Disinformation Crossing Spaces and Language Borders: A Contrastive Analysis of English and Lithuanian

Jūratė Ruzaitė (Vytautas Magnus University)

Abstract
This paper examines the language of disinformation on the topic of the COVID-19 pandemic in English and Lithuanian using the methods of contrastive corpus linguistics. The study not only reports research results but also addresses some methodological issues encountered in contrastive analysis of disinformation, a main one being the absence or limited amount of original content in Lithuanian disinformation texts. Since most of the Lithuanian content is translated or adapted from other sources, an important question is how likely it is that some distinct language-specific features will emerge in disinformation published in a lesser-used language. The content modifications in the Lithuanian texts range from very close translations of the source texts to highly abridged summaries of the original. A general trend is that almost all the texts are shorter in Lithuanian. Regarding the analysis of linguistic properties, the type-token ration (TTR) is very low in English texts but considerably higher in Lithuanian, which could be a result of typological differences between the two languages. Emphatics are almost equally distributed in both datasets; however, tentative language is more frequent in English. Such trends suggest that the language of disinformation tends to be simple, but Lithuanian false news aims at sensationalism by retaining the same frequency of emphatic wording but reducing the tentative tone.

Keywords: disinformation; contrastive analysis; corpus linguistics; English/Lithuanian

1. Introduction

This paper aims to examine the language of disinformation in English (as a global language) and Lithuanian (as a lesser-used language) by applying the methods of contrastive corpus linguistics. Contrastive analysis of disinformation features is not as straightforward as it may seem. The present study thus not only reports the results of the present research but also addresses some methodological issues encountered in contrastive

Ruzaitė, Jūratė. 2024. 'Disinformation Crossing Spaces and Language Borders: A Constrastive Analysis of English and Lithuanian.' *Nordic Journal of English Studies* 23(2): 268–298. https://doi.org/10.35360/njes.v23i2.39193.

analysis of disinformation. This research is part of a broader project focusing on Lithuanian propaganda and disinformation. In its initial stages, it became evident that a substantial portion of Lithuanian texts draws upon English-language sources, underscoring the importance of evaluating the impact of these sources on Lithuanian content.

To narrow down the scope of analysis, the present study focuses on one topic of disinformation—the COVID-19 pandemic, which has caused widespread uncertainty and anxiety, and has affected multiple facets of human life on a global scale. This has resulted in a significant proliferation of false information being widely shared in numerous languages on diverse social media platforms. Consequently, the pandemic offers researchers a chance to gain insights into the qualities of disinformation in various languages, which can be invaluable for advancing research and strategies for identifying and mitigating disinformation. During the COVID-19 pandemic, the United Nations (UN) made the oft-quoted statement expressing their concern that: 'In times of the Covid-19 health crisis, the spread of the "infodemic" can be as dangerous to human health and security as the pandemic itself'[1]. The term 'infodemic' was frequently used by various international organizations, including the World Health Organization (WHO) and the UN, to describe the excessive amount of information, both accurate and inaccurate, that was circulating during the pandemic.

Disinformation is generated and disseminated mainly in English, and research on disinformation texts primarily focuses on English data (e.g., Davoudi et al. 2022; Shu, Wang and Liu 2019; Grieve and Woodfield 2023). However, even though disinformation circulating in languages other than English is less abundant, it amplifies its influence and may reflect some regional specificities. Considering that culture, language, political views, and religion may influence the way that news (false or factual) is generated, perceived, and disseminated, it is important to examine disinformation in languages other than English to better perceive how it is produced, how it functions, and what socio-cultural aspects it reflects.

The present study is motivated by the lack of cross-linguistic research on disinformation. Generally, research on disinformation is ample,

[1] The specific source and date of this quote may vary, but it was a common theme in various UN and WHO communications throughout the pandemic in 2020 and 2021.

especially in such areas as media studies and computational linguistics. The main body of such research focuses on the differences between fake and factual news. However, systematic contrastive studies on linguistic features of disinformation are still highly limited (see Damstra et al. 2021).

The present analysis concerns the linguistic properties of Lithuanian and English disinformation by addressing the following research questions (RQs):

1. How is English content modified in Lithuanian disinformation texts?
2. What is the lexical diversity (type-token ratio, TTR) of the texts?
3. How are emphatics and tentative language used in Lithuanian and English data?

These RQs stem from some earlier research indicating that low lexical diversity, emphatics, and tentative language are strongly associated with disinformation (e.g., Bezerra 2021; see further section 2). Prior research suggests that disinformation is marked by simplicity, and thus TTR in disinformation texts is lower than in factual, or legitimate, news (Kumar and Vardhan 2021).

A more detailed review of research on disinformation is presented in section 2, which covers different terms associated with disinformation, offers an overview of previous studies exploring disinformation in languages other than English and employing contrastive analysis, and briefly touches upon research on tentative language and emphatics. Data and methods used in the empirical analysis are presented in section 3, and section 4 discusses the findings of the analysis. Section 5 generalises the main findings, points out some limitations of the present study and provides some suggestions for further research.

2. Background

2.1 Terminological issues

In contemporary debate in the field of journalism, media studies and political communication, there is a diversity of terms used to refer to false information including ‘misinformation’ and ‘disinformation’ (e.g., Lazer et al. 2018), ‘fake news’ (e.g., Gelfert 2018; Grieve and Woodfield 2023), ‘information pollution’ (Meel and Vishwakarma 2020), and ‘information disorder’ (Wardle and Derakshan 2017).

The concept of 'fake news' has perhaps generated greater discussion among journalists and the academic community than any other term in this set of concepts. Originally, the term 'fake news' was used to describe deliberately fabricated or false information presented as news. It is still rather widely used in this sense in current research, especially in research on automated detection of disinformation, which is commonly referred to as '(automated) fake news detection'. Over time, however, it has become a charged term, whose appropriateness is now often questioned for a variety of reasons.

The term is often disapproved of for its fuzzy meaning and especially its connotations. For instance, 'fake news' is criticised as a broad and often vague term that can encompass a wide range of deceptive information, from satirical content to misinformation and disinformation (ERGA 2020). It is also considered inappropriate as it has become a politically charged term, with different groups accusing their opponents of spreading it. This trend has led to disputes over what should be labelled as 'fake news' and has contributed to a decline in trust in media and discreditation of journalism (e.g., Freelon and Wells 2020).

In addition, it is argued that the term 'fake news' does not capture our new reality: most of the content which is part of 'information disorder' is not fake, but instead, is used out of context and thus becomes misleading; or it can mix falsehoods with some truth to look trustworthy (Wardle 2019: 6). Neither can such content be labelled as 'news', as very often it contains mere rumours, manipulated videos, or old photos re-shared as new (Wardle 2019: 6).

Finally, according to Habgood-Coot (2018), the term 'fake news' is not only redundant, given the abundance of vocabulary alternatives, but also legitimizes anti-democratic propaganda.

In response to these issues, some experts and organizations advocate for using more specific terms, such as 'misinformation', 'disinformation', or 'false information', to describe different types of deceptive content more precisely. 'Disinformation' is a term favoured not only by scholars but also by experts from the European Commission (de Cock Buning 2018), and thus will be used as the main term in the present paper, while the term 'fake news' will be avoided and will be employed only with reference to automated fake news detection.

