unidentified' topic being the most represented one, comprising 9,127 tweets. Then, after the hierarchical topic modeling merged topics with over 0.7 cosine similarity score, we ended up with a total of 132 topics. The 20 most common topics are represented with the custom topic labels on the visualisation in Figure 8.

Upon careful examination of the preprocessed data, we noticed that traces of both the Russian and English languages persist, potentially stemming from hashtags or proper names. Automatic differentiation between Russian and Ukrainian poses a significant challenge due to the close relationship between these languages within the Slavic language family. Moreover, the Russian spoken in Ukraine exhibits distinct characteristics from the Russian spoken

in Russia. Additionally, code-switching often involves mixed sociolects, further complicating automated language identification processes (Lyudovyk & Pylypenko 2014).

Despite the dominance of war-related themes in the Ukrainian corpus, including subjects like 'Russian invasion', 'Ukrainian forces', 'Air alarms', and 'Azov and Mariupol', a nuanced spectrum emerges. This spectrum encompasses seemingly unrelated topics on the more 'civilian' end, such as 'Cybersport' and 'Football'. Upon delving deeper into representative documents, the connection between these seemingly disparate themes and the overarching corpus theme becomes apparent. The topic of football, for instance, aligns with the corpus theme through expressions of support for Ukraine during football matches or charitable events designed to raise donations. Conversely, the presence of cybersport-related content seems accidental, resulting from the utilisation of war-related hashtags for tweet virality.

While hierarchical topic modeling has already been undertaken, visualisations reveal the convergence of two topics: 'Border policy' and 'Opening the borders'. These topics, absent from the Russian corpus, exhibit an interconnectedness that merits further investigation. Within the 'Border policy' topic, emphasis is placed on advocating for the opening of borders as a means of survival.

- Заїбали мусолити Маска. У нас в країні демократія перероста̀є в авиократію, урядом ігнорується конституція, співвідношення ціни/зп провальна, кордони закриті через що рівень життя людей, рівень економіки та довіри до уряду летить в тартарари, а нікому і діла нема?#відкрийтекордони [*original spelling preserved - L.M.*]/
  They are annoyingly bothering with Musk. In our country, democracy is turning into an autocracy, the government ignores the constitution, the price-to-income ratio is disastrous, borders are closed, leading to a decline in people's quality of life, the economy, and trust in the government plummeting, and no one seems to care? #openthebordersThe tweet begins with a tone of annoyance towards Elon Musk, highlighting the perceived distraction from pressing issues. It criticises the transition from democracy to autocracy and the government's constitutional violations, emphasising concerns about eroding democratic values. The reference to economic problems and closed borders underscores the impact on citizens' well-being, while the use of a rhetorical question at the end queries the lack of action or concern from those in power, urging a change in border policies.

- @APUkraine Випустіть людей яким є де вижити цю зиму, закрили без світла, тепла, грошей, можливостей ще й під бомбами!
  UAВільний народе вільної країни?🤷‍♀️🤦‍♀️ Дайте право ЖИТИ! #відкрийтекордони #УкрТві/
  @APUkraine Release people who have somewhere to survive this winter, they are trapped without electricity, heat, money, and even under bombs! UA Free people in a free country? 🤷‍♀️🤦‍♀️ Give them the right to LIVE! #openthebordersThis tweet conveys a plea to the official Twitter account of the Ukrainian Armed Forces (@APUkraine) to allow people to leave the country if they have a place to survive the winter. It paints a dire picture of the situation, highlighting the lack of necessities like electricity and

heat, as well as the threat of bombings. The use of emojis and hashtags (#opentheborders, #UkraineTwitter) adds emotional appeal and emphasises the urgency of the situation, urging authorities to grant people the right to live safely. In conclusion, we assume that the 'border policy' tweet is targeted at the Ukrainian government, while the pledge to open the borders is either motivated by general disagreement with the political decision to close them or is seen as a means of survival.

Conversely, 'Opening the borders' delves into more specific appeals, with users urging that male Ukrainians be allowed to go across borders.

- 🚧Поверніть нам наші права та свободи і можливість нормального існування. #відкрийтекордони #випустітьчоловіків @ZelenskyyUa @ZelenskaUA @r_stefanchuk @oleksiireznikov @Podolyak_M @DmytroKuleba @FedorovMykhailo @bihusinfo @sternenko @serhiyprytula @Luganskiy_Twi @serhiy_zhadan /🚧Give us back our rights and freedoms and the opportunity for a normal existence. #opentheborders #letmenout @ZelenskyyUa @ZelenskaUA @r_stefanchuk @oleksiireznikov @Podolyak_M @DmytroKuleba @FedorovMykhailo @bihusinfo @sternenko @serhiyprytula @Luganskiy_Twi @serhiy_zhadan

The author of the tweet makes a plea to various Ukrainian officials, including President Zelenskyy (@ZelenskyyUa) and First Lady Zelenska (@ZelenskaUA), to restore the rights and freedoms of the people. It emphasises the desire for a normal existence and includes hashtags and mentions of relevant individuals and organisations to draw attention to the call for action. The use of emojis and the construction of the tweet make it emotionally charged and urgent, conveying a strong desire for change.

