Political dogwhistles and community divergence in semantic change

Max Boholm* and **Asad Sayeed**[†]
*Gothenburg Research Institute (GRI), [†]Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{max.boholm, asad.sayeed}@gu.se

Abstract

We test whether the development of political dogwhistles can be observed using language change measures; specifically, does the development of a "hidden" message in a dogwhistle show up as differences in semantic change between communities over time? We take Swedish-language dogwhistles related to the on-going immigration debate and measure differences over time in their rate of semantic change between two Swedish-language community forums, *Flashback* and *Familjeliv*, the former representing an in-group for understanding the "hidden" meaning of the dogwhistles. We find that multiple measures are sensitive enough to detect differences over time, in that the meaning changes in *Flashback* over the relevant time period but not in *Familjeliv*. We also examine the sensitivity of multiple modeling approaches to semantic change in the matter of community divergence.

1 Introduction

As a type of manipulative communication, a political dogwhistle is a message with a controversial (or extreme) in-group meaning that is hidden to most of the public and only apprehended by a limited proportion of its audience, but at the same time communicates a less controversial (less extreme) out-group meaning to the wider audience who does not grasp the in-group meaning of the message (Haney-López, 2014; Stanley, 2015). An example is "inner city", which has a general meaning of "central section of a city" but has also been used with concealed derogatory racial reference to an area with a poor, African American population (Saul, 2018). Dogwhistles enable attracting some part of its audience who are appealed to by the extreme view, while at the same time not offending others (who do not get the hidden message). With concealed meanings, communicators can avoid accountability for expressing and approving of controversial views. Therefore dogwhistle communication can pose problems for representative democracy (Goodin and Saward, 2005; Stanley, 2015) and speech moderation online (Gavidia et al., 2022; Schmidt and Wiegand, 2017; Zhu and Bhat, 2021).[1]

By design, in-group meanings of dogwhistles evolve in parallel to existing out-group interpretations. Therefore semantic change is essential to the concept of the dogwhistle. However, little systematic attention has, in fact, been devoted to semantic change in dogwhistle expressions. This paper sets out to study this under-explored temporal dimension of dogwhistles through techniques from Natural Language Processing (NLP) to detect lexical semantic change (LSC). More precisely, the aim of this paper is to explore the role of community in the semantic change of set of known-to-be Swedish dogwhistle expressions (DWEs), identified in other work (Åkerlund, 2022; Hertzberg, 2022; Lindgren et al., 2023), including *kulturberika* (culture enrich) and *globalist* (described in more detail below).

In this work, we address the role of community in semantic change by studying the semantic change of DWEs in two online communities (Åkerlund, 2022; Bhat and Klein, 2020): *Flashback*, which is a discussion forum that is known for hosting controversial topics of discussion and for expression of controversial societal opinions (Åkerlund, 2021; Blomberg and Stier, 2019; Malmqvist, 2015); and *Familjeliv* ("family life" in English), which is a discussion forum that is expected to be very different from *Flashback*, with its focus on topics of parenting and family life, but also include discussions on politics and society (Hanell and Salö, 2017). We test *the isolated change of DWEs hypothesis*, i.e., that meaning change of dog-

[1] In democracies, political leaders get a mandate to govern through general elections. They get (re-)elected or replaced by their official proposals for collective action and policies. Dogwhistles obscure this legitimacy of the political mandate given by elections, since the promises are not what they seem to be.

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 53–65
December 6, 2023 ©2023 Association for Computational Linguistics

whistles is *community-dependent*. Here, this expectation is more precisely tested under the following formulation:

H1: The degree of semantic change of (selected) DWEs observed in the (highly politically polarized) online community *Flashback* is different from the degree of semantic change of the same terms (at the same period of time) in the (less polarized) community *Familjeliv*.

In recent years, several different approaches have been developed for modeling of LSC (Kutuzov et al., 2018; Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021; Tang, 2018). For a robust testing of H1, we test and compare results modeled by three different approaches: (1) the **SGNS** approach, which uses word embeddings built through a skip-gram with negative sampling (SGNS) model (Mikolov et al., 2013); (2) the **SBERT-PRT** approach which averages over contextual token embeddings from Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), so called "prototypes" (PRT) (Kutuzov and Giulianelli, 2020; Martinc et al., 2020a); and (3) the **SBERT-CLT** approach which, like the previous approach, uses contextual embeddings from SBERT, but instead of averaging, clusters token embeddings and compare distribution over clusters over time. We test H1 with respect to all three approaches (described in more detail below).

