A comparison of two text specificity measures analyzing a heterogenous text corpus

Anton Oleinik[1*]

[1] Memorial University of Newfoundland, St. John’s, Canada
[*] Corresponding author’s email: aoleynik@mun.ca

DOI: https://doi.org/10.53482/2023_54_404

ABSTRACT

The article compares the performance of two term specificity measures, Cohen’s d and Z-score, when analyzing political and media discourses on Russia’s war in Ukraine in four languages and five countries. In addition to linguistic and stylistic heterogeneity, 3,347 texts included in the corpus have variable length. The two measures display convergent validity, as confirmed by various performance metrics. It is argued that the measures can be adapted to a broader range of tasks in information retrieval and digital humanities, in addition to their usefulness for text mining and content analysis.

Keywords: text specificity, term specificity, text mining, content analysis

1 Introduction

Although information retrieval and text mining refer to separate fields of knowledge and lie at the origin of different application systems (search engines as opposed to text analytics programs), they remain closely interconnected (Zhai, Massung, 2016). The present article highlights an aspect of their interconnectivity related to text specificity measures. On the one hand, such measures assist in content analysis of texts helping to identify terms (words, n-grams) that distinguish one document from the other. On the other hand, in information retrieval text specificity measures may allow to assess the extent to which searchable documents are about the same thing as a search query treated as a text, although short (Savoy, 2019). Text similarity also plays a role in several other NLP (Natural Language Processing) based tasks, such as automatic question answering, machine translation, dialogue systems and document matching (Wang, Dong, 2020) and digital humanities, broadly understood.

The idea of comparing text specificity measures emerged when content analyzing a corpus of political and media discourses on Russia's war in Ukraine in the framework of a larger ongoing research project. It was necessary to identify terms that would help distinguish discourses in function of their source (countries, political leaders, and media). No text specificity measure emerged as an obvious and uncontested choice, however. The performance of two measures, one based on Cohen's d and the other introduced by Savoy (2016), Z-score, is thus compared. The proposed analysis aims to identify convergent and divergent patterns in the outcomes obtained with their help. The research question can be formulated as to whether d and Z can be used interchangeably, or they have their own areas of application.

2 Related work

The need for having text representation and calculating distances between texts exists in information retrieval and text mining alike (Wang, Dong, 2020). Text representation involves viewing a text as a set of numerical features, for instance, as a bag-of-words. The order of words in the 'bag' is deemed to be irrelevant. Only their frequencies count. Since it was found that the usefulness of a term for content representation increases with the frequency of the term in the document but decreases with the number of documents, various weighting schemes, such as TF*IDF (term frequency by inverse document frequency), are commonly used (Salton, McGill, 1983; Evans et al., 2007; Jurafsky, Martin, 2008; Manning, Raghavan, Schütze, 2008; Savoy, 2016; Diermeier et al., 2011).

The transformation of texts into vectors paves the way to assessing their similarity and, ultimately, calculating distances between them. The relative position of texts included in a corpus is thus determined. Distance measures include Euclidean distance, Cosine distance, the Jaccard index, etc.

In corpus-based approaches text representation always has a relative, as opposed to absolute, character. If the same text is compared with the other set of documents, its representation changes. This caveat needs to be borne in mind when determining specific terms that characterize a text. Specificity refers to terms used to distinguish the text content (Salton, McGill, 1983). Specificity does not belong to a given document but terms that can be used to discriminate between two (or more) text categories (e.g., authors, text genres, etc.). For instance, what are the terms and expressions that best characterize each source of political and media discourses on Russia's war in Ukraine?

An approach to operationalizing specificity consists in building a dictionary. This process can be either theory- or data-driven (Simon, Xenos, 2004). In addition to content words serving to name things, express relations, perceptions, states or actions, the data-driven approach will extract many functional words (and, or, above, etc.). Functional words are normally excluded from the analysis. A data-driven dictionary includes m most frequent content word types or lemmas, with m varying from 50 to 1,000

(Burrows, 2002; Savoy, 2017; Savoy, 2019; Juola, Mikros, Vinsick, 2019). All words outside of the top m are also excluded from the analysis as uninformative. A theory-driven dictionary contains terms identified with the help of literature review. For instance, McClelland's motive imagery model borrowed from psychology was used to identify textual signals prefiguring military threats in political discourses on the Islamic Republic of Iran (Hogenraad, Garagozov, 2014).