To avoid loaded terms, some scholars have also introduced alternative terms, such as 'information pollution' (Meel and Vishwakarma 2020) or

'false information' (Kapantai et al. 2020). The term 'false information' will also be used in this paper as a general term referring to different types of deception. However, 'information pollution' is too broad and refers to more types of contaminating information than necessary in this research, e.g., it refers not only to dis-, mis-, and mal-information, but also incomplete, inconsistent or irrelevant information (Orman 1984).

For the purposes of this study, I adopt the categorisation introduced by Wardle and Derakshan (2017; see also Wardle 2019), who suggest 'information disorder' as a hypernym to cover different types of inaccurate and harmful information:

- **Dis-information**. Information that is false and deliberately created to harm a person, social group, organization or country.
- **Mis-information**. Information that is false, but not created with the intention of causing harm.
- **Mal-information**. Information that is based on reality, used to inflict harm on a person, organization or country. (Wardle and Derakshan 2017: 20)

Despite their distinctive features, these concepts intersect to some extent, as represented by Wardle and Derakshan (2017: 20); see Figure 1.

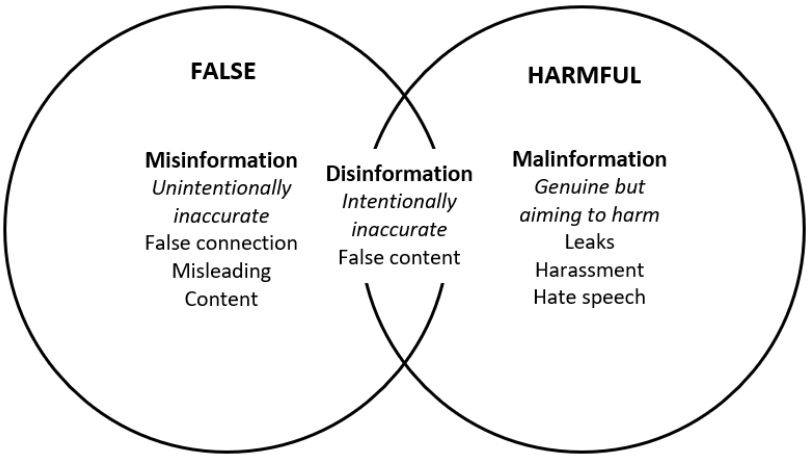

Figure 1: Information disorder: mis-, dis- and mal-information with regard to falseness and harm (adapted from Wardle and Derakshan 2017: 20)

Misinformation and disinformation, thus, both pertain to categories of inaccurate or false information, with the distinction being that disinformation is intentionally deceptive. Disinformation is distinct from mal- and misinformation in that it refers to content that is both harmful and false. Following Damstra et al., intentional deception is 'created and distributed with the aim of influencing political attitudes, behavior, or processes' (2021: 1948), and this is where the harm of disinformation manifests itself. Regarding the criterion of intentionality, it is important to note that the false information presented in the two datasets used in this research qualifies as disinformation since it has been sourced from widely recognised propaganda and disinformation websites.

It should be noted that this paper does not cover the category of satire, i.e., texts where humour is employed to create satirical news in an exaggerated way, clearly indicating that this information is not legitimate (del Pilar et al. 2017). When such news is shared on the internet, it appears decontextualised and, being similar in its form to legitimate news, it can deceive the reader just like disinformation. However, such instances did not occur in the current data.

2.2 Cross-linguistic research on disinformation

Research on disinformation being a fast-growing research area covers a diversity of approaches, methodologies, and research foci, which makes it impossible to cover all the trends in the present overview. Here, the main focus will be on research on languages other than English and especially on cross-linguistic studies of disinformation.

Though research on languages other than English is still in its initial stages, a number of attempts have been made mainly with the goal to develop models for automated detection of disinformation. Here just some examples are reported to show the range of languages covered so far. For a more detailed overview of the technical aspects of the models developed for non-English languages, see Zhou (2023).

In broad terms, studies in this area can be classified into those focussing on (a) individual vs. multiple languages, and (b) those aiming to develop language-independent vs. language-dependent models. A few studies have explored the detection of disinformation in multiple languages, e.g., English, Portuguese, and Spanish in Abonizio et al. (2020) and English, Portuguese, and Bulgarian (representing three language groups—Germanic, Latin, and Slavic) in Faustini and Covões (2020). In

research aiming to develop language-independent detection, the models are based on the analysis of several languages, but in these languages, the authors assess textual features that are not tied to a specific language (e.g., Abonizio et al. 2020).

Some aspects of such research could be relevant to contrastive analysis of disinformation texts; however, research papers in this area do not focus on linguistic analysis but technical aspects of model development. For instance, Abonizio et al. (2020) rely on complexity, stylometric, and psychological text features, which could be related to some linguistic categories examined in the present study, but they are not made explicit enough to constitute substantial cross-study comparisons.

Some studies develop language-dependent models for individual or multiple languages, e.g., Chinese (Zhu et al. 2018; Du et al. 2021), Urdu (Farooq et al. 2023), Indic languages (Kar et al. 2021), Spanish satirical news in Spain and Mexico (del Pilar et al. 2017), and French and English (Guibon et al. 2023).

In general, research focusing on automated fake news detection is undoubtedly important, but it is carried out by computer scientists and is thus different (in terms of methods and theoretical approaches) from contrastive (linguistic) analysis. In the trend of fake news detection, the models created are analysed and assessed regarding their accuracy in determining disinformation automatically, but a consistent and systematic analysis of linguistic text properties and typological differences/ similarities is lacking.

Some large-scale contrastive studies of disinformation have been carried out in the field of media and communication studies. For instance, Humprecht (2019) compares disinformation across English-speaking (US and UK) and German-speaking countries (Austria and Germany), but in this comparison the focus is on the content and not the linguistic properties of these texts.

One of the few, if not the only publication presenting a contrastive linguistic analysis of disinformation and mainstream media news texts is Sousa-Silva's (2022) research on English and Portuguese. One finding of his research that is relevant to the present study is that in both languages disinformation texts (in contrast to mainstream news reports) often use evaluative adverbs to emphasise the cruelty of the actions described in their texts.

The only attempt to compare more than two languages (English, Norwegian and Russian) from the linguistic perspective is the on-going project *Fakespeak—the language of fake news*[2] (for a report on the methodological considerations based on this project, see Põldvere, Kibisova and Alvestad 2023).

The research by Humprecht (2019) and Siwakoti et al. (2021) usefully indicates that there are important differences in the types of disinformation narratives across different parts of the world, but it remains unclear if there are important cross-linguistic differences in linguistic properties of fake news. Research attempting to compare Lithuanian to English or any other languages is non-existent.

