From Diachronic to Contextual Lexical Semantic Change: Introducing Semantic Difference Keywords (SDKs) for Discourse Studies

Isabelle Gribomont
KBR (Royal Library of Belgium)
CENTAL (Centre de Traitement Automatique du Langage)
UCLouvain
isabelle.gribomont@uclouvain.be

Abstract

This paper introduces the concept of Semantic Difference Keywords (SDKs). We define SDKs as keywords selected because of a comparatively high semantic difference between their use in two or more corpora. They are extracted by applying methods developed to identify diachronic Lexical Semantic Change. Like statistical keywords, most commonly used in quantitative discourse studies, SDKs capture the distinctiveness of a target corpus. However, they do not do so because they are used significantly more often or more consistently, but because they are used significantly differently. The case study presented in this paper shows that SDKs are successful in identifying concepts which are contested, i.e., sites of "semantic struggles" (Kranert, 2020). SDKs are therefore a useful contribution to (computational) discourse studies and text-based Digital Humanities more broadly.

1 Introduction

In discourse studies, a keyword is a central concept to the comparative study of corpora. However, the identification of such keywords is most often predetermined by the researcher, or, in the case of corpus linguistics studies, by statistical measures based on frequency and/or dispersion. Scholars have wondered how to identify keywords which are sites of "semantic struggles", i.e., which are at the centre of societal controversies and whose meaning is therefore contested (Jeffries and Walker, 2017).

This paper proposes the concept of Semantic Difference Keywords (SDKs), defined as words or multi-word expressions (MWEs) whose semantic difference between two or more corpora is comparatively large. SDKs are extracted with methods developed for the study of diachronic Lexical Semantic Change (LSC). This novel application of such methods is relevant to Computational and Digital Humanities.

As a case study, the discourse from Latin American guerrilla movements from the Cuban Revolution onward was investigated. More specifically, the words whose meaning in the discourse issued by the EZLN (Zapatista Army of National Liberation) most differs from their meaning in discourses issued by the other movements in the corpus were studied by training a Word2Vec model where two different embeddings were learned for candidate SDKs: one representing their use in the EZLN corpus, and one representing their use in the rest of the corpus. This analysis highlights that this concept shows promises to identify words which are sites of contestation within a specific discourse. It also underlines that high semantic difference can be explained by a variety of factors and that their relevance is therefore dependant on the initial research question. Stylistic markers, polysemy, context-dependant lexicon and ideological differences can all lead to variance in the context where words are being used by a specific group.

2 Background and related works

2.1 Quantitative and qualitative approaches to keywords in discourse studies

In corpus-driven discourse analysis, including computational literary studies, lexical “keyness” is a ubiquitous concept. It is most often based on frequency and represents the above-chance occurrence of the term in the corpus under investigation in comparison to another. Dispersion is another keyness measure which takes the distribution of the word across the corpus into account (Du et al., 2021; Egbert and Biber, 2019; Gries, 2008, 2021; Schöch et al., 2018).

From a computational discourse analysis perspective, keywords highlight what is distinctive at the lexical level in a target corpus. Through statistical keyword analysis, researchers have studied diplomatic letters (Pranoto and Yuwono, 2017),

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 153–160

court proceedings (Potts and Kjær, 2016), academic writing (Paquot and Bestgen, 2009), gender differences in language use (Newman et al., 2008), political manifestos (Skorczynska, 2016), online COVID discourse (Joharry, 2023), and the representation of minorities or events in the press (Baker et al., 2013; Mohammed et al., 2022; Taylor, 2014).

However, keywords in (qualitative) discourse studies more globally refer to words which are central to a discourse. Schröter (2008) argues that studying the semantics and use of such expressions is key to understanding these discourses, particularly the ways in which "the meaning of the word change relative to the group that uses it". For instance, Kranert (2020) identifies "populism" as one such sociopolitical keyword in politics, news coverage and academic discourse. These keywords are sites of contestation or "semantic struggle". Their rhetorical role is therefore highly context dependent.