It is at this stage that term specificity measures become necessary. They allow to quantitatively assess the specificity of terms included in the dictionary. The higher the value of a specificity measure for a term, the better it helps distinguish between texts included in a corpus. Several term specificity measures are known and used in application systems. A version of Cohen's d, a popular effect-size measure, is one (Shalak, 2004; Warner, 2013). It is implemented in several programs for content analysis, such as WordStat and VAAL. The other is Z-score, a version of the distance of a term score from the mean of a distribution expressed in unit-free terms (Savoy, 2016).

Both term specificity measures involve comparing the observed term frequency with its expected frequency calculated from corpus-level data. The calculation of the expected frequency is based on the assumption that the term is evenly distributed across all documents in the corpus. The idea of comparing the observed and the expected frequencies can be traced back to a generic chance-corrected measure (Goodman, Kruskal, 1954): $M_{cc} = \frac{M-E(M)}{M_{max}-E(M)}$, where M_{CC} is the chance-corrected measure, $M_{\max}$ is the maximal value M can reach, and $E(M)$ is the value expected for a null model. In the circumstances, the null model assumes that text category (e.g., authors, text genres, etc.) does not have an impact on the distribution of the term across documents.

The algorithms for calculating d and Z for the ith term denoted t_i differ, however.

(1) $$d\ (t_i) = \frac{\frac{tf_{i0}}{N_0} - \frac{tf_i}{N}}{\sqrt{\frac{\sum_i^m (\frac{tf_{i0}}{N_0} - \frac{tf_i}{N})^2}{m}}}$$

where m is the size of the dictionary (and also the query length in the context of information retrieval), tf_i – term i frequency, N_0 – the document word count, and N – the corpus word count.

There is a version of d adapted for the purposes of information retrieval, Standard Document Score, or SDS (Cummins, 2013):[1]

(2) $$SDS(m, D) = \frac{1}{\sqrt{|m|}} \sum_t^m \frac{(tf_{i0} - E[X^{t_i}])}{\sigma(X^{t_i})}$$

[1] The reviewer pointed out to some inconsistencies in this formula. However, it is kept in the version found in the source (Cummins, 2013, p. 114) since the possible errors do not affect the analysis in this article.

where $E[X^{t_i}]$ and $\sigma(X^{t_i})$ are the expected value and the standard deviation of the random variable X^{t_i} respectively, and |m| – the query length. In contrast to d, which is calculated at the term level, SDS is calculated at the document level. The SDS can be interpreted as the number of standard deviations a document is from the average document score for a specific query.

To calculate d, the difference between the observed and the expected *relative* frequencies is divided by the standard deviation. d has no lower or upper limit, whereas the cut-off values suggested in the literature are 0.8 for positive values and -0.8 for negative values (Warner, 2013).

(3)

$$Z\ (t_i) = \frac{tf_{i0} - N_0 \times \frac{tf_i}{N}}{\sqrt{N_0 \times \frac{tf_i}{N} \times (1 - \frac{tf_i}{N})}}$$

To calculate Z, the difference between the observed and the expected *absolute* frequencies is divided by the binomial variance. Z does not have a lower or upper limit either. The suggested cut-off values are 3 and -3: terms overused in a document have a Z score higher than 3 (and corresponding to 0.14% of the Gaussian distribution) whereas underused terms have a Z score lower than -3 (Savoy, 2016).

d is distinguishable from Z along the following lines:

1. Relative, as opposed to absolute, frequencies are used in numerator,
2. Standard deviation, as opposed to binomial variance, is used in denominator, and
3. The absolute value of d depends on m whereas that of Z does not depend on the size of the dictionary/query length.