2 Related work

2.1 The meaning of dogwhistles

Quaranto (2022) argues for the importance of linguistic practices in understanding dogwhistles. Essential to this account is the notion of community, since linguistic practices are defined in relation to some community who uphold the practice. At some level of analysis, the speech act of dog whistling *depends on specific lexical forms* embedded in particular linguistic practices (Henderson and McCready, 2018; Quaranto, 2022). While every usage of such DWEs does not perform a dogwhistle speech act – additional criteria are involved in performing the act of dogwhistling (Quaranto, 2022; Saul, 2018) – specific linguistic forms are necessary for conveying the in-group meaning.[2] As such, the link between DWEs and their in-group meanings are upheld by linguistic communities. Dogwhistle meanings in general and the meaning change of dogwhistles in particular are expected to be *community-dependent*. A stronger claim is that the semantic changes of DWEs observed in one community is unlikely to be observed in another community. Here, this expectation is discussed as *the isolated change of DWEs hypothesis*, which is more precisely tested under the formulation in H1. Note that the isolated change of DWEs hypothesis is a special case of a more general thesis that any lexical meaning and therefore also LSC more generally depends on the linguistic communities in which words are used (Clark, 1996).

2.2 Lexical semantic change detection

In accordance with the distributional hypothesis (Firth, 1957; Harris, 1954; Sahlgren, 2008), existing computational methods to analyze LSC apply unsupervised techniques to build numerical vector representations of words at different periods of time and then compare those vectors to determine how much, when and in what way words change (Tahmasebi et al., 2021). For the first two questions (how much and when), the semantic change of a word *w* in a transition from t_i to t_j, $\Delta_{t_i,t_j}(w)$, is the distance of *w*'s vector at t_i ($\overrightarrow{w}_{t_i}$) and its vector at t_j ($\overrightarrow{w}_{t_j}$):

$$\Delta_{t_i,t_j}(w) = distance(\overrightarrow{w}_{t_i}, \overrightarrow{w}_{t_i})$$

Both *static* word embeddings, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), and *contextualized* word embeddings, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have been used to vectorize words in LSC. With static word embeddings, *w*'s meaning is represented by *one* vector that generalizes over its usages. There are two common measures of the distance of static word embeddings: cosine distance (Hamilton et al., 2016) and angular distance (Kim et al., 2014). With contextualized word embeddings, the procedure for word representations over time is somewhat more elaborate than for static embeddings (Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020; Martinc et al., 2020a; Vani et al., 2020). First, contextual word embeddings, such as BERT and ELMo, are multi-layered, multidimensional representations that for every token have a *L* × *N* vector representations, where *L* is the number of layers and *N* is the number of dimensions. Selecting the top layer or averaging over (top) layers

[2] This might be too strong a claim, since symbols other than words have been claimed to function as dogwhistles, as exemplified by the Willie Horton campaign (Mendelberg, 1997).

is usually applied when comparing vectors over time. Second, with contextualized embeddings, there is no single representation of w at each time period to be compared. Rather, a word is associated with sets of token vectors at t_i and t_j. In order to arrive at a single measure of change of a word in transition from t_i to t_j, there are two main solutions. In a *prototype approach* the distance between the average token vectors at t_i and t_j is measured by cosine distance or angular distance. These average token vectors are referred to as "prototypes" in previous work. In a *clustering approach* token vectors in t_i and t_j are clustered and then the distance of the distributions of clusters are compared by some measure for comparing probability distributions, for example, Jensen-Shannon distance (Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020; Martinc et al., 2020b; Vani et al., 2020).

Comparisons of methods for LSC detection show mixed findings. The best performing models of SemEval-2020 shared task on unsupervised LSC detection used static word embeddings (Schlechtweg et al., 2020). However, reported findings include contextualized approaches outperforming static embeddings (Kutuzov and Giulianelli, 2020); clustering of contextual embeddings performing worse than approaches that average contextual embeddings (Laicher et al., 2021) and approaches with static embeddings (Martinc et al., 2020b); and clustering contextualized embeddings performing better than averaging over them (Martinc et al., 2020a). Moreover, performance is often different for different languages (Kutuzov and Giulianelli, 2020; Martinc et al., 2020b; Vani et al., 2020). Performance on Swedish data is sometimes found to be worse than, for example, English and German (Laicher et al., 2021; Martinc et al., 2020b), sometimes better (Vani et al., 2020).