Previous research projects have combined quantitative and qualitative understandings of keywords. Kranert (2020) uses corpus linguistics methods to examine the pre-selected sociopolitical keyword "populism". Jeffries and Walker (2017), similarly drawing from research on cultural/sociopolitical keywords (Williams, 2014) and corpus linguistics keywords (O'Halloran, 2010; Stubbs, 2001), propose to identify keywords of interest by filtering statistical keywords, instead of focusing on a predetermined selection. To do so, they explicitly filter out statistical keywords that were "uncontested", "uncontroversial" and "least likely to actually demonstrate a change in their semantics between the two corpora".

This paper proposes a method that partially fulfills the same goals as the methodology proposed in Jeffries and Walker (2017). However, instead of filtering statistical keywords manually, relying on contextual knowledge and close readings of concordances and collocation lists, keywords are automatically extracted using NLP methods developed to recognise semantic difference.

2.2 Word embeddings and discourse studies

Because of their ability to map and formalise relationships between words within specific discourses, word embeddings are increasingly used in the field of Critical Discourse Analysis. See Wiedemann and Fedtke (2022) for a relevant survey of the topic. Such studies usually focus on one corpus (see, for instance Mandenaki et al. (2022) and Durrheim et al. (2023)). When the study is comparative, it usually investigates diachronic discourse change. For instance, Rodman (2020) tracked the changing meanings of political concepts in a dataset of 161 years of newspaper coverage and Viola and Verheul (2020) studied the evolution of the concept of migration in *The Times Archive* from 1900 to 2000.

Comparative synchronic semantic change analyses in discourse studies are rare. However, Schlechtweg et al. (2019) argue for the relevance of LSC for synchronic studies and apply it to detect sense divergence in domain-specific corpora (see also (Ferrari et al., 2017)). In addition, Gruppi et al. (2023) utilize semantic shift as an indicator of agreement among synchronic sources in the context of a method for news veracity classification. Yehezkel Lubin et al. (2019) use the concept of top changing words between synchronic corpora in the context of the alignment of vector spaces with noisy supervised lexicons, and Yin et al. (2018) investigate domain-specific linguistic shifts using word embeddings, also in the context of the development of a new vector space alignment method.

In the context of discourse studies specifically, a notable contribution is Dénigot and Burnett (2021), who use word embeddings to compare the discourses of the supporters and detractors of the legalisation of same-sex marriage at the French Assemblée Nationale in 2013. They conclude by arguing that embeddings have potential for the comparative analysis of synchronic corpora. The concept of SDK is part of the same impulse to expand the exploitation of word embeddings for discourse studies to synchronic investigation.

3 Defining Semantic Difference Keywords

SDKs are terms whose meaning differs substantially between two corpora. Otherwise stated, they are the words for which the semantic difference between their manifestation in one corpus and another is largest. In analogy with frequency or dispersion-based keywords, SDKs capture the distinctiveness of a target corpus, not because they are used significantly more often or more consistently, but because they are used significantly differently. Like frequency and dispersion-based keywords, they highlight the locations where the language of the two corpora differs at the lexical level. However, words which have similar relative frequencies and dispersions in two corpora, by definition, will not be

selected as key, even if they are used in largely different ways, and therefore hold clues to fundamental differences in language use between the two corpora (Dénigot and Burnett, 2021).

In corpus linguistics, collocations are used to contrast how a word has different meanings or connotations in different contexts, but the words whose collocations are investigated are selected either because of underlying research questions, or according to frequency or statistical keyness criteria. In addition, they do not allow to measure how stable the meaning of the word under investigation is between the two corpora.

The concept of SDK is therefore useful to automatically extract words or phrases whose meaning is most unstable across two or more corpora. Not only can they contribute to identifying terms around which sociopolitical debates take place, but, like quantitative keywords, they could be leveraged for literary analysis, stylistic studies, authorship attribution, etc.