Those differences suggest that Z may be more sensitive to small counts than d, whereas d appear to be suitable to analyze large corpora. To compare the behavior of d and Z when calculated for the same texts, their relative performance was assessed using standard performance metrics, precision, recall, accuracy, F_1-measure and Cohen’s Kappa (Evans et al., 2007; Jurafsky, Martin, 2008; Zhai, Massung, 2016; Vellino, Alberts, 2016; Khan, Qamar, Bashir, 2017; Dourado et al., 2019). Although human input is normally used as the gold standard in computer science (DiMaggio, 2015), in this case the other term specificity measure plays this role. In other words, an attempt to cross-validate d and Z was made. If their performance has convergent patterns, it is indicative of a consensus in the determination of term specificity (Mathet, Widlöcher, Métivier, 2015). Divergent patterns, on the other hand, would lead us to believe that d and Z cannot be used interchangeably, and it is up to the user to choose which one best serves her needs. Like many other problems in information retrieval and text mining, the choice of the term specificity measure is empirically defined. Which one works better cannot be answered by pure analytical reasoning or mathematical proofs (Zhai, Massung, 2016).

3 Political and media discourses on Russia's war in Ukraine

A highly heterogenous corpus was used for the purpose of comparing the performance of d and Z. It includes 3,347 texts in four languages of variable length, from 232 words to 145,237 words (55,599,283 words in total), discussing Russia's war in Ukraine. In addition to their variable length, texts also exemplify different genres: speeches of political leaders (political discourse) and news items (media discourse). The corpus covers five counties: two belligerents, the United States, and two European counties, the United Kingdom and France. In the case of Russia, Ukraine and the United States the coverage is more comprehensive since three media were monitored in each of those countries (*Kommersant*, *Izvestia*, and *Gazeta.ru*; *Ukrainska Pravda*, *RBC-Ukraina* and *Liga.net*; *New York Times*, *Washington Post* and *CNN* respectively). In the United Kingdom and France, one media was selected (*The Times* and *Le Monde* respectively). Transcripts of political leaders' speeches were retrieved from their official websites. Transcripts of speeches of members of legislative bodies were retrieved from the legislative bodies' official websites (the Russian Duma, the Ukrainian Rada, the US Congress, the British Parliament, and the French Assemblée Nationale). In total, the data came from 21 source.

Raw term frequencies (terms are only sequence of letters or short sequence of words, n-gram) were calculated using WordStat computer program. All subsequent calculations were performed with the help of custom algorithms.

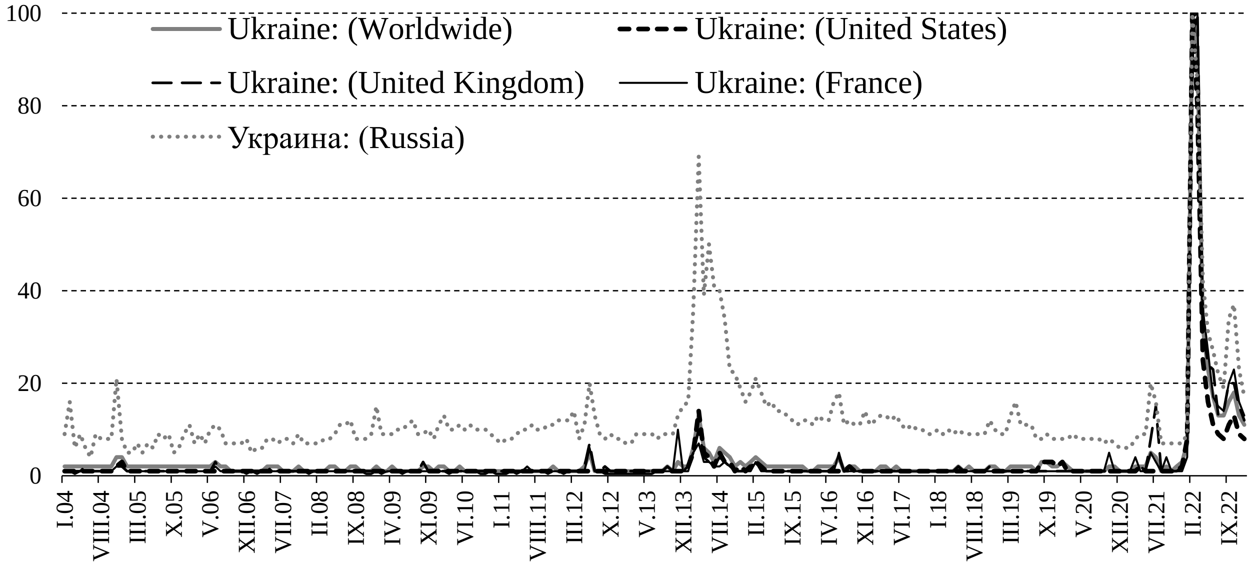

Figure 1: Relative popularity of 'Ukraine' as an internet search term in the US, the UK, France, Russia and worldwide, 2004-2022 (Source: Google Trends; a value of 100 is the peak popularity for the term).