- Наскільки гірше росія, яка прийшла нас всіх вбити, за UAуряд який не дозволяє нам від цих ракет, цієї небезпеки врятуватись? Філософське питання, чи не так?🤔 #відкрийтекордони #випустітьчоловіків #УкрТві/How much worse is Russia, which came to kill us all, than the UA government that doesn't allow us to escape from these missiles and this danger? A philosophical question, isn't it? 🤔 #opentheborders #letmenout #UkraineTwitter

The tweet above raises a philosophical question comparing the perceived threat posed by Russia and the restrictions imposed by the Ukrainian government. It suggests that both Russia's actions and the government's limitations are seen as endangering the population. The use of emojis and hashtags conveys a sense of contemplation and urgency, encouraging discussion of the topic. As we can see, the 'open the borders' topic employs a victim narrative, emotional appeal, blame attribution, urgency, and a touch of philosophy to engage and mobilise readers. It aims to sway public opinion, elicit support, and provoke discussion on the pressing issue of closed borders and its impact on Ukrainian citizens.

The Ukrainian corpora also uncovered topics centred around the foundational aspects of Ukrainian identity and the pivotal role of language within it. This thematic exploration is evident in the topics 'Ukr. [Ukrainian] Independence', 'Ukr. [Ukrainian] identity', 'Speak Ukr. [*Ukrainian - L.M.*], and 'Patriotic exclamations'.

- українці UA неймовірні https://t.co/G4WFKV8XZL/
  Ukrainians UA are incredible.

The author expresses admiration and praise for Ukrainians, likely in response to a positive or inspiring event or action associated with Ukraine. The use of the Ukrainian flag emoji adds a sense of national pride and solidarity. It is a concise and positive statement that celebrates the people of Ukraine.

- Чернетка Акту проголошення незалежності України, написана Левком Лук'яненком
  Першим варіантом було «відновлення»
  Виправлено на «проголошення» https://t.co/AbBSL2Jw9l/
  A draft of the Act of Independence of Ukraine, written by Levko Lukyanenko. The initial version was 'restoration'. Corrected to 'declaration'.

The mention of the initial word 'restoration' and its correction to 'declaration' suggests a nuanced understanding of Ukraine's history, potentially alluding to the country's periods of independence as well as external rule in its past. This historical context underscores the significance of Ukraine's declaration of independence and its enduring struggle for sovereignty.

Notably, the topic 'Speak Ukr. [*Ukrainian - L.M.*]' features direct calls to employ the Ukrainian language (говоріть українською/speak Ukrainian; українською перевіряйте/check in Ukrainian). In this context, the Ukrainian language is regarded as an integral facet of national identity, distinguishing the local Ukrainian population from Russian soldiers. The Ukrainian language serves as a litmus test (українською перевіряйте/check in Ukrainian), a means to discern whether a soldier is Ukrainian or Russian in times of uncertainty.

- @ua_industrial Окупанти перевдягаються в нашу форму і нападають підло. Говоріть з усіма українською, перевіряйте документи
  Репост!
  #Україна #Ukraine #російська_агресія
  #StopWar #StopRussia #StandWithUkraine #UkraineRussia #CloseTheSky #StopPutin #NATOINUKRAINENOW #NATOmustbeinUkraine/
  @ua_industrial Occupiers are dressing in our uniforms and attacking treacherously. Speak to everyone in Ukrainian, check documents. Repost! #Ukraine #Ukraine #RussianAggression #StopWar #StopRussia #StandWithUkraine #UkraineRussia #CloseTheSky #StopPutin #NATOINUKRAINENOW #NATOmustbeinUkraine

This tweet highlights a security concern, urging vigilance in verifying the identity of individuals, especially those wearing Ukrainian uniforms. It also emphasises the importance of using the Ukrainian language and checking documents to prevent infiltration by hostile forces. The use of relevant hashtags and mentions underscores the broader context of Russian aggression in Ukraine and calls for international support from NATO. The tweet aims to raise awareness and promote safety measures in the face of ongoing conflict.

In summary, the analysis of the Ukrainian corpus, using the ukr-models/xlm-roberta-base-uk language model and BERTopic clustering, has highlighted the profound meanings

embedded in the topics. Topics such as ‘Open the Borders’ underscore the urgent human appeals for safety, while discussions on Ukrainian identity and language reveal the significance of cultural preservation amidst the war. These topics collectively depict a narrative of resilience, identity, and the human impact of the conflict, offering crucial insights into the multifaceted concerns and aspirations of the Ukrainian population during this tumultuous period.

