Sentiment Analysis of Russian Political Discourse: Does Translation Matter?

Lindy Comstock[1,*], Brandon Soung[2], Priyanshu Sharma[3]

1 Department of Psychiatry & Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior,
University of California, Los Angeles, Los Angeles, California, USA

2 Department of Political Science, University of California, Los Angeles, Los Angeles, California, USA

3 Department of Linguistics, University of California, Los Angeles, Los Angeles, California, USA

* corresponding author: lbcomstock@ucla.edu

1 Introduction

Current trends in the automated analysis of media texts endeavor to identify 'misinformation', i.e., the spread of misleading information. Emotional and subjective language are often exploited with the aim of intentional disinformation; however, misinformation may also result unintentionally when speaker intent is decoded incorrectly, in instances of 'pragmatic failure' (Thomas, 1983). The risk of pragmatic failure is compounded in cross-cultural communication, when speaker intent may be misinterpreted due to the transformation of meaning that occurs in translation (Lotman, 1990). Other factors that induce pragmatic failure include political or economic context (Clayman et al., 2006), sociolinguistic specificity (Arseniev-Koehler and Foster, 2020), and diachronic linguistic change (Fernández-Cruz and Moreno-Ortiz, 2023).

Automated analyses struggle to classify texts on the level of discourse pragmatics. Few authors question how pragmatic systems may be encoded across languages (Comstock, 2015), and whether this will affect the interpretation of their model outputs (Araújo et al., 2020; Balahur and Turchi, 2014). Utilizing a corpus of questions posed by journalists to the Russian president at international summits, this paper problematizes the assumption that a sentiment analysis performed on a source text and its translation will be equivalent. More generally, readers should be aware that sentiment analyses may have limited utility when the model does not account for text-specific pragmatic factors.

We collected the expected and observed lemma frequencies for the original Russian transcripts and their English translations. We then compared the sentiment classifications by (i) language, (ii) political context, and (iii) across presidential terms. We found significant differences in all three categories, underscoring the linguistic, contextual, and temporal specificity of sentiment analyses.

2 Related work

There is a growing body of work on sentiment analysis as a tool for identifying misinformation (Alonso et al., 2021; Kušen and Strembeck, 2018), including with Russian language data (Pocyte, 2019; Yaqub et al., 2020). However, even methods that adopt a sophisticated approach may assume that sentiment analyses reveal stable relationships representative of the language as a whole. Authors who attempt multilingual sentiment analysis often classify small, formulaic texts, such as business reviews (Abdalla and Hirst, 2017) and sentences (Araújo et al., 2020), or rely on abstract metrics to measure successful classification, such as the comparison of source and translation BLEU scores (Balahur and Turchi, 2014).

The existing literature acknowledges that determining whether the pragmatics of the translated text align more closely with the source language or the target language is a major problem in the field (Araujo et al., 2016; Sergey, 2020). Some authors argue that simple sentiment analyses using target language texts annotated for polarity items still outperform complex machine learning algorithms (AR et al., 2013; Basiri and Kabiri, 2017). Our approach adopts a simple analysis technique to ensure the relationship between the method and the findings remains transparent and interpretable and that we are assessing fundamental characteristics of the source and translated texts.

Yet the successful classification of discourse-level phenomena requires the synthesis of multiple linguistic features and domains (Becker et al., 2020). The strategic use of positive, negative, and subjective assessments is thought to underlie tactics used to spread disinformation, such as playing on emotions for political gain (Carrasco-Farré, 2022). Thus, we will classify co-occuring markers of polarity and subjectivity, which have a high potential to isolate contexts where misinformation may arise.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 268-272.
Irvine, California, June 27-29, 2024

Summit	Term	Russian	English
G8	2000-2003	757	874
	2004-2007	2129	2611
	2008-2011	1412	1709
	2012-2015	611	737
G20	2000-2003	–	–
	2004-2007	–	–
	2008-2011	1598	1887
	2012-2015	2241	2474
Total		12338	14667

Table 1: The number of words collected in each summit for the Russian transcripts and English translations.

3 Methods

The corpus comprised all the publicly available written transcripts of press conferences held by the Russian president at G8 and G20 summits from 2000-2015 (Comstock, 2023). Russian and English transcripts were accessed at the Kremlin online press archives. All questions were originally posed to the president in Russian, either directly or via a human translator. All transcripts were compiled by human translators from video recordings. Due to the different language types (analytic vs. synthetic), the English transcripts are ∼19% longer. Although the corpus is small in scope, it is comprehensive; therefore, it cannot be considered an insufficient sample size of an underlying distribution.