Although Russia's aggression against Ukraine started in 2014, it transformed into an all-out war eight years later, on February 24, 2022. The corpus covers first eight months of the full-fledged war which significantly increased the demand for information about Ukraine and the situation in this country at the international level (Figure 1). During this period, Ukraine managed to contain Russia's initial attacks and started to progressively regain control over the territories occupied by Russia. Political and media discourses included in the corpus represent an informational dimension of the war. As Lasswell (1938) once observed, efforts to control opinions constitute one of the three chief implements of warfare, along with military pressure and economic pressure.

When studying informational warfare, one needs to identify the terms and expressions that can best characterize the discourse in each country directly or indirectly involved in the conflict. If such terms differ across the countries, then the hypothesis that war coverage does not allow establishing the truth finds some support (Lasswell, 1938; Knightley, 2003). One of the research questions addressed in the larger project is whether the differences in war coverage are territorially segregated according to national boundaries in the case of Russia's war in Ukraine.

A dictionary was built for the purpose of content analyzing the corpus. When assessing the alternative research strategies, including topic analysis (DiMaggio, Nag, Blei, 2013; Zhai, Massung, 2016), the use of the dictionary was deemed to be a better option. Although Multilingual Probabilistic Topic Models allow discovering topics in corpora composed of texts written in different languages (Lind et al., 2022), they do not lessen the other requirement of homogeneity. In the circumstances, text lengths and genres (news items as opposed to speeches of political leaders) vary significantly. The quadrilingual dictionary includes 246 categories (about 400 words in each version – Ukrainian, Russian, English and French – since some categories include more than one word). The dictionary was compiled using a combination of theory- and data-driven approaches. Along with most frequent terms, it contains those commonly discussed in the extant literature on war coverage. For example, 'war' is often described in terms of 'special military operation' (Lukin, 2013), as in the Russian case. The mandated substitution of 'special military operation' for 'war' helps members of Russia's power elite to create a perception of the aggression as limited in scope and not inherently violent. The 246 categories were weighted by TF*IDF whereas texts lengths were normalized. The TF*IDF values were calculated for the entire corpus using the formula $TF * IDF_i = tf_i \times \log\left(\frac{D}{df_i}\right)$, where D is the total number of documents in the corpus and df_i – the number of documents containing term i (Manning, Raghavan, Schütze, 2008; Jiang, Li, 2012).

4 Two measures of term specificity compared

The two lists generated for the group of leaders are discussed in detail for the purpose of illustration. As in all other subsamples, lengths of their speeches devoted to Russia's war in Ukraine vary from 8,115 words (French President Macron) to 322,596 words (Ukraine's President Zelensky). President

Zelensky delivered at least one address (sometimes up to four) to the national and the international audiences every day during the period under consideration. Lengths of war-related speeches of Russian President Putin's (49,169 words), US President Biden (31,175), and the then UK Prime Minister Johnson (25,242) lie in between.