3.4 Dynamic topic modeling of the Ukrainian corpus

The visual representations provided in figures 9 and 10 illustrate the evolution of topics within the Ukrainian corpus over time. Notably, a conspicuous surge in tweet activity concerning the topic of ‘Russian invasion’ emerged in the spring of 2022. This surge is subsequently followed by a gradual decline and relative dormancy in the discussion of the topic. This pattern suggests a potential reallocation of war-related news to other thematic areas, including ‘Ukrainian forces’, ‘Air alert’, and ‘Stop the occupation’.

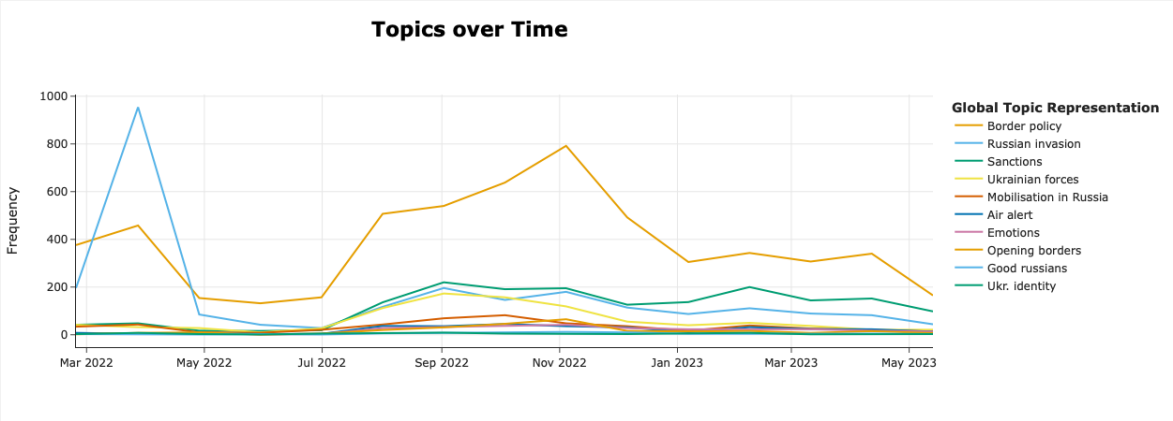

Figure 9: Development of topics in the Ukrainian corpus over time. Part 1

Additionally, the topic of ‘Border policy’ stands out as one of the most frequently discussed subjects, with noticeable spikes in both the spring and autumn of 2022. These spikes in discourse signal heightened engagement with matters pertaining to border policies during these periods. The progression of these topics corresponds closely with the broader news timeline. For instance, the ‘Azov and Mariupol’ topic exhibits two notable peaks, with the initial surge reaching its zenith in May 2022. This timing coincides with the moment when the defenders of Azovstal laid down their weapons between the 16th and 20th of May. The second peak can be attributed to news of the release of Azov commanders from captivity in September 2022, sparking renewed interest in the topic.

Likewise, the topic of ‘Bakhmut’ garnered increased attention during periods of heightened military activity around the city. This alignment between topic dynamics and significant events in the region underscores the responsiveness of online discussions to developments in the real world.

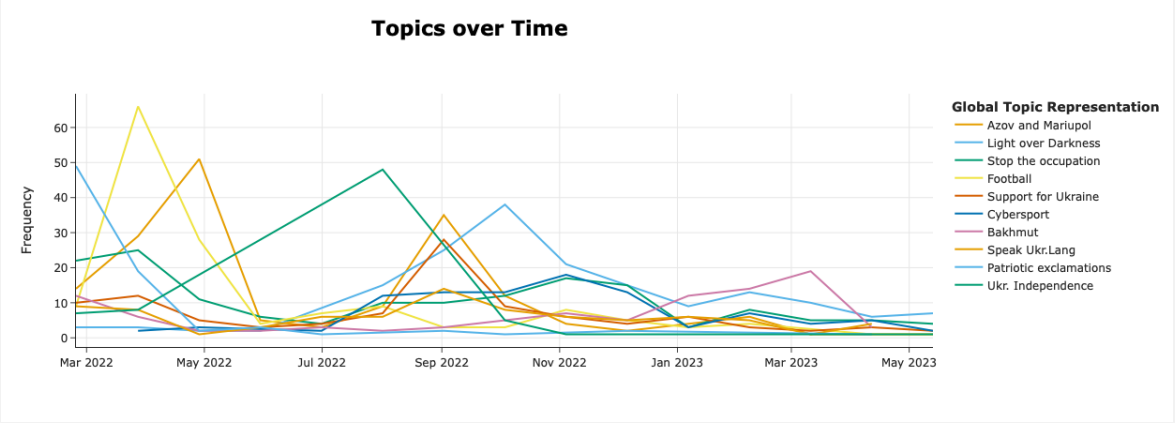

Figure 10: Development of topics in the Ukrainian corpus over time. Part 2

3.5 Topic modeling of the English corpus

To analyse the English tweets, DistilBERT embeddings were employed. DistilBERT, developed by Hugging Face, is a lighter and faster version of the original BERT model while maintaining comparable performance. It achieves this efficiency by distilling the knowledge from the larger model through a teacher-student training process. By utilising DistilBERT embeddings, we aimed to capture the contextual representation of the text, enabling more effective topic clustering and analysis across the English-language tweets in our dataset.