3 Data

3.1 Data sets

Two online communities are explored here: *Flashback* and *Familjeliv*. As mentioned above, *Flashback* is a discussion forum on a wide range of topics organized in "threads" under 15 general sections (e.g., drugs, economy, lifestyle and politics). As of 3 August, 2023, the website claims to have over 1.5 million members and almost 80 million posts. *Flashback* support anonymity of users, which enables discussion of controversial topics and expression of controversial opinions, including discrimination and racism (Åkerlund, 2021; Blomberg and Stier, 2019; Malmqvist, 2015). While threats and hate speech are not allowed by the rules of *Flashback*, the website clearly contains offensive language. In a recent survey from 2021, 26% of male and 21% of female social media users in Sweden reported using *Flashback* within the last 12 months (Internetstiftelsen, 2021).

The discussion forum *Familjeliv* is organized in threads of 20 general categories (with several subtopics), where most topics focus on family and parenting (e.g., adoption, pregnancy, and pets), but also include topics of society, economy and law. In 2014, *Familjeliv* had about 700 000 visitors every week (Hanell and Salö, 2017). The forum is explicitly claimed to be a meeting place for women (Hanell and Salö, 2017), which is confirmed by survey data from 2021: 4% of male and 8% of female social media users in Sweden reported using *Familjeliv* within the last 12 months (Internetstiftelsen, 2021).

The corpora we use are collected from the Swedish national language data processing infrastructure Språkbanken Text.[3] The *Flashback* data hosted by them range from 2000 to 2022. In total, *Flashback* data contain 49M sentences (posts) and 785M words. On average, there are 2.1M sentences (*SD* = 1.4M) and 34.1M words (*SD* = 21.7M) per year. The *Familjeliv* data range from 2003 to 2022 and contain 19M sentences (*M* = 0.9M, *SD* = 0.9M) and 305M words (*M* = 15.2M, *SD* = 14.3M).

3.2 A selection of Swedish dogwhistle expressions

A sample of known-to-be Swedish DWEs are investigated (Åkerlund, 2022; Hertzberg, 2022; Lindgren et al., 2023), henceforth referred to as *S-DWE*:

(*S-DWE*) *berika* (enrich, verb), *kulturberika* (culture enrich, verb), *kulturberikare* (culture enricher, noun), *globalist* (globalist, noun), *återvandra* (re-migrate, verb), *återvandring* (re-migration, noun), and *hälpa på plats* (help at site, verb phrase).

This set is identified through exploration of frequent morphological variation of a set of "base forms" in corpus data, resulting in adjectives, nouns and verbs: "återvandr" (as in the verb *återvandra* 're-migration'), "(culture) berika" ([culture] enrich), "globalist" (globalist) and "hjälpa på plats" (help at site). With the exception of the VP *hjälpa*

[3] See: https://spraakbanken.gu.se/en

på plats, which is here explored as a fixed phrase (ignoring inflectional variation), *S-DWE* is a set of lexemes, i.e., abstractions over inflectional forms.

The in-group meanings of the terms in *S-DWE* can be listed at a general level, related to their base forms (Lindgren et al., 2023). This discussion ignores the systematic meaning variation resulting from morphological modifications of the base forms, for example, *kulturberika* (process) → *kulturberikare* (agent of that process). The terms related to re-migration are assumed to have in-group and out-group meanings based on the (in)voluntariness of the process, with a voluntary act as the out-group meaning, while 'deportation' is the in-group meaning. The DWE of *berika* (and its related terms) is a result of malevolent irony, in response to the positive opinions about multiculturalism. The in-group meaning of *berika* (and its related terms) is the opposite of enrichment (i.e. the out-group meaning), namely criminal and destructive activities (by immigrants). In a Swedish context and elsewhere, *globalist* (and related DWEs) is used with several different in-group meanings, including an anti-Semitic reference to Jews and a nationalistic reference to anti-nationalists (i.e., opponents of nationalism). Finally, *hjälpa på plats* (help at site) has as its in-group meaning non-acceptance of refugees coming to Sweden.

Below we present examples of the words *berika* and *återvandring* in context. The examples are selected from years of transitions where the terms exemplified have a higher rate of semantic change in *Flashback* than *Familjeliv*; i.e., transitions where there is a divergence of semantic change of the (potential) DWE in the two corpora. Examples are taken from the top five sentences that are most similar to the the average vector of the SBERT-PRT approach, as defined in detail below, where the similarity of the average vector and sentence representations has been measured by cosine similarity.

1. "jag tycker att relationen till min sambos ursprung **berikar** mig enormt!" (*Familjeliv*, 2004)
   (I think that the relationship to my partner's origin enriches me enormously!)