2.3 Disinformation markers: Tentative language and emphatics

In this study, emphatics and tentative language are focused on as they serve as useful indicators of bias (Recasens, Danescu-Niculescu-Mizil and Jurafsky 2013): tentative language (e.g., such adverbs as *just* and *only*) is a category of epistemological bias, and emphatics (e.g., *very* and *absolutely*) are a category of framing bias (Recasens, Danescu-Niculescu-Mizil and Jurafsky 2013). Additionally, emphatics are associated with sensationalism and are used as part of propaganda techniques aiming to persuade users on an emotional rather than cognitive level (Damstra et al. 2021; Staender et al. 2021). Emphatics and softeners are also strongly linked to the propaganda strategy of exaggeration and minimisation (e.g., Da San Martino et al. 2019).

The decision to focus specifically on tentative language and emphatics is further motivated by Grieve and Woodfield's (2023) research. Their study examined twenty-eight grammatical features in both real and fake news texts, aiming to identify markers that distinguish between the two categories. Among these features, fifteen, including emphatics and downtoners, were found to be considerably more prevalent in fake news compared to factual news. According to Grieve and Woodfield, emphatics, alongside many other adverbs and adjectives, contribute 'inconsequential information' and consequently reduce 'the information density of newspaper writing' (2023: 54). As a result, they are used primarily for expressive and evaluative rather than informative purposes.

[2] More information is available at https://www.hf.uio.no/ilos/english/research/projects/fakespeak/.

Linguistic research on deceptive language further indicates that, in general, this type of language tends to include a higher occurrence of tentative, angry, and emotionally charged words compared to truthful language, as observed in Asubiaro and Rubin's (2018) study (see also Faris et al. 2017; Benkler, Faris and Roberts 2018). Therefore, it is hypothesised in this research that tentative language and emphatics will be of similar frequency in both English and Lithuanian data.

3. Data and methods

The Lithuanian data collected for this study includes disinformation texts about COVID-19 obtained from one Lithuanian website well-known for publishing propaganda and disinformation—*minfo.lt*. The website *minfo.lt*, which stopped its activity in August 2023, became known as an unreliable news website after it was bought in 2018 by Marius Gabrilavičius, a controversial figure who is also involved in alternative therapeutic practices of treating different addictions.

Being a first attempt to examine Lithuanian disinformation from the perspective of contrastive linguistics, the present research is a small-scale pilot study, based on a limited amount of data. The total amount of data used for this study initially included 40 disinformation texts (18,111 tokens) from the Lithuanian website; these were all the disinformation texts available on the topic of the COVID-19 pandemic on the website under investigation. In this study, the texts were identified as disinformation through fact-checking conducted by the author of the research herself, and some of the fake claims had already been debunked by mainstream media.

When processing the Lithuanian corpus, it appeared that 32 of the texts (80%) were based on English-language sources, and only 8 out of 40 texts (20%) were original content not republished from other sources. Unexpectedly, it appeared that the Lithuanian texts were mainly adapted reposts from two major propaganda portals: the far-right international newspaper *The Epoch Times*[3] and the conspiracy and pseudoscience website *Technocracy News*[4]. The media bias resource *Media Bias/Fact Check*[5] lists both websites as questionable sources of low credibility,

[3] *The Epoch Times*, https://www.theepochtimes.com/.
[4] *Technocracy News*, https://www.technocracy.news/.
[5] *Media Bias/Fact Check*, https://mediabiasfactcheck.com/.

Technocracy News being classified as a source of very low factual reporting, and *The Epoch Times* categorised as a source reporting mixed—false and factual—information.

To examine how the English texts were exploited for Lithuanian disinformation and to perform a comparative analysis of linguistic features in the two languages, the original English texts were collected for further comparative analysis (32 texts totalling 52,264 tokens). The sizes of both datasets are represented in Table 1. The eight non-translated texts were excluded from further investigation due to their substantially different nature compared to the translated ones. While they could potentially offer value for comparison, their limited quantity within the current dataset makes them unsuitable for consistent analysis.

Table 1: Size of the corpus used for contrastive analysis

	Number of text units	**Number of tokens**
English dataset	32	52,264
Lithuanian dataset	32	12,810

An interesting and unexpected finding that emerged already in the initial stages of data collection is that Lithuanian texts are considerably shorter than the English texts: only 40% of English content is retained in Lithuanian. This can be partly explained by linguistic typology: English is more analytical, while Lithuanian is a synthetic language and therefore uses a smaller number of function words. Synthetic languages tend to use inflections and affixes to convey meaning, which can lead to fewer words and shorter sentences compared to analytical languages, which often rely on word order and function words. However, this inherent linguistic typology cannot explain the shortening of texts by more than a half.

As some prior research shows (e.g., Kasseropoulos and Tjortjis 2021; Kumar and Vardhan 2021), the size of a text is a crucial factor in identifying disinformation. Typically, spreaders of false information tend to opt for concise messages when communicating their content. Their aim in crafting brief messages is to draw in users by using uncomplicated language and minimal content (Kumar and Vardhan 2021: 201). It can hence be hypothesised that disinformation recycled in a new source and/or another language will be shorter than the original. However, this still needs to be tested on a larger and more varied dataset, as the trend observed in this study could be a feature of this particular website.

The present analysis focuses on the following categories and text properties of disinformation texts in English and Lithuanian:

- text length,
- lexical complexity,
- emphatic language, and
- tentative language.

These categories emerged as important indicators of disinformation in some earlier research, as will be further explained.

To assess lexical complexity, the study employs the type-token ratio (TTR), a linguistic measure used to analyse the diversity of vocabulary in a text or corpus. It is calculated by comparing the number of unique words (types) to the total number of words (tokens) in the text or corpus. The formula for calculating the type-token ratio is: TTR = (Number of Types) / (Number of Tokens). The choice to consider the TTR measure is motivated by the findings in prior research suggesting that a typical feature of disinformation is simplicity, and thus TTR in disinformation texts is lower than in factual texts (Kumar and Vardhan 2021). However, some studies report opposite results. For instance, Abonizio et al. (2020) observe that the TTR values on all three corpora (English, Portuguese, and Spanish) are lower for legitimate news, followed by disinformation or satirical news.

Emphatics and tentative language (or downtoners in some research, e.g., Grieve and Woodfield 2023) are focused on as they are not only easy-to-identify surface-level features, but also serve as useful indicators of bias: as has already been noted, tentative language is a category of epistemological bias, and emphatics are a category of framing bias (Recasens, Danescu-Niculescu-Mizil and Jurafsky 2013).

Tentative language is studied here by mainly focusing on hedges, defined as words and phrases employed to express uncertainty and lessen the speaker's commitment to the accuracy of a statement, thereby avoiding assertive predictions or statements, e.g., *tikėtina* ('likely'), *galbūt* ('probably'). This category also includes modal verbs, such as *galėti* ('can'), when they reduce the certainty of an utterance.

Emphatics are defined here as linguistic units used to intensify the degree of certainty and intensity of an utterance. The analysis of emphatics mainly covers subjective intensifiers defined as 'adjectives or adverbs that

add (subjective) force to the meaning of a phrase or proposition' (Recasens, Danescu-Niculescu-Mizil and Jurafsky 2013: 1653), e.g., *visiškas* ('complete'/'entire'), *labai* ('very'). In addition, the modal *must* is also added to this category when it gives emphasis to a statement/opinion.