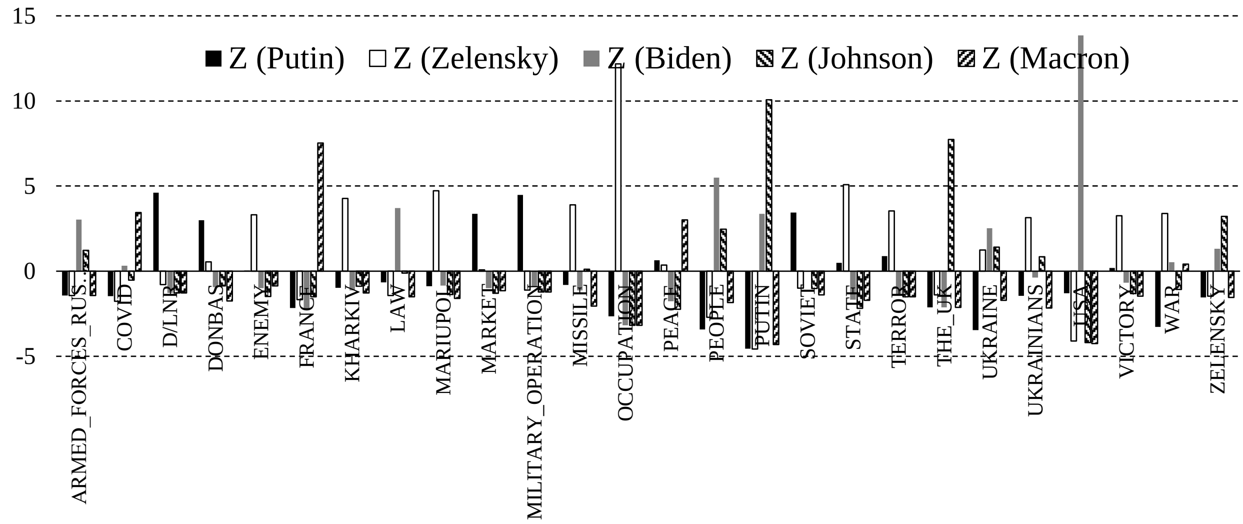

Figure 2: Z scores for the most specific terms of five political leaders.

Since in the cases of France and the UK one media only was monitored, the performance of d and Z was compared on the basis of four subsamples containing five sources of data each: the group of five leaders (Putin, Zelensky, Biden, Johnson and Macron), Russia, Ukraine, and the US. Most specific terms were identified for each subsample using d and Z, after which their lists so compiled were cross-checked. Six performance metrics, precision, recall, accuracy, F_1-measure, Cohen's Kappa and Pearson's r, informed the comparison of five pairs of the lists of most specific terms. The addition of Pearson's r allowed disregarding cut-off values that in the case of d may be arbitrary to some extent (they are set by convention): correlations were run between raw scores of d and Z.

The list of terms whose Z-scores exceed |3| includes 26 items (Figure 2). The list of terms whose d-values exceed |0.8| contains 39 items (Figure 3). Those lists overlaps to a significant extent. 25 terms are present on both lists: Covid, D/LNR (the acronyms for the Donetsk and the Luhansk People's Republics, two entities created in 2014 and supported by Russia in Eastern Donbas), Donbas, enemy, France, Kharkiv, law, Mariupol', market, military operation, missile, occupation, peace, people, Putin, Soviet, the State, terror, the UK, Ukraine, Ukrainians, USA, victory, war, and Zelensky.