A composite list of all positive, negative, and subjective words was compiled from the Harvard IV-4, Loughran, McDonald, and Lexicoder sentiment dictionaries and lemmatized. Each lemma was then restricted to one of three sentiment lists (positive, negative, subjective) to ensure that cross-listed words would not force a correspondence between sentiment classification results. When used with a summit-specific meaning (e.g., “leader", “minister",“agreement", “good morning"), words were removed from the dataset to avoid inflating their representation in the dataset. The translation accuracy of the composite and sentiment lists was confirmed by a professional Russian translator.

The resulting sentiment lists were subdivided by political context and presidential term. Data was taken from international summits assumed to represent very similar contexts except for one dimension: the G8 summit is more exclusive than the G20 summit. The Russian president was Vladimir Putin from 2000-2007 and again in 2013-2105. The president was Dmitry Medvedev from 2008-2011.

We first compiled (i) the number of lemmas in each language dataset. We then calculated (ii) the observed frequency of each lemma per dataset as a percentage of the total words, and (iii) the expected frequency of each lemma as a percentage of its observed frequency in a larger corpus assumed to be representative of wider language norms. The expected frequencies were calculated using the subcorpus of media texts from the Russian National Corpus and the English Google Ngrams corpus. The total word counts, observed frequency, and expected frequencies were recorded for language, summit, and term. We performed an ANVOA and Tukey ASD test in JASP (JASP Team, 2024) for each analysis. A subset of relevant findings are reported below; the full set can be found on OSF.

4 Results

We observe in Figure 1 that the analyses do not remain consistent across similar political venues and short time differences. Positive and subjective lemma counts are greater in the G8 summit than in the G20 summit, seen as a marginally significant increase in positive lemmas between summits and a significant difference between positive and negative lemmas for the G8 summit only (see Tables 2-3). Although G8 data spans four terms and G20 data comprises only the last two terms, this difference cannot account for the finding: positive lemmas are at their highest level in the second two terms.

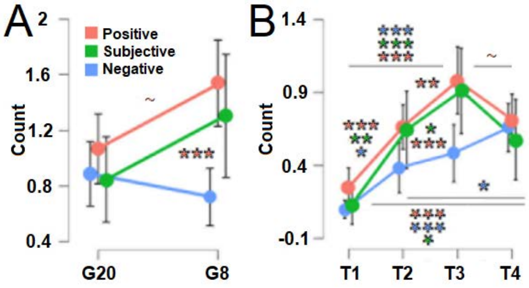

Figure 1: Sentiment analysis by summit and term. (A) Total positive, subjective, and negative lemma counts by summit (G8, G20). (B) Total positive, subjective, and negative lemma counts by presidential term (T1-T4).

A general increase in polarity and subjectivity is observed across terms with significant differences found between individual terms. Yet a significant difference between negative and both positive and subjective lemma counts appears only in the third term, which was presided over by Medvedev (See Tables 4-5. Only significant findings are shown).

Table 2: ANOVA: Fig.1A

	df	F	p
Summit	1	3.875	0.050
Sentiment	2	7.000	0.001
Summit*Sentiment	2	3.229	0.040
Residuals	504	1.809	

Table 3: Post Hoc Comparisons: Fig.1A

Summit	SE	t	p_{tukey}
G8-G20	0.129	−1.968	0.050
Sentiment	SE	t	p_{tukey}
Neg.-Pos.	0.133	−3.741	$< .001$
Neg.-Sub.	0.172	−1.570	0.260
Pos.-Sub.	0.165	1.371	0.357
Sentiment	SE	t	p_{tukey}
Pos. G8-G20	0.174	−2.699	0.077

Table 4: ANOVA: Fig.1B

	df	F	p
Summit	2	9.564	$< .001$
Term	3	44.877	$< .001$
Summit*Term	6	3.187	0.004
Residuals	2419		

A reduction in positive and subjective lemmas occurs from the third to fourth term. Both counts are significantly different from the negative count in the third but not fourth term. The increase and decrease in positive lemmas is significant and marginally significant, respectively.

The sentiment analyses in Figure 2 also reveal significant language-specific findings. Positive lemmas appear less frequently than expected in the original Russian data, but more frequently than expected in English translation. This cannot be explained by the differences between datasets: with a higher total word count, we would expect the reverse pattern in the absence of a real trend.

While we observe that the subjective lemmas appear to partially replicate this trend, in the sense that they appear less frequently than anticipated in the Russian data, the finding did not reach significance in this small dataset (see Figures 6-7. Only significant findings in poshoc tests shown).