The words used as emphatics and softeners were manually selected from a frequency list generated with *AntConc*. In cases of ambiguity, the use of the word was analysed in context by examining concordance lines to determine whether it has an emphatic or softening effect. For instance, *about* can be used as an approximator before a noun phrase (and thus is an instance of tentative language) or as a preposition (and thus is irrelevant in this analysis). All the latter uses were excluded from further analysis.

Since the two corpora are of different sizes, raw frequencies of the categories under study have been normalized to 10,000 words to make the quantitative results comparable. The statistical significance of the compared frequencies is assessed by applying the log-likelihood test (LL) with the critical value of 3.84 or higher at the level of $p < 0.05$.

4. Results

4.1 Text length

The analysis in this section addresses RQ1: How is English content modified in Lithuanian disinformation texts? Content modifications in the Lithuanian texts are examined by considering the text length in the two datasets.

The results have revealed that Lithuanian disinformation texts range from very close translations to highly abridged summaries of the source texts. This is reflected in the text length of Lithuanian texts, which, very evidently, are shorter than the source texts. Figure 2 gives an overview of how much of the original text is retained in Lithuanian texts (see also Figure A in the Appendix for a more detailed representation of the text length in the two datasets).

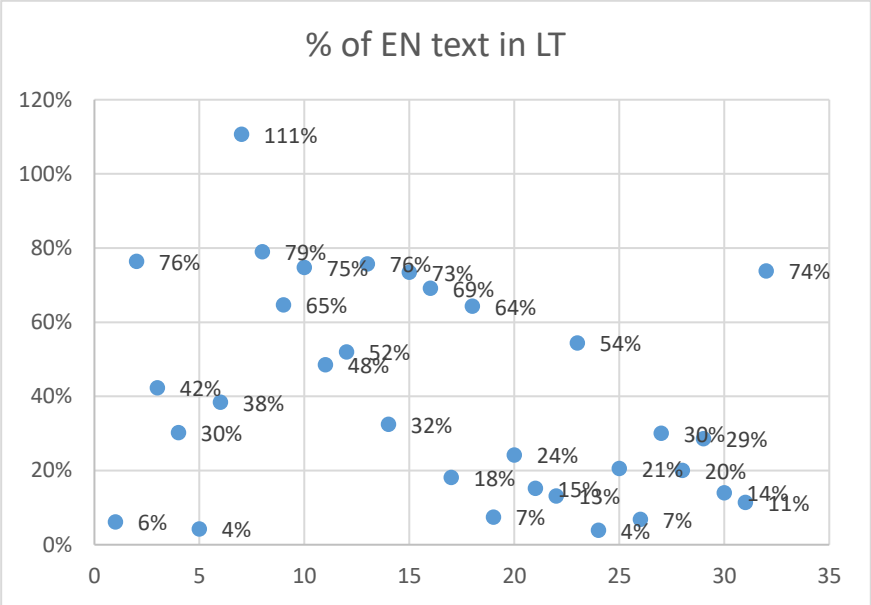

Figure 2: Proportions of the English source text in Lithuanian texts (the percentages indicate how much text is retained in the Lithuanian text)

As the findings indicate, in nearly all cases, Lithuanian texts are shorter, but to varying degrees, with lengths ranging from 79% down to as little as 4% of the source text. The text length exhibits significant variation, with nine texts falling within the 80-60% range, four texts within the 40-60% range, seven texts within the 20-40% range, and ten texts within the 0-20% range. Interestingly, as demonstrated in Figure 2, one text is longer than the source text (111%).

Texts retaining over half of the source text can be assumed to be relatively faithful translations, whereas those in the lower range can be expected to be highly condensed summaries. The latter category of texts slightly predominates. Interestingly, none of the texts includes any additions that would elaborate on a Lithuanian perspective.

As explained in section 3, shorter Lithuanian texts are preconditioned by the inherent typological properties of Lithuanian, which uses fewer function words than English. However, some texts are abridged by omitting certain portions of the source text. Typically, these are (1) lists of references, (2) section titles, and (3) some passages containing supposedly

unnecessary or superfluous information. It is interesting, though perhaps not completely surprising, that lists of references are considered superfluous and are thus excluded in Lithuanian disinformation texts. In cases where some passages were omitted, it was difficult to determine a clear pattern and to suggest a reason why they were considered unnecessary.

4.2 Lexical diversity

This section addresses RQ2: What is the lexical diversity (type-token ratio, TTR) of the texts? To answer this question, TTR indices have been calculated following the principles explained in section 3.

As demonstrated in Table 2, TTR is very low in English texts (0.14) but is considerably higher in Lithuanian (0.39), which again could be a result of typological differences between the two languages.

Table 2: Lexical diversity (type-token ratio, TTR)

	EN	**LT**
Total No. of Word Types	7,056	4,826
Total No. of Word Tokens	52,154	12,358
TTR	0.14	0.39

As previously mentioned, Lithuanian is an analytic language, meaning that a single word can have a wider array of grammatical forms, leading to a greater number of types. Therefore, a straightforward interpretation of the current TTR results is impossible (see also Abonizio et al. 2020 for a similar observation regarding English, Portuguese, and Spanish). Considering these differences in typology, the TTR should be interpreted with caution, and the obtained results need to be tested in future research on the basis of lemmatised data to solve the problem where different grammatical forms are counted as different types.

Another concern with TTR is the instability of this index in the case of short texts (for a discussion, see, e.g., Jarvis 2002). Therefore, a considerably larger database could yield more reliable findings, or a different measurement should be applied.

Despite the difference between the two datasets, both TTR indices are still very low. As previous research consistently shows, lower lexical diversity is characteristic of intentionally deceptive communication in different settings (e.g., Newman et al. 2003; Zhou et al. 2004; Fuller, Biros

and Delen 2011). Some studies (e.g., Ahmed, Traore and Saad 2018), though not all, indicate that this holds true also for news media: disinformation texts tend to be lexically less varied, and thus, as Horne and Adali (2017: 763) explain, disinformation texts require a slightly lower education level to read:

> They seem to be filled with less substantial information, which is demonstrated by a high amount of redundancy, more adverbs, fewer nouns, fewer analytic words, and fewer quotes. (Horne and Adali 2017: 763)

However, such findings are not fully consistent. For example, Abonizio et al. (2020) find higher TTR values for disinformation corpora compared to factual news.

4.3. Emphatic vs. tentative language

This part of analysis shifts its focus to some lexical preferences in the two datasets and aims to answer RQ3: How are emphatics and tentative language used in Lithuanian and English data?

As the quantitative analysis shows (see Table 3), emphatics are almost equally distributed in both datasets (92 instances in English and 96 occurrences in Lithuanian). Despite a slightly higher relative frequency in Lithuanian, the LL test indicates that there is no significant difference between the two datasets (LL = 0.02; Log Ratio = -0.02). While the Log Ratio of -0.02 is negative, it is very close to zero, which suggests that the difference in the frequency of emphatics between the two corpora is so minimal that, without further analysis, it might not have practical implications.