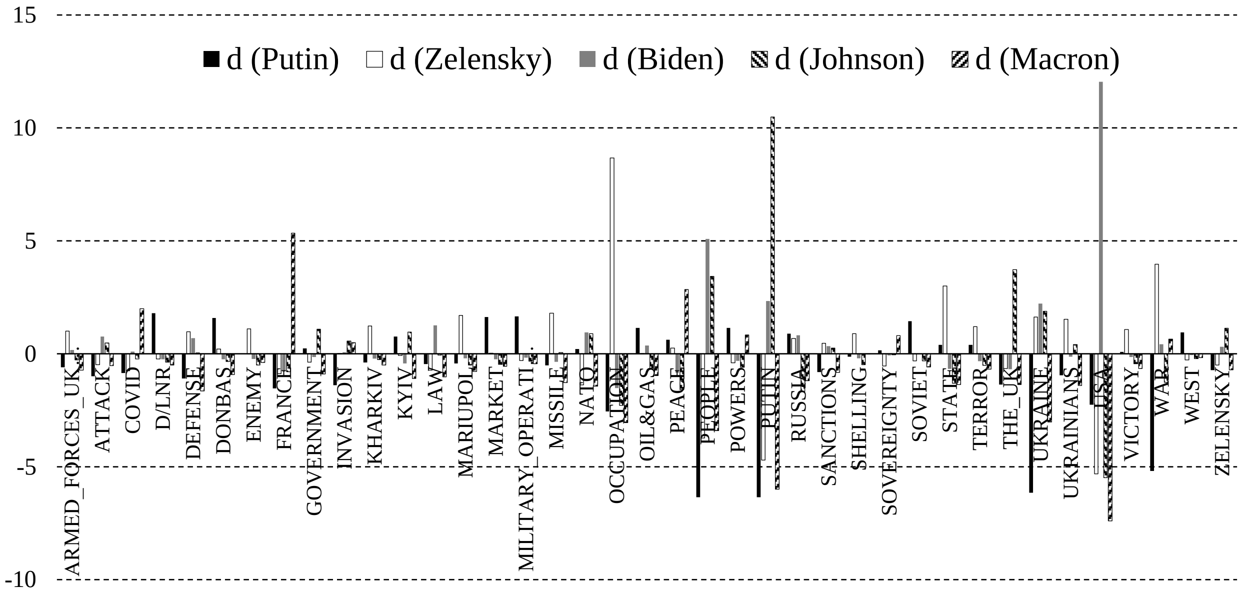

Figure 3: 'd values for the most specific terms of five political leaders'.

The 25 terms constitute markers of the presidents' discourses. When a president overuses a term, it becomes a positive marker of his discourse. When a president underuses a term, it is deemed to be a negative marker of his discourse. A negative character of a marker does not mean that a negative connotation is attached to it. The term is simply used by a president significantly less frequently than by his fellows. Vice versa, the term-positive marker has no value judgment attached to it. Sentiment analysis would be needed to discern value-judgments attached to the markers.

Positive markers for Putin are D/LNR, Donbas, market, military operation, Soviet; negative – people, Putin, Ukraine, and war. Waging a war against Ukraine, Putin nevertheless avoids naming his opponent and framing the aggression as a war. Positive markers for Zelensky are enemy, Kharkiv, Mariupol', occupation, the State, terror, Ukrainians, victory, war; negative – Putin and the US. His discourse has more local detail than in other cases and is centered on the war's impact on Ukraine instead of geopolitics. Positive markers for Biden are law, people, Putin, and the US; negative – occupation. Positive markers for Johnson are Putin, the UK, and Zelensky; negative – occupation and the US. Positive markers for Macron are Covid, France, and peace; negative – occupation, Putin, and the US.

Since the use of the term 'military operation' is mandated in Russia and the term 'war' – banned, the fact that the first is consistently (as per relevant d-value and Z-score) overused and the second – underused in this country comes as no surprise. The consistent overuse of the term 'the State' by Zelensky is more noteworthy (tf_0=1,974, tf=2,135, d=3.01, Z=5.08). Although the process of state-building started in Ukraine almost from scratch after the declaration of its independence in 1991 (Harasymiw, 2002), it appears that the ongoing war provided additional and powerful incentives to intensify this process: '36 days! 36! This is how long our State, our people have been able to stand against the army which was deemed to be among the best in the world' (Zelensky, March 31). 'I want to thank separately the inhabitants of our city of Energodar. Those brave Ukrainians who went down to the streets today to defend their city, to defend our State' (Zelensky, April 2).

There is one term with the Z-score exceeding the cut-off value yet with the d-value not reaching it, 'Russian Armed Forces.' It can relatively frequently be found in President Biden's speeches (tf$_0$=25, tf=37, d=3.04, Z=0.65): 'Thanks to the aid we've provided, Russian forces have been forced to retreat from Kyiv' (Biden, April 28).