4 Case Study

As a case study, the discourse of Latin American leftist armed movements from the Cuban Revolution onward is investigated. The language of Latin American guerrilla discourse is relatively repetitive and heavily relies on fixed expressions and clichés. However, it has been argued that the EZLN (Zapatista Army of National Liberation), active from 1994 until today, offered a renovation of revolutionary leftist language in Latin America (Marcos and Le Bot, 1997; Gribomont, 2019). Identifying SDKs by comparing the EZLN corpus and a comparison corpus of texts written by other Latin American guerrilla movements from 1953 onward contributes to assess the ways in which this renovation takes place.

4.1 Data

The corpus was assembled by scraping the CeDeMa archive (Centro de Documentacion de los movimientos armados),[1] documents issued by the 26th of July Movement (the leading organisation of the Cuban Revolution),[2] and the archive of the Zapatista Army of National Liberation (EZLN).[3] The corpus totals more than 26 million Spanish words, of which more than 4 millions belong to the EZLN corpus. As part of pre-processing, the corpus was lower-cased, lemmatised and segmented into sentences.

4.2 Method

In theory, all methods developed to identify semantic change can be adapted to extract such sites of "semantic struggle". For a general survey of computational approaches to lexical semantic change, see Tahmasebi et al. (2021). See Kutuzov et al. (2018) for a survey focused on word embeddings.

The approach selected for this experiment relies on static word embeddings. With this method, a Word2Vec model (Mikolov et al., 2013) is trained with the whole data, but we append a context specific string to target words, i.e., words which are pre-selected as potential SDKs. This method is equivalent to the Temporal Referencing method described in Dubossarsky et al. (2019), where time-specific tokens are added to target words to model LSC. As demonstrated in the paper, this method has advantages in comparison to other embedding-based methods which learn a different semantic vector space for each time period before aligning them so as to minimise the distances between the time-specific embeddings of the same word (Hamilton et al., 2016). In addition, it shows that Temporal Referencing leads to models which are less noisy in comparison to alignment-based embeddings methods (Levy et al., 2015). Finally, it is more likely to perform well with smaller corpora since the words which are not selected for referencing are learned once, thereby minimizing the robustness issues caused by low frequency word embeddings (Dubossarsky et al., 2019). However, it does not account for the potential semantic difference between different contextual uses of the context words, which likely introduces biases into the semantic space.

To select the potential SDKs, the corpus was compared to the general Web Spanish corpus esTenTen18 available in Sketch Engine which contains 16.9 billion words of both European and American Spanish (Kilgarriff and Renau, 2013; Kilgarriff et al., 2014).[4] The words which obtained a simple maths keyness score higher than 1 (Kilgarriff, 2009) and whose frequency was greater than 400 in the EZLN corpus and 1000 in the rest of corpus were selected, resulting in 151 words.

Instead of the time period, the context is ref-

[1] https://cedema.org/digital_items
[2] http://www.fidelcastro.cu/es/biblioteca/documentos/
[3] https://enlacezapatista.ezln.org.mx/
[4] http://www.sketchengine.eu

erenced. In this case, the string *_EZLN* was appended to the pre-selected words so that different embeddings are learned for their manifestation in the EZLN corpus and the comparison corpus. The cosine similarity between vector pairs is calculated for all potential SDKs. They are then ranked from smallest to highest similarity.[5]

4.3 Results

Table 1 shows the top ten SDKs, i.e., the ten words for which the cosine similarity between the embedding pairs is the smallest. A full analysis is beyond the scope of this paper. However, the first word, "revolution" is particularly interesting. Its ten nearest neighbours include "independence", "1910", "1810" and "PRI". PRI is the acronym of the Mexican Institutional Revolutionary Party, a right-wing party which has been in power from 1929 to 2000. This party co-opted the imagery of the 1910 Mexican Revolution, which included a peasant rebellion against unjust agrarian laws, thereby "institutionalising" the concept of revolution. In doing so, they have rendered the word unusable for the EZLN. By naming their party "revolutionary", the PRI essentially altered the meaning of the word "revolution" for a segment of the Mexican population. Within the Mexican context, "revolution" is a site of semantic struggle at the centre of societal conflicts.