Table 3: Frequency of emphatics and tentative language

	Emphatics		**Tentative language**	
	Raw freq.	**f/10,000**	**Raw freq.**	**f/10,000**
EN	478	92	448	86
LT	119	96	67	54
	LL=0.02 *(no significant difference)*		LL=16.04 *(significantly higher use in EN)*	

Tentative language is of a considerably higher frequency in English, with 86 occurrences compared to 54 occurrences in Lithuanian. The LL value indicates significantly higher use of tentative language in English (LL=16.04), and the positive Log Ratio value (Log Ratio = 0.71) further supports the association of tentative language with the English corpus over Lithuanian. Though the Log Ratio value is moderate and not extensive, this finding is still important and unexpected, particularly in light of earlier results indicating both emphatics and downtoners as characteristic features of disinformation (e.g., Grieve and Woodfield 2023).

These results (namely an almost identical frequency of emphatics and a significantly lower frequency of tentative language in Lithuanian) suggest that Lithuanian disinformation texts are perhaps more marked by sensationalism (for an overview of the concept, see, e.g., Grabe, Zhou and Barnett 2001). Sensationalism in media texts is characterised by content produced to evoke strong reactions or the use of exaggerated, attention-grabbing, or emotionally charged language. Characteristic linguistic features of sensationalism include, for example, exaggeration, emotional language, shocking details, and conflict emphasis. On the other hand, this could be influenced by certain cross-linguistic differences. For example, the current findings are consistent with Marín-Arrese's (2015) study on stance comparing English and Spanish, which suggests that English shows a more tentative style in argumentative discourse strategies compared to Spanish.

4.3.1 Tentative language

Regarding the repertoire used to express tentativeness, the analysis has revealed that there is a larger diversity of items used for this purpose in English. Including single occurrences in both datasets, in total there are 35 types of softeners in English, compared to 25 types in Lithuanian.

The list of the most frequent softeners, shown in Table 4, includes all items that occur at least twice per 10,000 words in the data. Items that occur in both languages are highlighted in grey. It should be noted, however, that the issue of translation equivalence is not addressed here, in order to limit the scope of the analysis, and the correspondence between the two lists is relative.

Table 4: Linguistic repertoire of tentative language in English and Lithuanian

Raw freq.	f/10,000	Softener	Raw freq.	f/10,000	Softener
126	24	can	12	10	galėti (‘can’)
48	9	may	10	8	maždaug (‘about’*)
46	9	could	8	7	beveik (‘nearly’)
37	7	most	6	5	gana (‘rather/quite’)
29	6	likely	6	5	palyginti (‘relatively’)
24	5	actually	3	2	galbūt (‘probably’)
16	3	rather	2	2	atrodo (‘seems’)
13	3	quite	2	2	bent (‘at least’)
11	2	almost	2	2	tikėtina (‘likely’)
11	2	basically			
10	2	about*			

*These numbers only include instances where *about* is used as an approximator.

As can be seen in Table 4, many items coincide in the two languages, but they are used with different frequencies. The main item employed in both English and Lithuanian is the modal *can*/*galėti* (‘can’). Interestingly, its frequency in English is more than twice as high as in Lithuanian (24 and 10 occurrences respectively).

The concordance lines of *can* show that this modal is often used to refer to potential risks and dangers of vaccines, e.g., ***can*** *cause tumors*; ***can*** *also be a risk factor*; *there* ***can*** *be some serious longer-term consequences*; and ***can*** *ultimately lead to neurodegeneration*. Such uses are also abundant in Lithuanian data, e.g., *vakcinos nuo kovido ligos* ***gali*** *sukelti neurodegeneracines ligas* (‘covid vaccines can cause neurodegenerative diseases’) and *vėžys—****gali*** *grėsti tiems, kam buvo suleistos vakcinos* (‘cancer can be a risk for those who have been vaccinated since March’).

In scientific texts, it is common to use *can* when actual possibilities are discussed, e.g., *X can cause Y* (as in *Smoking can cause cancer*.). In such uses, the possibility referred to is realistic though may not be an obligatory consequence of X; such structures can be paraphrased as *X sometimes causes Y*. In disinformation texts, such uses are usually part of scare tactics and do not refer to actual risks, or if they do, the argumentation lacks context or is deficient in some other ways. Being non-categorical statements, they also function as shields when sheer speculations are provided.

Though rarely, the modal *can* is also used in impersonal hedging structures with the *dummy-it*, as in example (1).

(1) *It can be argued* that the loss of a sense of smell and/or taste in association with COVID-19 is a sign of a Parkinsonian link …

The entire fragment quoted in example (1) is omitted in the Lithuanian version. In general, such uses with modals are not common in the Lithuanian dataset: no impersonal structures (e.g., *galima manyti/teigti* etc.) that would correspond to the English structures (*It can be argued/stated/claimed* etc.) can be found in the Lithuanian data.

An important subcategory that emerges in Table 4 is that of items used for approximation, e.g., *about*, *almost*, *maždaug* ('about'), and *beveik* ('nearly'). Numerical references are typically associated with objectivity and scientific style (Ruzaitė 2007) and thus may be exploited in disinformation to achieve different manipulative goals, including building trustworthiness and imitating serious genres. It is characteristic of disinformation creators to selectively choose data that supports their narrative, provide fabricated or misleading statistics to support their claims, and use exaggerated or alarming numbers to elicit strong emotional responses from readers. The latter use seems to be especially important in both datasets.

To show how numerical references are used for manipulative purposes, approximations with *maždaug* ('about'), *nearly* and *about* will be further examined in more detail. Examples (2)–(3) illustrate how large and thus alarming numbers are used to amplify readers' negative emotions. In such cases, they are employed as part of appeals to fear, which are very typical of propaganda (cf. Da San Martino et al. 2019):

(2) a. Federal authorities have received *nearly 800* reports of heart inflammation in people who received a COVID-19 vaccine.
b. JAV ligų kontrolės ir prevencijos centras pranešė, kad nuo koronaviruso vakcinų šalyje *maždaug 800* žmonių kilo miokardo infarktai.

(3) a. … but with an unusual spike from March to April with *about 140* cases per month rather than 100.

b. … tačiau nuo kovo iki balandžio su neįprastu šuoliu—*maždaug* 140 atvejų per mėnesį, o ne 100, kaip įprasta.

In these instances, the approximated rounded numbers in English are accurately preserved in the Lithuanian version (the accuracy of the reported numbers, though, remains another issue).

However, sometimes the Lithuanian version provides numbers with an approximator when there is no corresponding numerical expression in the original text, as in example (4).

(4) Tačiau manoma, tai realus skaičius gali būti *maždaug dešimt* kartų didesnis.
'But it is believed that the real number may be *about ten* times higher'

The Lithuanian version of the text where this example was encountered has undergone considerable modifications when compared to the English original, so it is rather difficult to pinpoint which elements in the two texts should be paralleled. Content-wise, in the original text, there are two possible fragments equivalent to the Lithuanian wording: *meaning that actual deaths are much higher* or *totals may be much higher*. In neither of the two is the numeral *ten* used; instead, a more general quantification is employed (***much*** *higher*). Thus, here the Lithuanian choice to use a specific fabricated number is clearly manipulative.