For this group of tweets, we were able to identify 363 topics with 45,643 tweets being left in the 'uncategorised' topic. The 'uncategorised' topic often contains documents that exhibit diverse or heterogeneous content, making it challenging to assign them to a specific topic. The 'uncategorised' topic is essentially a catch-all category that captures documents that do not have a strong affinity to any predefined topic or do not meet the threshold for assignment to a specific theme.

The BERTopic model was configured with the following settings for the research paper: an n-gram range of 2 to 3; a minimum topic size of 75 documents, ensuring that topics with fewer documents are not created; and the selection of the top 5 words as representative keywords for each topic, providing concise and interpretable summaries of the main themes. We were able to cluster the original 363 topics into 34 larger topics with hierarchical topic modeling, described in section 2, by merging the total of 324 sublists of topics together. For the hierarchical clustering, we set up the cosine distance as higher than 0.7.

The twenty larger topics identified and visualised in Figure 11 (next page) encompass a range of emotionally charged experiences, such as accounts of personal encounters, reflections on the fate of Ukrainian children, and discussions of military developments. Additionally, topics emerged regarding the food crisis, the situation surrounding the nuclear power plant, military aid, regional fire alerts, and NATO membership.

- @antonioguterres @HoxhaFer @BWoodward_UN @USAmbKyiv @NDeRiviere Russian invaders have created a global food crisis. Russian leaders don't care if there is some food on the tables of people in BD or UA. Don't believe Russia! More sanctions on Russia! #RussiaIsATerroristState https://t.co/mUnkpDtRF5