Table 5: Post Hoc Comparisons: Fig.1B

All Factors	SE	t	p_{tukey}
Neg. T1-Sub. T1	0.094	−0.206	1.000
Neg. T1-Pos. T1	0.071	−1.575	0.918
Sub. T1-Pos. T1	0.088	−1.049	0.996
Neg. T2-Sub. T2	0.094	−2.133	0.599
Neg. T2-Pos. T2	0.071	−2.502	0.339
Sub. T2-Pos. T2	0.088	0.258	1.000
Neg. T3-Sub. T3	0.094	−3.554	0.020
Neg. T3-Pos. T3	0.071	−4.909	$< .001$
Sub. T3-Pos. T3	0.088	−0.167	1.000
Neg. T4-Sub. T4	0.094	1.247	0.985
Neg. T4-Pos. T4	0.071	0.270	1.000
Sub. T4-Pos. T4	0.089	−1.111	0.994
Neg. T1-Neg. T2	0.078	−3.408	0.032
Neg. T1-Neg. T3	0.078	−4.501	$< .001$
Neg. T1-Neg. T4	0.078	−6.687	$< .001$
Neg. T2-Neg. T4	0.078	−3.279	0.049
Sub. T1-Sub. T2	0.108	−4.160	0.002
Sub. T1-Sub. T3	0.108	−6.196	$< .001$
Sub. T1-Sub. T4	0.108	−3.575	0.018
Pos. T1-Pos. T2	0.063	−5.254	$< .001$
Pos. T1-Pos. T3	0.063	−9.312	$< .001$
Pos. T1-Pos. T4	0.063	−6.190	$< .001$
Pos. T2-Pos. T3	0.063	−4.058	0.003
Pos. T3-Pos. T4	0.063	3.121	0.078

The negative lemmas are observed at roughly the expected frequency in both languages. Notably, in the English translation, all lemmas are observed at roughly the expected frequency, across all terms, despite the specific political context.

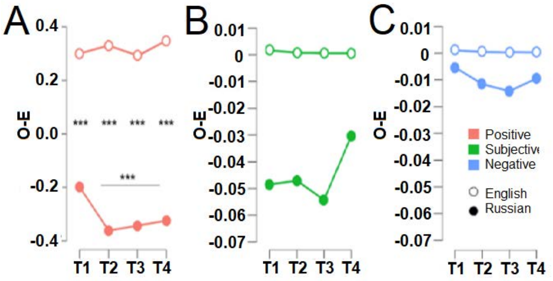

Figure 2: Sentiment analysis by sentiment type. Each graph depicts the difference between the expected and observed frequencies for lemmas of (A) positive, (B) subjective, and (C) negative sentiment types, reported as the difference between observed and expected values.

Table 6: ANOVA: Fig2

	df	F	p
Lang	1	5.595	$< .001$
Summit	2	0.029	0.441
Term	3	0.017	0.687
Lang*Term	3	0.013	0.771
Summit*Term	6	0.014	0.878
Lang*Summit*Term	6	0.022	0.701
Residuals	802		

Table 7: Post Hoc Comparisons: Fig.2

Language	SE	t	p_{tukey}
Eng.-Rus.	0.018	12.641	$< .001$
Sentiment	**SE**	**t**	p_{tukey}
Neg.-Sub.	0.025	0.702	0.762
Neg.-Pos.	0.018	-0.558	0.842
Sub.-Pos.	0.022	-1.255	0.421
All Factors	**SE**	**t**	p_{tukey}
Pos. Eng.-Rus.	0.020	31.102	$< .001$
Pos. T1 Eng.-Rus.	0.057	8.715	$< .001$
Pos. T2 Eng.-Rus.	0.033	20.714	$< .001$
Pos. T3 Eng.-Rus.	0.031	20.327	$< .001$
Pos. T4 Eng.-Rus.	0.033	20.352	$< .001$

5 Discussion

Our analyses illustrate that sentiment analyses of a source text and its translation can not be relied upon to produce equivalent findings. The translated text largely reproduced the expected distribution of words with emotional and subjective content. In the given example, this tendency resulted in a notable increase in positive lemmas in translation, when the observed frequency of positive lemmas was in fact less than expected in the Russian source.

Positive and subjective lemmas co-occured in the text. The trend was more apparent in lemma counts when no language distinction was made; Russian-specific analyses found a non-significant trend between their observed and expected frequencies. This may reflect language-specific pragmatic norms in English to upgrade positive assessments and minimize negative ones (Lindström and Sorjonen, 2012; Markkanen and Schröder, 1997).

Translators employed fewer lemmas: 36% and 37% fewer in the positive and subjective lists, respectively, and 19% fewer in the negative list. While this correspondence may in part underlie the observed trends, we also note that the total count differed substantively by sentiment: 168 Eng./262 Rus. positive lemmas, 122 Eng./151 Rus. negative lemmas, 58 Eng./92 Rus. subjective lemmas.

These findings do not mean a human translator is less ideal. Such discrepancies can be advantageous: a text will read more naturally when it conforms to target language norms. However, this practice also changes the emotional tone of the text, which could lead to misleading conclusions (Mohammad et al., 2016); Russian journalists have been accused of adopting a lenient tone in questioning their president (Comstock, 2009, 2023), when in fact their questions carry significantly less positive emotional content than observed in the official translations.