Interestingly, an opposite trend can also be observed in the data: in some instances, numerical references are provided in the source text but are not transferred to the Lithuanian version. For instance, in (5), an alarming number is provided, which could be potentially useful in Lithuanian manipulative content, but the entire sentence with this number is still omitted in the translation.

(5) A Scandinavian study concluded *about 40%* of post-jab deaths among seniors in assisted living homes are directly due to the injection.

The article where example (5) occurs claims that vaccines are particularly dangerous for senior people. The Reuters Fact Check[6] has determined that approved vaccines are not more dangerous than COVID-19, and they are not exceptionally dangerous for the elderly—the numbers provided in articles suggesting the opposite lack context (for more detailed argumentation, see Reuters Fact Check).[7]

Typically, as could be expected, approximators precede large quantities, e.g., *prilygsta bendram* ***maždaug*** *98–140 atvejų pertekliui* ('equates to a total excess of approximately 98–140 cases') and *Iš užfiksuotų sužeidimų* ***maždaug*** *pusė (968 870) yra rimti* ('Of the recorded injuries, approximately half (968,870) are serious'). However, this is characteristic not only of disinformation but also other interaction (e.g., academic communication as demonstrated in Ruzaitė (2007)).

As the discussion so far has shown, predominantly, softeners are used when discussing effects of vaccines and, in particular, when outlining their potential threats. Paradoxically, even when mitigatory linguistic items are used, sensationalism in such mitigated claims is not avoided in either English or Lithuanian texts. However, sensationalism is even more evident when emphatics are employed.

4.3.2 Emphatics

As already mentioned, in terms of their overall frequency, emphatics are used with no significant difference between the two languages. As demonstrated in Table 5, where the most frequent intensifiers are provided, there are many correspondences in the repertoire of linguistic items used for intensification in both languages (highlighted in grey).

In total, there are 48 types of items used for intensification in English and as many as 52 types in Lithuanian. This is an unexpected outcome: considering that the Lithuanian data is substantially smaller, lower lexical variation was predicted in Lithuanian than English.

---

[6] *Reuters Fact Check*, https://www.reuters.com/article/factcheck-vaccines-safe-idUSL1N2P51T3.

[7] *Reuters Fact Check*, https://www.reuters.com/article/uk-factcheck-norway-idUSKBN29P2R1.

Table 5: Top items used for emphasis

Raw freq.	f/10,000	Emphatic	Raw freq.	f/10,000	Emphatic
71	14	very	17	14	labai (‘very’)
70	13	just	16	13	visiškai (‘completely’)
65	12	only	13	11	**daug (‘many/much’)**
50	10	**many**	10	8	ypač (‘especially’)
39	8	really	6	5	**daugelis (‘great number’)**
33	6	**much**	5	4	gerokai (‘considerably’/’significantly’)
13	3	must	5	4	**dauguma (‘majority’)**
10	2	especially	4	2	**daugybė (‘multitude’)**
9	2	real	4	3	itin (‘especially’)
9	2	significantly	3	2	būtent (‘exactly’)
8	2	clearly	3	2	būtina (‘necessary’)
8	2	entire	3	2	tikrai (‘really’)
8	2	major	2	2	aišku (‘certainly/clearly’)
			2	2	esminis (‘essential’)
			2	2	smarkiai (‘markedly’ / ’highly’)
			2	2	stipriai (‘heavily’)
			2	2	vien (‘only’/’just’)
			2	2	visiškas (‘complete’/‘entire’)

The most frequent intensifier in both languages is *very*/*labai* (‘very’), which tends to occur in highly evaluative contexts, e.g., in reference to conspiracy theories, such as the Big Pharma conspiracy theory, as in (6):

(6) a. It would be *very* foolish to dismiss Pfizer as simply incompetent …

b. Būtų *labai* kvaila vertinti korporaciją "Pfizer" kaip tiesiog nekompetetingą ...

In this example, an accusatory speculation in the source text and its Lithuanian translation is strengthened with the intensifier to evaluate sinister intentions of pharmaceutical companies.

In the set of emphatics (presented in bold in Table 5), the category of quantifiers (e.g., *daug* 'many'/'much' and *daugelis* 'great number') is of special importance. In general, non-numerical references to quantities are often used to evaluate and interpret quantities as being large or small. Multal quantifiers, which refer to large quantities, prevail in the current data and are especially frequent in Lithuanian, as shown in Table 6.

Table 6: Multal quantifiers

	Tokens (raw)	**Tokens (f/10,000)**	**f/10,000**	**Types**
EN	96	18	18	5 unique forms and lexemes
LT	42	34	33	22 unique forms; 9 different lexemes

The LL value (LL = 9.01; Log Ratio = -0.84) indicates moderate overuse of multal quantifiers in Lithuanian, which further suggests that Lithuanian disinformation tends to favour a more emphatic style.

A more extended example of a hyperbolic context achieved through intensification is represented in (7); here the angle brackets substitute elements which were omitted when rendering the passage in Lithuanian.

(7) a. There are *many* reasons to be wary of the COVID-19 vaccines, which have been rushed to market […] aggressively promoted. […] we will likely see an alarming increase in several major neurodegenerative diseases, including Parkinson's disease, […] Alzheimer's, and these diseases will show up with increasing prevalence among younger and younger populations, in years to come. Unfortunately, we won't know whether the vaccines caused this increase, because there will usually be a long time separation between the vaccination event and the disease diagnosis. *Very*

> convenient for the vaccine manufacturers, who stand to make *huge* profits off of our misfortunes …
> b. Yra *daug* priežasčių vengti kovido ligos vakcinų, kurios užplūdo rinką savo agresyvia taktika. Šios vakcinos gali sukelti pagrindines neurogeneracines ligas, tarp kurių—Parkinsono ir Alzheimerio ligos. Manoma, kad šios ligos vis jaunės. Deja, nežinome, ar šių ligų daugės dėl vakcinų, nes tarp skiepijimo ir ligos diagnozės praeis gana daug laiko. Tai *labai* naudinga vakcinų gamintojams, kurie *daug* uždirba iš mūsų bėdų.

The statements are made emphatic through the use of the adverb *very* and the quantifier *many* in English and Lithuanian and the adjective *huge* in English (translated as *daug* 'many'). The quantifiers here emphasise the harm caused by pharmaceutical companies and their sinister nature, but at the same time, by being highly non-specific, they make the claims elusive and vague.

Similarly to the present study, del Pilar's et al. (2017) research on satirical news in Mexican and Spanish data demonstrates that satirical tweets contain more quantifiers than non-satirical ones; however, there are no differences between Mexican and Spanish satirical texts. They also observe that categories such as certainty and quantifiers, among others, are strongly related to hyperbole (McCarthy and Carter 2004), which is used for extreme exaggerations in satirical news. Adverbs, which are predominantly used in their data to exaggerate or minimize a statement, also appeared as a reliable indicator of satirical tweets.

5. Conclusion

Even though the results of this study are of limited generalisability due to the sizes of the datasets used, this investigation serves as a useful pilot study that can help outline the possibilities and potential pitfalls of (contrastive) corpus-assisted analysis of disinformation. The findings of this study carry important implications regarding several major issues which emerged in the phase of data collection and in the empirical analysis related to the three research questions raised in this investigation.