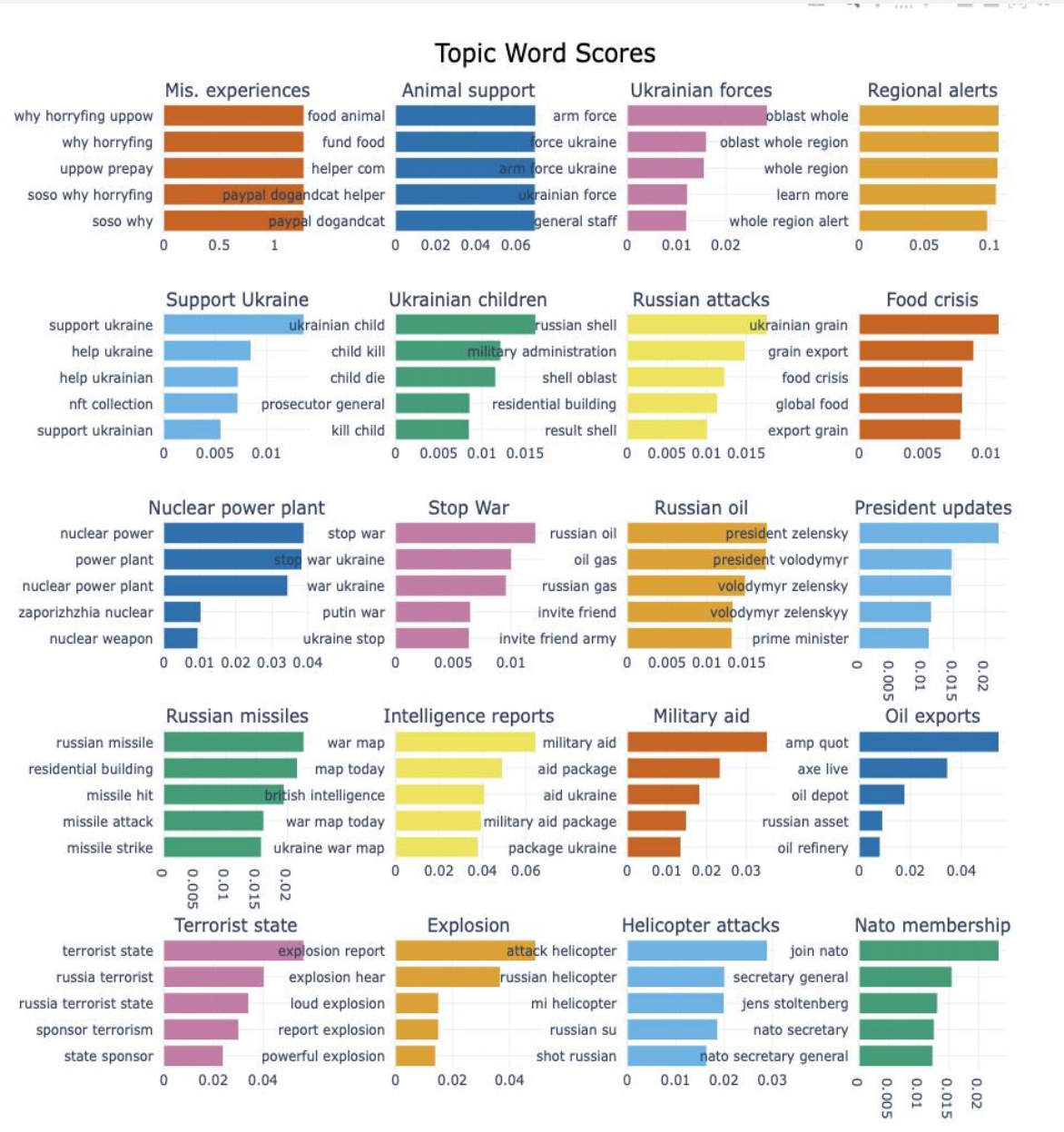

Figure 11: Distribution of the topics within tweets in English

The tweet employs a rhetorical strategy to draw attention to the alleged impact of Russian actions on global food security. By tagging international figures and organisations, the message aims to rally support for its claims. The emotive language, including the term ‘Russian invaders,' portrays Russia negatively. The use of emoticons of flags (🇧🇩 and 🇺🇦) and a table symbolises the consequences of Russia’s actions on people’s lives and suggests a disregard for their well-being. The hashtag #RussiaIsATerroristState further amplifies the accusatory tone, while the call for more sanctions on Russia encourages international responses against Russia.

- This is the first step to ensure security of #ZNPP and its demilitarisation from Russian armed forces.
  Video: @RafaelMGrossi

#StandUpForUkraine #StandWithUkraine #GenocideOfUkrainians #ArmUkraineNow #zaporizhzhia #nuclearpowerplant #DATTALION

The use of hashtags like #StandUpForUkraine, #StandWithUkraine, #GenocideOfUkrainians, #ArmUkraineNow, and specific location-related hashtags (#zaporizhzhia, #nuclearpowerplant) indicates the tweet's alignment with the Ukrainian perspective and conveys messages of solidarity, urgency, and concern for Ukraine's security. Overall, the tweet aims to raise awareness about the situation, emphasise the significance of the discussed action, and engage users who support Ukraine's stance.

A noteworthy observation is that the topics identified in English-language tweets tend to focus on the global impact of the war and exhibit more extrinsic orientations than do the topics found in Ukrainian and Russian tweets.

Discourse analysis reveals that a significant portion of the English corpus is dedicated to the examination of military operations on Ukrainian territory. Within this domain, a discernible trend emerges, highlighting the differentiation between attacks targeting civilians and those taking place on the battlefield. The topic of Russian attacks, for instance, highlights instances of residential buildings being targeted and entire regions being subjected to shelling. Notably, this topic exhibits a high level of emotional intensity, as evidenced by emotionally charged tweets.

- Again and again Ukrainian civilian people die from Russian shelling, and these are not military objects. 🤬
  @amnesty why are you silent?
  @UN @UNICEF @BBC @CNN @nytimes @guardian @POTUS @OlafScholz @EmmanuelMacron #RussiaIsATerroristState #Mykolayiv https://t.co/0vc8BQagQp
- Ukrainian civilian people die Russian shelling, military objects. 🤬
  silent

The tweets above strongly condemn the loss of civilian lives in Ukraine due to Russian shelling, expressing frustration and anger through emoticons and profanity. By addressing global organisations and media outlets, the commenter seeks to draw attention to the perceived inaction of entities like @amnesty, @UN, and others. The use of hashtags like #RussiaIsATerroristState signals an appeal for international condemnation of Russia's actions. This comment demonstrates emotional outrage and employs a persuasive strategy to evoke empathy and action from a broader audience.

In contrast, topics such as 'Helicopter attacks', 'Explosion', 'Intelligence reports', and 'Russian missiles' employ more concise and factual language, focusing on information derived from the battlefield and intelligence sources.

- Russian Helicopter Shot Ukrainians Used Stinger Destroy Ka- RUSSIA-UKRAINE WAR
- Russian-Belarusian grouping troops hardly corresponds declared number, - British intelligence
- This comes at the time when Russian missiles and drones are targeting Kyiv and other cities https://t.co/kPliQOqiLg

These tweets seem to present a factual news update regarding the Russia-Ukraine war. Unlike the previous emotionally charged tweets, these messages appear fairly objective in tone. The pragmatic intent here seems to be informative, aiming to provide updates on the war's developments rather than invoking emotional reactions.