The second word, "class", reveals the EZLN's departure from the dominant language and ideology of Latin American guerrilla discourse. It is most commonly used in the context of "class struggle", "class conscience" and "working class" in the reference corpus, but used mostly to refer to the "political class" in the EZLN discourse. The approach proposed by the EZLN is intersectional and the redefinition of the word class is part of an abandonment of stereotypical Marxist vocabulary, symptomatic of a detachment from past guerrilla movements.

The third word, "citizen", is used to address "citizen rights", "citizen security" and "innocent citizens" in the reference corpus. In the EZLN corpus, it refers to "citizen initiatives", "citizen organizations" and "citizen movements". This semantic shift reflects the discrepancy in the perceived role of citizens in the social struggles and, more specifically, the key role of civil society in the Zapatista movement. In effect, the word "citizen" means something different in the two discourses, but semantics reflects a diverging ideology and mode of action.

Word	**Translation**	**cosine sim.**
revolución	revolution	0.2183
clase	class	0.2239
ciudadano	citizen	0.2299
plan	plan	0.2304
comandante	commandant	0.2351
frente	front	0.2462
terreno	piece of land/field	0.2540
dirección	direction/address/ management	0.2600
pensamiento	thought	0.2641
principio	principle/beginning	0.2657

Table 1: Top ten SDKs for the EZLN corpus and the comparison corpus.

It is also interesting to note that several words from this list are polysemic. It is the case of *principio*, for instance, which is used most often in the sense of "values" or "norms" in the reference corpus and in the sense of "beginning" in the EZLN corpus. This difference is again symptomatic of the Zapatistas' rhetoric, which is based on revolutionary practices more than revolutionary principles, but it also reflects the more narrative and oral writing styles adopted by Zapatista representatives.

The words whose meaning is the least different in the two corpora (negative SDKs) reveal the area where there is a strong continuity between the EZLN language and other movements. "Hand", "blood", "land", "money" and "wealth" are the top five negative keywords. For each of these, the nearest neighbour for the EZLN vector is the corresponding vector in the comparison corpus. Conversely, the EZLN vector is in the top five nearest neighbours of the comparison corpus vectors.

5 Discussion and future work

The method is successful in pointing to words which are used differently in different corpora. However, for the sake of illustrating the concept of SDKs, this pilot study relied on a single model and did not address potential robustness issues linked to the variability of word embeddings. For a fully-fledged analysis of the corpus, additional steps will be undertaken to increase reliability. First, im-

[5]The content of the scraped websites/archives is publicly available, but cannot be reproduced elsewhere. However, the code, frequency files and link to the resulting Word2Vec embeddings are available on GitHub: https://github.com/isag91/Semantic-Difference-Keywords.

plementing recommendations proposed by scholars who investigated the instability of Word2Vec models (Zhou et al., 2020; Pierrejean and Tanguy, 2018), especially those learned from comparatively small amount of data, will contribute to mitigate this issue. Antoniak and Mimno (2018) demonstrated that word embeddings are sensitive to small variations in the source documents, including their position in the corpora, suggesting that they are not trustworthy to study word associations. They recommend to average distance calculations over multiple bootstrap samples instead of relying on a single model. In addition, finetuning an existing model trained on a larger corpus instead of training a new model from scratch has proven to be a useful measure (Howard and Ruder, 2018).

Second, the influence of the algorithm, hyperparameters, word frequencies and length difference between sub-corpora on the results should be investigated. To truly assess performance and validate results, the creation of ground-truth datasets for such tasks would be valuable, whether via the annotation or existing data or the creation of simulated data (Hengchen et al., 2021). See Rodman (2020) for details on the creation of a gold standard for the evolution of meanings of political concepts.