First, as demonstrated in the initial stages of data collection, the very size of disinformation corpora is a major concern. It is relatively uncomplicated to collect a corpus of disinformation texts in English;

however, when lesser-used languages like Lithuanian are examined, the number of texts available is highly restricted.

A second important issue that emerged in the data collection process is the limited amount of original content in Lithuanian disinformation. Since most of the Lithuanian content comes from English sources, it may be questionable whether any distinct language-specific features are likely to emerge in Lithuanian disinformation texts. Still, the current analysis shows that even when texts are translated and/or summarised, some important differences between two languages can come to light.

On the other hand, when analysing such fluid phenomena as disinformation, the role of language arguably becomes just instrumental: English texts are generated fast and profusely, can be easily translated by disinformation producers into different languages and disseminated in fluid media spaces, often making it impossible for the researcher to know the original source, the translation tool, or the author's personal input into the output text. When large amounts of data are collected for corpus analysis, this meta-information can be difficult to control, but in contrastive linguistics these are central variables. One potential solution to this problem is to gather texts that explicitly indicate the author and to avoid those that are republished and/or lack clear authorship attribution.

Regarding the research questions raised in the empirical analysis, several conclusions can be drawn about the content and linguistic properties of English and Lithuanian disinformation. Concerning RQ1, which sought to evaluate the content adjustments in Lithuanian texts in comparison to the English source texts, the Lithuanian texts vary from closely matching translations to significantly condensed summaries of the source texts. Another notable trend is the comparatively shorter length of disinformation texts in Lithuanian. Unlike other trends identified in the current study, this aspect may not primarily stem from cross-linguistic differences. Instead, it may be that the frequency of disinformation posts is more important than the amount of information in each text, its argumentation and development. However, a more specific interpretation of this pattern of behaviour would require a different type of research instrument, such as surveys or interviews with disinformation conveyors, in order to know their real motivations for such choices.

Regarding RQ2, the type-token ratio (TTR) is very low in the English data (0.14) but higher in Lithuanian (0.39), which could be a result of typological differences between the two languages. In Lithuanian, as a

synthetic language, words can have more grammatical forms than in English, which results in more types in non-lemmatised data. Another reason may be the smaller size of the Lithuanian corpus, which gives a higher chance of encountering items that may appear as rare or unique words. However, overall, the TTR is very low in both English and Lithuanian as might have been predicted considering the results in earlier research. Disinformation texts in both datasets use uncomplicated language and minimal content; the latter quality is especially characteristic of Lithuanian data.

My third and final RQ aimed to assess how emphatics and tentative language are used in Lithuanian and English data. The study has shown that intensifiers are almost equally distributed in both datasets, but tentative language is of a considerably higher frequency in English. Such trends suggest that the language of disinformation tends to be simple; however, Lithuanian disinformation texts prioritise emphasis by retaining the same frequency of emphatic wording while reducing the tentative tone.

In further research, it is important to expand the corpus and develop more representative databases, which would include a larger diversity of disinformation outlets. It is also important to examine a larger number of lexical and syntactic categories for a more substantial comparison of the two languages. To assess the lexical diversity in Lithuanian disinformation texts in a more reliable way and to solve the issue of a large diversity of word forms, in future the TTR indices should be calculated for lemmatised data and adjusted for text length.

Acknowledgements

This study was carried out as a pilot study within the framework of the research project *Propaganda and disinformation research: automatic recognition using machine learning methods, 2023-2026* (No. S-VIS-23-8) funded by the Research Council of Lithuania (Priority Research Programme 'Strengthening Societal Resilience and Crisis Management in the Context of Contemporary Geopolitical Events').

References

Abonizio, Hugo Queiroz, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. 2020. Language-independent fake news detection: English, Portuguese, and Spanish mutual features. *Future Internet* 12(5): 1–18. doi:10.3390/fi12050087.

Ahmed, Hadeer, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy* 1(1): 1–15. doi: 10.1002/spy2.9.

Asubiaro, Victoria T., and Toluwase Victor Rubin. 2018. Comparing features of fabricated and legitimate political news in digital environments (2016–2017). *Proceedings of the Association for Information Science and Technology* 55(1): 747–750. doi: 10.1002/pra2.2018.14505501100.

Benkler, Yochai, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. New York: Oxford University Press. doi: 10.1093/oso/9780190923624.001.0001.

Bezerra, Ribeiro, and Jose Fabio. 2021. Content-based fake news classification through modified voting ensemble. *Journal of Information and Telecommunication* 5(4): 499–513. doi:10.1080/24751839.2021.1963912.

Damstra, Alyt, Hajo G. Boomgaarden, Elena Broda, Elina Lindgren, Jesper Strömbäck, Yariv Tsfati, and Rens Vliegenthart. 2021. What does fake look like? *A Review of the Literature on Intentional Deception in the News and on Social Media, Journalism Studies*, 22(14): 1947–1963. doi: 10.1080/1461670X.2021.1979423.

Da San Martino, Giovanni, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by Sebastian Padó, and Ruihong Huang, 5636–5646. Association for Computational Linguistics. doi:10.18653/v1/D19-1565.

Davoudi, Mansour, Mohammad R. Moosavi, and Mohammad Hadi Sadreddini. 2022. A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Systems with Applications* 198: 116635. doi.org/10.1016/j.eswa.2022.116635.

de Cock Buning, Madeleine. 2018. A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation. Publications Office of the European Union, https://data.europa.eu/doi/10.2759/739290.

del Pilar Salas-Zárate, María, Mario Andrés Paredes-Valverde, Miguel Ángel Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems* 128: 20–33. doi:10.1016/j.knosys.2017.04.009.

Du, Jiangshu, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and Philip S. Yu. 2021. Cross-lingual COVID-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand*, 859–862. doi.org/10.1109/ICDMW53433.2021.00110.

ERGA Report: Notions of disinformation and related concepts. 2020. https://erga-online.eu/wp-content/uploads/2021/03/ERGA-SG2-Report-2020-Notions-of-disinformation-and-related-concepts-final.pdf.

Faris, Robert M., Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center for Internet & Society Research Paper*. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33759251.

Farooq, Muhammad Shoaib, Naseem Ansar, Rustam Furqan, and Ashraf, Imran. 2023. Fake news detection in Urdu language using machine learning. *PeerJ Computer Science* 9: e1353. doi: 10.7717/peerj-cs.1353.

Faustini, Pedro Henrique Arruda, and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications* 158: 1–9. doi.org/10.1016/j.eswa.2020.113503.

Freelon, Deen, and Chris Wells. 2020. Disinformation as political communication. *Political Communication* 37: 145–156.

Fuller, Christie, David Biros, and Dursun Delen. 2011. An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications* 38(7): 8392–8398. doi:10.1016/j.eswa.2011.01.032.

Gelfert, Axel. 2018. Fake news: A definition. *Informal Logic* 38(1): 84–117.

Grabe, Maria, Shuhua Zhou, and Brooke Barnett. 2001. Explicating sensationalism in TV news: Content and the bells and whistles of form. *Journal of Broadcasting and Electronic Media* 45: 635–655.