Our analysis was relatively simple quantitatively. The question arises whether a more sophisticated model would yield improved findings. We note that a surprising number of authors continue to base their methods on simple natural language processing algorithms and that deriving more complex relationships between data that is fundamentally different will not erase underlying distinctions: any model based on word vectors of semantic or distributional information will be compromised.

6 Conclusion and limitations

Two limitations should be noted: we utilized a media-specific corpus for expected Russian lemma frequencies, but a corpus that was not constrained by genre for expected English lemma frequencies; the Google Ngrams corpus is also substantially larger. While a larger corpus may yield additional linguistic features for the training of sophisticated machine learning models, we believe a smaller but well-curated corpus can effectively represent word frequencies (Stubbs, 2004). In any corpus, the frequencies represented are contingent on the quality and applicability of the source texts included.

Overall, researchers must consider how the choice to utilize translated texts may influence the assessment of linguistic features and subsequent findings. Despite breakthroughs in AI and machine learning technologies, it is essential to understand the pragmatic specificity of corpora in order to develop reliable methodologies that ensure accurate interpretations of foreign language discourse.

7 Acknowledgements

We thank Mikhail Belov for his assistance in the gathering of Russian language frequency data.

8 Data Availability

All data and analyses are available at OSF to allow reproducibility under the CC BY 4.0 license.

References

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. *arXiv preprint*.

Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. 2021. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348.

Balamurali AR, Mitesh M Khapra, and Pushpak Bhattacharyya. 2013. Lost in translation: viability of machine translation for cross language sentiment analysis. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14*, pages 38–49. Springer.

Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.

Matheus Araujo, Julio Reis, Adriana Pereira, and Fabricio Benevenuto. 2016. An evaluation of machine translation for multilingual sentence-level sentiment analysis. pages 1140–1145.

Alina Arseniev-Koehler and Jacob G Foster. 2020. Sociolinguistic properties of word embeddings.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Mohammad Ehsan Basiri and Arman Kabiri. 2017. Translation is not enough: comparing lexicon-based methods for sentiment analysis in persian. In *2017 international symposium on computer science and software engineering conference (CSSE)*, pages 36–41. IEEE.

Maria Becker, Michael Bender, and Marcus Müller. 2020. Classifying heuristic textual practices in academic discourse: A deep learning approach to pragmatics. *International Journal of Corpus Linguistics*, 25(4):426–460.

Carlos Carrasco-Farré. 2022. The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, 9(1):1–18.

Steven E Clayman, Marc N Elliott, John Heritage, and Laurie L McDonald. 2006. Historical trends in questioning presidents, 1953-2000. *Presidential Studies Quarterly*, 36(4):561–583.

Lindy Comstock. 2009. Russian political interviews.

Lindy Comstock. 2023. Journalistic practice in the international press corps: Adversarial questioning of the russian president. *Journal of Language Aggression and Conflict*, 11(2):145–175.

Lindy B Comstock. 2015. Facilitating active engagement in intercultural teleconferences: A pragmalinguistic study of russian and irish participation frameworks. *Intercultural pragmatics*, 12(4):481–514.

Javier Fernández-Cruz and Antonio Moreno-Ortiz. 2023. Tracking diachronic sentiment change of economic terms in times of crisis: Connotative fluctuations of 'inflation'in the news discourse. *Plos one*, 18(11):e0287688.

JASP Team. 2024. JASP (Version 0.18.3)[Computer software].

Ema Kušen and Mark Strembeck. 2018. Politics, sentiments, and misinformation: An analysis of the twitter discussion on the 2016 austrian presidential elections. *Online Social Networks and Media*, 5:37–50.

Anna Lindström and Marja-Leena Sorjonen. 2012. Affiliation in conversation. *The handbook of conversation analysis*, pages 250–369.

Yuri Lotman. 1990. *Universe of the mind: A semiotic theory of culture*. Indiana University Press.

Raija Markkanen and Hartmut Schröder. 1997. Hedging: A challenge for pragmatics and discourse analysis. *Research in Text Theory*, pages 3–20.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Agniete Pocyte. 2019. From russia with fear: The presence of emotion in russian disinformation tweets.

Smetanin Sergey. 2020. The applications of sentiment analysis for russian language texts: Current challenges and future perspectives. 8:110693–110719.

Michael Stubbs. 2004. Language corpora. *The handbook of applied linguistics*, pages 106–132.

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied linguistics*, 4(2):91–112.

Ussama Yaqub, Mujtaba Ali Malik, and Salma Zaman. 2020. Sentiment analysis of russian ira troll messages on twitter during us presidential elections of 2016.