The list of terms with d-values exceeding the cut-off value and Z-scores below the cut-off value contains 14 items: Ukrainian Armed Forces (Zelensky, tf$_o$=223, tf=239, Z=2.9, d=1.01), attack (Putin, tf$_o$=6, tf=176, Z=-1.63, d=-0.99), defense (Zelensky, tf$_o$=995, tf=1,195, Z=1.6, d=0.97), government (Johnson, tf$_o$=38, tf=250, Z=2.26, d=1.08), invasion (Putin, tf$_o$=3, tf=230, Z=-2, d=-1.4), Kyiv (Johnson, tf$_o$=37, tf=381, Z=1.71, d=0.97), NATO (Biden, tf$_o$=60, tf=265, Z=2.14, d=0.95), oil & gas (Putin, tf$_o$=60, tf=352, Z=1.5, d=1.14), powers (Putin, tf$_o$=93, tf=221, Z=2.21, d=1.15), Russia (Putin, tf$_o$=363, tf=3,136, Z=0.76, d=0.89), sanctions (Macron, tf$_o$=5, tf=591, Z=-1.1, d=-0.83), shelling (Zelensky, tf$_o$=236, tf=251, Z=2.72, d=0.89), sovereignty (Macron, tf$_o$=17, tf=174, Z=1.6, d=0.81), and the West (Putin, tf$_o$=51, tf=99, Z=2.4, d=0.95). For instance, references to sovereignty are common in the discourse of France's President Macron: 'France and Europe responded to this flagrant violation of the territorial integrity and the sovereignty of a European country with no delay and with resolution' (Macron, March 2).

Overall, the two term specificity measures show more convergency than divergency (Table 1). The average value of F$_1$, 0.64, is within the acceptable range. For instance, in a study of 11,089 front-page news articles using a dictionary of 20 categories, the reported F$_1$, 0.68, was similar (Burscher, Vliegenthart, De Vreese, 2015). The average value of Cohen's Kappa can be interpreted as substantial since it falls within the range from 0.61 to 0.8 (Warner, 2013). The average value of Pearson's r is also indicative of a substantial to strong relationship. One needs to bear in mind that the choice of the reference point, d or Z, affects only the values of precision and recall (precision becomes recall and vice versa), whereas the other performance metrics remain the same.

Table 1: Average values of recall, precision, accuracy, F$_1$, Cohen's Kappa and Pearson's r for four subsamples.

precision	0.7065
recall	0.7697
accuracy	0.9547
F$_1$	0.6412
Cohen's Kappa	0.6225
Pearson's r	0.8745

Since d and Z show convergent validity, they appear to measure the same thing, term specificity. At the same time, a look at the instances of misclassification in which one term specificity measure exceeds the cut-off point whereas the other does not suggests that Z tends to be more sensitive to small counts than d, whereas d – more suitable to analyze large texts. There are relatively more cases with small tf_0 when Z exceeds the cut-off value whereas d does not than when d exceeds the cut-off value whereas Z does not, although more tests are needed to confirm this pattern.

5 Conclusion

Two term specificity measures, d and Z, show convergent validity. They are not perfectly interchangeable, however. The assumption that Z is more sensitive than d when small wordcounts are imputed in their calculation needs further testing. It remains to be seen if text length should be taken into account when choosing between the two measures indeed.

The other promising direction for further research refers to the adaptation of d and/or Z to tasks in information retrieval. Although SDS suggests that it can be done, computational complexity constitutes an obstacle. The author of this measure was able to run tests at the price of its significant simplification as a result of imputing nominal-level data instead of ratio-level (Cummins, 2013). The underlying intuition is that the calculation of d and Z can be thought of as a method of outlier detection. The larger the values of d, Z and SDS, the further from an average score a term or a document deviates. Inversely, the smaller the values of specificity measures, the closer to an average score a term or a document is. The identification of centroids is important in information retrieval. For instance, by identifying documents with small SDS values for a given dictionary (query), it is possible to retrieve the closest matches. Under this scenario, instead of focusing on documents with largest values of specificity measures, principal attention is devoted to those with smallest values. They likely contain the information the author of a query is looking for. By typing a query, the user creates a dictionary with the help of which the aboutness of searchable documents, to use Cummins's term, is measured.

Acknowledgments

The author is grateful to one of the anonymous reviewers for useful suggestions and to the editorial office for the prompt handling of the submission. The remaining errors and inaccuracies are the author's.

References

Burscher, B., Vliegenthart, R., De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The Annals of the American Academy of Political and Social Science*, 659(1), pp. 122-131.

Cummins, R. (2013). A Standard Document Score for Information Retrieval. In: Kurland, O., Metzler, D., Lioma, C., Larsen, B., Ingwersen, P. (Eds.). *ICTIR'13: Proceedings of the 2013 Conference on the Theory of Information Retrieval*, Copenhagen Denmark 29 September 2013-2 October 2013, pp. 113-116. The Association for Computing Machinery.