English-language tweets about the Russia-Ukraine conflict primarily emphasise its global impact and adopt more extrinsic viewpoints than Ukrainian and Russian tweets. Emotionally charged topics accuse Russia of causing a global food crisis and denounce Russian attacks on Ukrainian civilians, using emotive language, hashtags, and international mentions to garner support. In contrast, factual and objective topics deliver updates on military operations, such as helicopter attacks and missile strikes, with a focus on informative reporting. This analysis underscores the diverse discourse surrounding the invasion in English tweets, blending emotional appeals with factual updates and shedding light on varied communication strategies used by stakeholders.

3.6 Dynamic topic modeling of the English corpus

The visual representations provided in figures 12 and 13 depicts the temporal dynamics and prevalence of the top 10 English corpus topics. The topics within the English corpus are relatively broad compared to the Russian and Ukrainian ones. Therefore, to conduct a more profound analysis of the dynamics we have focused on the largest 10 topics instead of visualising 15 as in the case of a Russian corpus or 20 as in the case of the Ukrainian one. Notably, the 'Regional alerts' topic exhibited a substantial surge during the initial phase of the observation period, coinciding with heightened activity during the invasion. However, its occurrence gradually declined during comparatively calmer periods. This trend aligns with the notion that regional alerts garnered increased attention when the conflict was more intense but waned as stability improved.

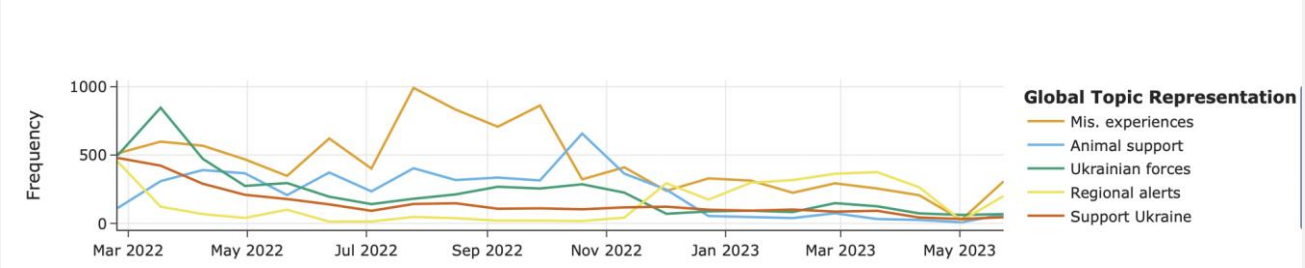

Figure 12: Development of topics in the English corpus over time. Part 1

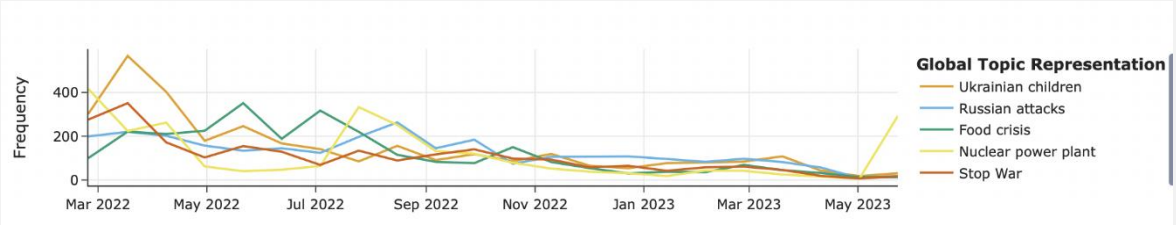

Figure 13: Development of topics in the English corpus over time. Part 2

Similarly, the ‘Support Ukraine’ and ‘Stop war’ topics demonstrated a downward trajectory over time. This decline could be attributed to a combination of factors, including diminishing media interest and a gradual waning of social media campaigns advocating for support for Ukraine. The decrease in attention and engagement surrounding these topics suggests a potential shift in public discourse and priorities, which commonly occurs as conflicts evolve and media narratives evolve accordingly.

In contrast to the aforementioned topics, the topics ‘Food crisis’, ‘Animal support’, and ‘Nuclear power plant’ displayed noticeable fluctuations in their trajectory, which suggests that they are driven by the ongoing issues causing sudden spikes in the spread of these topics. Nevertheless, though sometimes talked less about, they are constantly present in the public discourse, indicating the sustained relevance and significance of these themes. The discourse surrounding the ‘Food crisis’ topic likely reflects concerns about access to adequate food supplies in the context of the Grain Agreement between Ukraine and Russia, while the ‘Animal support’ topic may be indicative of discussions related to the well-being and protection of animals in the war zone. Furthermore, the ‘Nuclear power plant’ topic suggests ongoing attention to the potential risks and consequences associated with nuclear facilities in the temporarily occupied region. The continued relevance of this topic underscores the enduring societal and environmental concerns surrounding the situation.