As mentioned above, this pilot study aimed at illustrating the concept of SDKs. However, by limiting the contextual referencing to the EZLN corpus, the power of the methodology is limited. In future works, potential SDKs will be referenced for all movements (frequency permitting) and divided into three periods informed by historical research of Latin American leftist guerrilla movements (Wickham-Crowley, 2014). Some movements have been active for several decades and significantly evolved over time. This more granular referencing will be used to identify ideological clusters as well as patterns of continuity and rupture in the discourse of insurgency in Latin America (Chasteen, 1993). From a methodological viewpoint, by calculating all pairwise semantic similarities for potential SDKs, we will be able to extract keywords which are most susceptible to semantic variability across the board, rather than focusing on one movement. In addition, when focusing on one movement, it will be interesting to look at words whose pairwise distances is abnormally large in comparison to the pairwise distances involving the other movements.

Beyond this specific adaptation of LSC metholologies, relying on contextualised instead of static embeddings to investigate semantic difference (see Wiedemann and Fedtke (2022); Montanelli and Periti (2023)) would be productive for this area of research, since it would allow for the assessment of the stability of word meaning within one (sub-)corpus as well as across different corpora. For instance, examining the variance within different sub-corpora would be useful to track patterns of influence and cross-fertilisation between different social groups.

6 Conclusions

This paper introduced the concept of SDKs, i.e., keywords or key terms which are used most distinctively between two or more corpora. This concept is useful for the field of discourse studies, where researchers are interested in the ways in which terms are leveraged for differing rhetorical purposes by different groups. The extraction of SDKs bypasses the need for a predetermined shortlist of keywords. Nevertheless, the reason behind semantic difference cannot be assumed and close reading is necessary to interpret results.

In addition, researchers in the humanities and social sciences have to be wary of the potential instability or word embeddings (Sommerauer and Fokkens, 2019). Implementing recommended mitigating measures and reporting on variability metrics is key (Antoniak and Mimno, 2018). Ultimately, for this avenue of research to grow, the creation of more ground-truth datasets would be helpful.

Finally, like Dénigot and Burnett (2021), this paper wishes to argue that methods developed for the identification of LSC can productively be used for synchronic semantic difference in discourse studies as they have unique capabilities to extract language patterns which would be difficult to decipher with other quantitative or qualitative discourse studies methods.

Acknowledgements

This work was supported by a FED-tWIN grant (Prf-2020-026_KBR-DLL) funded by BELSPO (Belgian Science Policy). I thank my colleagues at CENTAL for the helful discussions and the three anonymous reviewers for their feedback and suggestions for improvements.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Paul Baker, Costas Gabrielatos, and Tony McEnery. 2013. Sketching muslims: A corpus driven analysis of representations around the word 'muslim'in the british press 1998–2009. *Applied linguistics*, 34(3):255–278.

John Charles Chasteen. 1993. Fighting words: the discourse of insurgency in latin american history. *Latin American Research Review*, 28(3):83–111.

Quentin Dénigot and Heather Burnett. 2021. Using word embeddings to uncover discourses. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 298–312.

Keli Du, Julia Dudar, Cora Rok, and Christof Schöch. 2021. Zeta & eta: An exploration and evaluation of two dispersion-based measures of distinctiveness. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Kevin Durrheim, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2023. Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1):617–629.

Jesse Egbert and Doug Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104.

Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting domain-specific ambiguities: An nlp approach based on wikipedia crawling and word embeddings. *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 393–399.

Isabelle Gribomont. 2019. The zapatista linguistic revolution: A corpus-assisted analysis. *Discourses from Latin America and the Caribbean: Current Concepts and Challenges*, pages 139–171.

Stefan Th Gries. 2008. Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*, 13(4):403–437.

Stefan Th Gries. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2):1–33.

Maurício Gruppi, Panayiotis Smeros, Sibel Adalı, Carlos Castillo, and Karl Aberer. 2023. Scilander: Mapping the scientific news landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 269–280.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. *Computational approaches to semantic change*, 6:341.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Lesley Jeffries and Brian Walker. 2017. *Keywords in the press: The New Labour years*. Bloomsbury Publishing.