Diermeier, D., Godbout, J.-F., Yu, B., Kaufman, S. (2011). Language and Ideology in Congress. *British Journal of Political Science*. 42(1), pp. 31-55.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2).

DiMaggio, P., Nag, M., Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), pp. 570-606.

Dourado, Í. C., Galante, R., Gonçalves, M. A., da Silva Torres, R. (2019). Bag of Textual Graphs (BoTG): A General Graph-Based Text Representation Model, *Journal of the Association for Information Science and Technology*, 70(8), pp. 817-829.

Evans, M., McIntosh, W., Lin, J., Cates, C. (2007). Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. *Journal of Empirical Legal Studies*, 4(4), pp. 1007-1039.

Goodman, L. A., Kruskal, W. H. (1954). Measures of Association for Cross Classifications, *Journal of the American Statistical Association*, 49(268), pp. 732-764.

Harasymiw, B. (2002). *Post-Communist Ukraine*. Edmonton and Toronto: Canadian Institute of Ukrainian Studies Press.

Hogenraad, R. L., Garagozov, R. R. (2014). Textual fingerprints of risk of war. *Literary and Linguistic Computing*, 29(1), pp. 41-55.

Jiang, H., Li, W. (2012). Improved Algorithm Based on TFIDF in Text Classification. *Advanced Materials Research*, 403, pp. 1791-1794.

Juola, P., Mikros, G. K., Vinsick, S. (2019). A Comparative Assessment of the Difficulty of Authorship Attribution in Greek and in English. *Journal of the Association for Information Science and Technology*, 70(1), pp. 61-70.

Jurafsky, D., Martin, J. H. (2008). *Speech and Language Processing*. 2nd edition. Upper Saddle River, NJ: Pearson-Prentice Hall.

Khan, F. H., Qamar, U., Bashir, S. (2017). A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. *Knowledge and Information Systems*, 51(3), pp. 851-872.

Knightley, P. (2003). *The First Casualty: The War Correspondent as Hero, Propagandist, and Myth-maker from the Crimea to the Gulf War II*. London: André Deutsch.

Lasswell, H. D. (1938). *Propaganda technique in the World War*. New York: Peter Smith.

Lind, F., Eberl, J.-M., Eisele, O., Heidenreich, T., Galyga, S., Boomgaarden, H. G. (2022). Building the Bridge: Topic Modeling for Comparative Research. *Communication Methods and Measures*, 16(2), pp. 96-114.

Lukin, A. (2013). The meanings of 'war': From lexis to context. *Journal of Language and Politics*, 12(3), pp. 424-444.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mathet, Y., Widlöcher, A., Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3), pp. 437-479.

Salton, G., McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Savoy, J. (2019). Text Categorization with Style. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (Eds.). *Advances in Information Retrieval: 41st European Conference on IR Research*, ECIR 2019 Cologne, Germany, April 14–18, 2019. Proceedings, Part II, pp. 408-409. Springer.

Savoy, J. (2017). Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries. *Glottometrics*, 38, pp. 55-76.

Savoy, J. (2016). Text Representation Strategies: An Example With the State of the Union Addresses. *Journal of the Association for Information Science and Technology*, 67(8), pp. 1858-1870.

Shalak, V. I. (2004). *Kontent-analiz. Prilozhenija v oblasti politologii, psihologii, sociologii, kul'turologii, jekonomiki, reklamy* [Content analysis and its applications to political sciences, psychology, sociology, culturology, economic sciences and advertising]. Moscow: Omega-L.

Simon, A. F., Xenos, M. (2004). Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis, *Political Analysis*, 12(1), pp. 63-75.

Vellino, A., Alberts, I. (2016). Assisting the appraisal of e-mail records with automatic classification. *Records Management Journal*, 26(3), pp. 293-313.

Wang, J., Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information*, 11(9), p. 421.

Warner, R. M. (2013). *Applied Statistics: From Bivariate Through Multivariate Techniques*. 2nd edition. Thousand Oaks, CA: Sage.

Zhai, C., Massung, S. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. ACM Books.