Conclusions

Through the application of BERTopic for topic modeling, our analysis centred on war-related tweets originating from Ukraine. Across Russian, English, and Ukrainian tweets, common themes emerged, yet distinct nuances within each dataset underscored their unique pragmatic objectives. English tweets exhibited a global perspective, reflecting an intent to garner international attention. These tweets encompassed topics extending beyond immediate conflict issues, to areas like food crises, animal welfare, nuclear power plants, humanitarian concerns, and geopolitical matters such as NATO membership and military aid.

Conversely, Ukrainian tweets demonstrated a more internally focused discourse, encompassing news updates, military developments, and issues pertaining to national identity and culture. Furthermore, they sometimes conveyed disagreement with Ukrainian government policies and a call for change.

Russian tweets, in contrast, targeted the Russian population with a different agenda. These tweets aimed to dissuade the Russian population from participation in the conflict and to disseminate information. Our findings support the hypothesis that language choice on Twitter correlates with distinct communicative goals, leading to diverse topic compositions in the datasets. While the possibility of bot activity promoting a pro-Russian stance was considered, all topics within our corpora displayed an unmistakable anti-war focus.

We have identified that topic dynamics within all three corpora are reflective of the media timeline and main information occasions. Even though the topics of the tweets in English, Russian and Ukrainian differ greatly, we can trace the spikes in tweets on specific topics to the dates a specific event occurred and was discussed in the news and in social media. Nevertheless, the topic dynamics for the Russian corpus appears to be rather flat with almost

identical development of the topics present in the corpus, which we attribute to the initially small size of the topics.

It is crucial to acknowledge the imbalance in the sizes of the corpora, which influenced the efficacy of the topic modeling process. To mitigate this challenge, we employed language-specific techniques, including tailored encodings, adjusted minimum topic sizes, and lemmatisation, to optimise topic identification, particularly in smaller datasets.

Our analysis indicated that the corpus size discrepancy did not result from external influences, suggesting that there was no deliberate shift from using Russian to Ukrainian in online communication. English remained the most consistently used language throughout our research timeline. This underscores the complex interplay between language choice, communicative goals, and the discourse surrounding war-related topics on Twitter.

References

ALASH, Hayder M. – AL-SULTANY, Ghaidaa A. 2020. "Improve topic modeling algorithms based on Twitter hashtags". *Journal of Physics: Conference Series* 1660, 1, 012100.

ANONYMOUS. Retrieved 06 September 2023. "How Russia attempts to widen its arsenal of pro-war propaganda". *Medium*. <https://medium.com/dfrlab/how-russia-attempts-to-widen-its-arsenal-of-pro-war-propaganda-c0a6181b4efb>.

BADAWY, Adam – FERRARA, Emilio – LERMAN, Kristina. 2019. "Who falls for online political manipulation?". In Liu, Ling – White, Ryen (eds), *Companion Proceedings of the 2019 World Wide Web Conference*, 162–168. New York: Association for Computing Machinery.

BARRINGTON, Lowell. 2022. "A new look at region, language, ethnicity and civic national identity in Ukraine". *Europe-Asia Studies* 74(3), 360–381.

BLEI, David M. – NG, Andrew Y. – JORDAN, Michael I. 2003. "Latent dirichlet allocation". *Journal of Machine Learning Research* 3, 993–1022.

CHAYINSKA, Maria – KENDE, Anna – JA WOHL, Michael. 2022. "National identity and beliefs about historical linguicide are associated with support for exclusive language policies among the Ukrainian linguistic majority". *Group Processes & Intergroup Relations* 25(4), 924–940.

CHEN, Emily – FERRARA, Emilio. 2023. "Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia". In Lin, Yu-Ru – Mejova, Yelena – Cha, Meeyoung (eds.), *Proceedings of the International AAAI Conference on Web and Social Media 17*, 1006–1013. Washington: AAAI Press.

COHEN, Elliot A. 2022. "Die Ukraine gewinnt den Krieg…". *Zeitschrift Osteuropa* 72(1–3), 179–183.

ERAS, Laura. 2022. "War, identity politics, and attitudes toward a linguistic minority: Prejudice against Russian-speaking Ukrainians in Ukraine between 1995 and 2018". *Nationalities Papers* 51(1), 1–22.

FOONG, Wai Ng. Retrieved 30 August 2023. "Introduction to google's compact language detector V3 in Python". *Medium*. <https://towardsdatascience.com/introduction-to-googles-compact-language-detector-v3-in-python-b6887101ae47>.