Siti Aeisha Joharry. 2023. Faith in the time of coronavirus: A corpus-assisted discourse analysis. *Intellectual Discourse*, 31(1):211–232.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, volume 6.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Adam Kilgarriff and Irene Renau. 2013. estenten, a vast web corpus of peninsular and american spanish. *Procedia-Social and Behavioral Sciences*, 95:12–19.

Michael Kranert. 2020. When populists call populists populists:'populism'and 'populist'as political keywords in german and british political discourse. *Discursive Approaches to Populism Across Disciplines: The Return of Populists and the People*, pages 31–60.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Katerina Mandenaki, Catherine Sotirakou, Constantinos Mourlas, and Spiros Moschonas. 2022. Topic models and word embeddings for ideological analysis: A case study in neoliberal discourse. *The International Journal of Communication and Linguistic Studies*, 21(1):37.

Subcomandante Marcos and Yvon Le Bot. 1997. El sueño zapatista. *Entrevistas con el Subcomandante Marcos, el*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Tawffeek AS Mohammed, Felix Banda, and Mahmoud Patel. 2022. The topoi of mandela's death in the arabic speaking media: A corpus-based political discourse analysis. *Frontiers in Communication*, 7:849748.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *arXiv preprint arXiv:2304.01666*.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236.

Kieran O'Halloran. 2010. How to use corpus linguistics in the study of media discourse. In *The Routledge handbook of corpus linguistics*, pages 563–577. Routledge.

Magali Paquot and Yves Bestgen. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In *Corpora: Pragmatics and discourse*, pages 247–269. Brill.

Bénédicte Pierrejean and Ludovic Tanguy. 2018. Predicting word embeddings variability. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 154–159.

Amanda Potts and Anne Lise Kjær. 2016. Constructing achievement in the international criminal tribunal for the former yugoslavia (icty): A corpus-based critical discourse analysis. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 29:525–555.

Budi Eko Pranoto and Untung Yuwono. 2017. Leader's attitude towards terrorism: A critical discourse analysis of dr. mahathir mohamad's diplomatic letters. In *Cultural dynamics in a globalized world*, pages 65–73. Routledge.

Emma Rodman. 2020. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1):87–111.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Christof Schöch, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho. 2018. Burrows' zeta: Exploring and evaluating variants and parameters. In *DH*, pages 274–277.

Melani Schröter. 2008. Discourse in a nutshell: Key words in public discourse and lexicography. *German as a foreign language*, (2):42–57.

Hanna Skorczynska. 2016. The emerging parties' manifestos for the 2015 spanish general elections: a comparative analysis of lexical choices. *EPiC Series in Language and Linguistics*, 1:372–380.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

Michael Stubbs. 2001. *Words and phrases: Corpus studies of lexical semantics*. John Wiley & Sons.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).

Charlotte Taylor. 2014. Investigating the representation of migrants in the uk and italian press: A cross-linguistic corpus-assisted discourse analysis. *International journal of corpus linguistics*, 19(3):368–400.

Lorella Viola and Jaap Verheul. 2020. One hundred years of migration discourse in the times: A discourse-historical word vector space approach to the construction of meaning. *Frontiers in Artificial Intelligence*, page 64.

Timothy Wickham-Crowley. 2014. Two "waves" of guerrilla-movement organizing in latin america, 1956–1990. *Comparative Studies in Society and History*, 56(1):215–242.

Gregor Wiedemann and Cornelia Fedtke. 2022. From frequency counts to contextualized word embeddings: The saussurean turn in automatic content analysis. In *Handbook of Computational Social Science, Volume 2*. Taylor & Francis.

Raymond Williams. 2014. *Keywords: A vocabulary of culture and society*. Oxford University Press.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. 2019. Aligning vector-spaces with noisy supervised lexicon. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 460–465, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. *Advances in neural information processing systems*, 31.

Xuhui Zhou, Zaixiang Zheng, and Shujian Huang. 2020. Rpd: a distance function between word embeddings. *arXiv preprint arXiv:2005.08113*.