Guibon, Gaël, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. 2023. Multilingual fake news detection with satire. In *Computational linguistics and intelligent text processing*, edited by Alexander Gelbukh, 392–402. Cham: Springer. doi:10.1007/978-3-031-24340-0_29.

Habgood-Coote, Joshua. 2019. Stop talking about fake news! *Inquiry* 62: 1033–1065.

Horne, Benjamin, and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media* 11(1): 759–766. doi:10.1609/icwsm.v11i1.14976.

Humprecht, Edda. 2019. Where 'fake news' flourishes: A comparison across four western democracies. *Information, Communication and Society* 22(13): 1973–1988.

Jarvis, Scott. 2002. Short texts, best-tting curves and new measures of lexical diversity. *Language Testing* 19(1): 57–84.

Kapantai, Eleni, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2020. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society* 23(5): 1301–1326. doi:10.1177/1461444820959296.

Kasseropoulos, Dimitrios Panagiotis, and Christos Tjortjis. 2021. An approach utilizing linguistic features for fake news detection. In *Artificial Intelligence applications and innovations (AICT 677),* edited by Ilias Maglogiannis, Lazaros Iliadis, Antonios Papaleonidas, and Ioannis Chochliouros, 646–658. Cham: Springer. doi.org/10.1007/978-3-030-79150-6_51.

Kar, Debanjana, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2021. No rumours please! A multi-Indic-lingual approach for COVID fake-tweet detection. In *2021 Grace Hopper celebration India (GHCI)*, 1–5. doi.org/10.1109/ghci50508.2021.9514012.

Kumar, Tirupathi B., and Vishnu B. Vardhan. 2021. A stylistic feature based approach for fake news spreaders detection. *Journal of Tianjin University Science and Technology* 54(9): 190–209. doi: 10.17605/OSF.IO/Z9DW5.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359: 1094–1096.

Marín-Arrese, Juana I. 2015. Epistemicity and stance: A cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish. *Discourse Studies* 17(2): 210–225.

McCarthy, Michael, and Ronald Carter. 2004. 'There's millions of them': Hyperbole in everyday conversation. *Journal of Pragmatics* 36(2): 149–184.

Meel, Priyanka, and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications* 153: 112986. doi.org/10.1016/j.eswa.2019.112986.

Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29(5): 665–675.

Orman, Levent. 1984. Fighting information pollution with decision support systems. *Journal of Management Information Systems* 1(2): 64–71. doi.org/10.1080/07421222.1984.11517704.

Põldvere, Nele, Elizaveta Kibisova, and Silje Susanne Alvestad. 2023. Investigating the language of fake news across cultures. In *The Routledge handbook of discourse and disinformation*, edited by Stefania Maci, Massimiliano Demata, Mark McGlashan, and Philip Seargeant, 153–165. London: Routledge. doi.org/10.4324/9781003224495-11.

Recasens, Marta, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Hinrich Schuetze, Pascale Fung, and Massimo Poesio, 1650–1659. Sofia: Association for Computational Linguistics. https://aclanthology.org/P13-1162.

Ruzaitė, Jūratė. 2007. *Vague language in educational settings: Quantifiers and approximators in British and American English*. Frankfurt am Main: Peter Lang.

Shu, Kai, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312–320: New York: Assocation for Computing Machinery. doi.org/10.1145/3289600.3290994.

Siwakoti, Samikshya, Kamya Yadav, Isra Thange, Nicola Bariletto, Luca Zanotti, Alaa Ghoneim, and Jacob N. Shapiro. 2021. Localized misinformation in a global pandemic: Report on COVID-19 narratives around the world. https://drive.google.com/file/d/1LqoK1lK4ufIQ3OisLVPBT99ikAD87wfX/view.

Sousa-Silva, Rui. 2022. Fighting the fake: A forensic linguistic analysis to fake news detection. *International Journal for the Semiotics of Law* 35: 2409–2433.

Staender, Anna, Edda Humprecht, Frank Esser, Sophie Morosoli, and Peter Van Aelst. 2021. Is sensationalist disinformation more effective? Three facilitating factors at the national, individual, and situational level. *Digital Journalism* 10(6): 976–996. doi:10.1080/21670811.2021.196631.

Wardle, Claire, and Hossein Derakshan. 2017. Information disorder: An interdisciplinary framework for research and policy for the Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Zhou, Lina, Jie Tao, and Dongsong Zhang. 2023. Does fake news in different languages tell the same story? An analysis of multi-level thematic and emotional characteristics of news about COVID-19. *Information Systems Frontiers* 25: 493–512. doi.org/10.1007/s10796-022-10329-7.

Zhu, Yonghua, Xun Gao, Weilin Zhang, Shenkai Liu, and Yuanyuan Zhang. 2018. Bi-directional LSTM-CNN model with attention for aspect-level text classification. *Future Internet* 10(12): 116. doi.org/10.3390/fi10120116.

Appendix

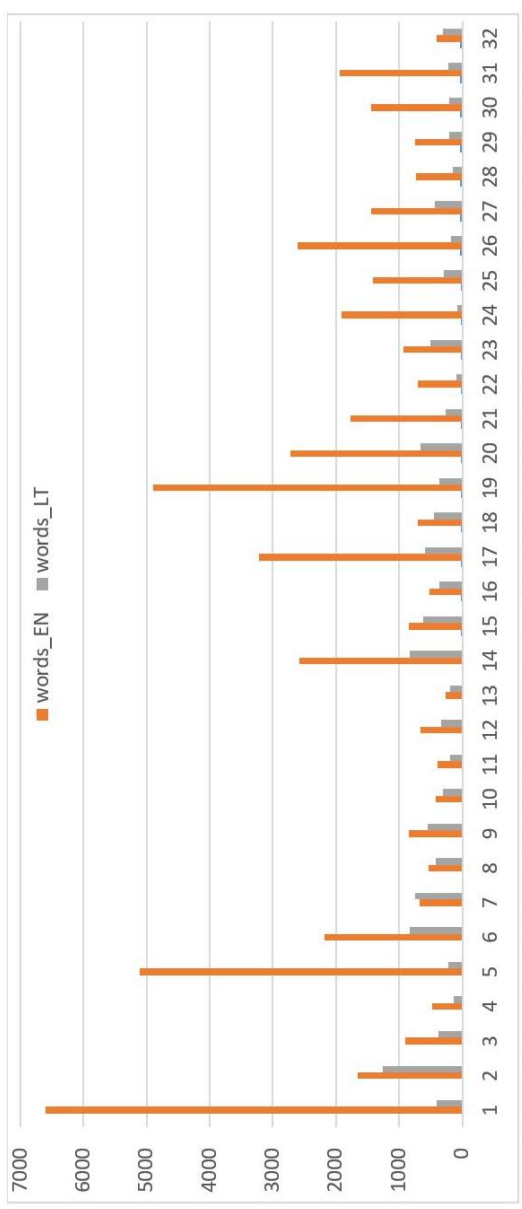

Figure A: Number of words in English (EN) and Lithuanian (LT) texts (vertical axis). The numbers on the horizontal axis are the identity numbers assigned to the corpus texts.