FUNG, Yi R. – JI, Heng. 2022. "A weibo dataset for the 2022 Russo-Ukrainian crisis." *arXiv preprint*, arXiv:2203.05967. <https://arxiv.org/pdf/2203.05967>.

GONZÁLEZ-PIZARRO, Felippe – ALAVI, Soheil. Not dated. "MultiModalTopicExplorer: A visual text analytics system for exploring a collection of multi-modal online conversations". <https://www.cs.ubc.ca/~tmm/courses/547-21/projects/felipe-soheil/report.pdf> (retrieved 22 July 2024).

GROOTENDORST, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". *arXiv preprint*, arXiv:2203.05794. <https://arxiv.org/abs/2203.05794>.

——. Retrieved 30 August 2023. "Dynamic topic modeling, BERTopic". <https://maartengr.github.io/BERTopic/getting_started/topicsovertime/topicsovertime.html>.

HENTSCHEL, Gerd – PALINSKA, Olesya. 2022. "The linguistic situation on the Ukrainian Black Sea coast–Ukrainian, Russian and Suržyk as 'native language', 'primary code', frequently used codes and codes of linguistic socialization during childhood". *Russian Linguistics* 46(3), 259–290.

KARPINA, Olena – CHEN, Justin. 2022. "Topic modeling and emotion analysis of the tweets of British and American politicians on the topic of war in Ukraine". *East European Journal of Psycholinguistics* 9(2), 41–66.

KOSTIUČENKO, Anastasija. 2023. "Surzhyk in Ukraine: Between language ideology and usage". *Ukrainian Analytical Digest* 1, 15–17.

LYUDOVYK, Tetyana – PYLYPENKO, Valeriy. 2014. "Code-switching speech recognition for closely related languages". In Karpov, A. A. (eds.), *SLTU-2014 4th International Workshop on Spoken Language Technologies for Under-resourced Languages*, 188–193. St. Petersburg.

MATHAYOMCHAN, Boonyanit – TAECHARUNGROJ, Viriya – WATTANACHAROENSIL, Walanchalee. 2023. "Evolution of COVID-19 tweets about Southeast Asian Countries: Topic modeling and sentiment analyses". *Place Branding and Public Diplomacy* 19(3), 317–334.

PAVLENKO, Aneta – BLACKLEDGE, Adrian. 2001. "Negotiation of identities in multilingual contexts". *International Journal of Bilingualism* 5(3), 243–257.

PETRIV, Olha. 2022. "УКРАЇНСЬКА МОВА У ВОЄННИЙ ЧАС". *Наукові записки Національного університету «Острозька академія»: Серія «Філологія»* 14(82), 13–16.

QIN, Zhikang – RONCHIERI, Elisabetta. 2022. "Exploring pandemics events on Twitter by using sentiment analysis and topic modeling". *Applied Sciences* 12(23), 1–21.

RACEK, Daniel – DAVIDSON, Brittany I. – THURNER, Paul W. – ZHU, Xiao Xiang – KAUERMANN, Göran. 2023. "The politics of language choice: How the Russian-Ukrainian war influences Ukrainians' language use on Twitter". *arXiv preprint*, arXiv:2305.02770. <https://arxiv.org/abs/2305.02770>.

SCIPY. Retrieved 06 September 2023. "scipy.cluster.hierarchy.linkage". *SciPy* v1.7.3 Documentation. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy. linkage.html>.

SHULTZ, Benjamin. 2023. "In the spotlight: The Russian government's use of official Twitter accounts to influence discussions about its war in Ukraine". In Cuccovillo, Luca – Ionescu, Bagdan – Kordopatis-Zilos, Giorgos – Papadopoulos, Symeon – Popescu, Adrina

(eds.), *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, 45–51. New York: Association for Computing Machinery.

STAALESEN, A. Retrieved 30 August 2023. "The Nazi-inspired symbol used by Russia in war against Ukraine finds way to downtown Murmansk". *The Independent Barents Observer*. <https://thebarentsobserver.com/en/security/2022/03/nazi-inspired-symbol-used-russia-war-against-ukraine-finds-way-downtown-murmansk>.

TAO, Wei – PENG, Yingtong. 2023. "Differentiation and unity: A Cross-platform comparison analysis of online posts 'Semantics of the Russian–Ukrainian war based on Weibo and Twitter". *Communication and the Public* 8(2), 105–124.

WICKE, Philipp – BOLOGNESI, Marianna M. 2020. "Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter". *PloS one* 15(9), e0240010.

YAGODKINA, Maryana V. 2020. "Технологии пропаганды в современной коммуникативной среде". In Eremeev, S. G. (ed.), *XXIV Царскосельские чтения. 75-летие Победы в Великой Отечественной войне*, 295–298. LGU.