2. "olikheter **berikar** också" (*Familjeliv*, 2005)
   (differences enrich also)

3. "det har ju bildat en hel politisk / facklig rörelse uttryckligen med syftet att ta ifrån andra och **berika** sig själva" (*Flashback*, 2004)
   (It has made a whole political / trade-union movement explicitly with the objective to take from others and enrich themselves)

4. "dessutom kan det ju vara så att detta inte är första gången någon **berikare berikar** en infödd" (*Flashback*, 2005)
   (In addition, it can be the case that this not is the first time that some enricher enriches a native)

5. "i dessa fall, och det är många , så är jag övertygad att det samhällsekonomiskt är bäst att satsa på **återvandring**" (*Familjeliv*, 2021)
   (In these cases, and those are many, I am convinced that it is socioeconomically best to go for re-migration)

6. "jag har skrivit det förr i en annan tråd: inom tio år är det '**återvandring**' som är modeordet nummer ett inom svensk politik ." (*Familjeliv*, 2022)
   (I have written that before in another thread: within ten years it is 're-migration' that is the number one buzzword in Swedish politics)

7. "det viktigaste är att vi får **återvandring**, inte hur politiker motiverar det imho" (*Flashback*, 2021)
   (The most important is that we get re-migration, not how politicians motivates it IMHO [i.e. English loan of In My Humble Opinion])

8. "sd talar om frivillig **återvandring**, men det som behövs är forcerad *återvandring*" (*Flashback*, 2022)
   (SD [i.e., the Sweden Democrats] speaks of voluntary re-migration, but what is needed is forced re-migration)

While not sufficient for systematic analysis, these examples still illustrate potential shifts in meaning in *Flashback*, but not in *Familjeliv*. We interpret example 4 as a case of the malevolent irony characteristic of the in-group meaning of enrichment dogwhistles but not present in examples 1-3. Moreover, in example 8, re-migration is associated with (in)voluntariness, where the author argues for the need of deportation. This (in)voluntariness is not present in examples 5-7.

3.3 Frequency distributions

Three observations of the frequency distributions of the terms in *S-DWE* in the present data need

	Flashback			*Familjeliv*		
DWE	Total	*M*	*SD*	Total	*M*	*SD*
berika	20936	27.92	12.18	2047	8.02	2.94
globalist	31156	32.07	39.62	122	1.77	3.15
hjälpa på plats	1150	1.14	1.50	453	1.99	2.88
kulturberika	2445	2.88	2.75	101	0.21	0.38
kulturberikare	6133	9.88	8.41	202	0.42	0.58
återvandra	1449	1.51	1.84	66	0.12	0.25
återvandring	12999	13.19	22.20	384	3.27	5.73

Table 1: Total frequency and mean frequency per million per year

mentioning (Table 1). First, compared with each other they are very different in frequency. Second, their frequencies are very different in different years, reflected by high standard deviations. Third, the terms are more common in the *Flashback* data than in the *Familjeliv* data.

For semantic change of words in general, previous work has observed a correlation with word frequency (Hamilton et al., 2016). Also in the present data there are correlations of LSC and word frequency (see Appendix A). However, three comments can be made in this regard. First, LSC and frequency are not (significantly) related for all terms in *S-DWE*. Second, correlation measures are not consistent over the three approaches here explored to model semantic change (see next section for details). For example, for SBERT-CLT, there is only significant correlation between LSC and word frequency for one of the terms in *S-DWE*. Third, as expected, with the rectified measure of change to control for noise (defined below), fewer terms in *S-DWE* show a significant correlation of frequency and semantic change rates (Noble et al., 2021; Dubossarsky et al., 2017). So although frequency is a factor for LSC modelled here, these points suggest that our findings on semantic change of DWEs are not solely due to word frequency and corpus sizes. See Noble et al. (2021) for other factors than word frequency that can drive semantic change in online communities.

3.4 Preprocessing

Data for all experiments (SGNS, SBERT-PRT and SBERT-CLT) have been preprocessed by lowercasing and removing URLs and emojis. Data for the SGNS approach has been further processed by removal of numbers and punctuation; separation of compounds that have a term in *S-DWE* as its left-hand element, for example, "globalistelit" is replaced by "globalist elit" (with space); and lemmatization of terms in *S-DWE*, for example, "globalisten" (definite form of *globalist*) is replaced by "globalist" (lemma form). Regular expressions were used for lemmatization and separation of compounds. For the SBERT approaches, there is no additional step of preprocessing to the general steps listed above. However, the analysis still implements generalizations similar to those of lemmatization by pairing every sentence with with its "lexemes" in *S-DWE*, thereby generalizing over inflection and compounding. Again, regular expressions were used for this.[4]

4 Semantic change modeling

4.1 The SGNS approach

A corpus is a collection of sentences. Let C be a diachronic corpus that covers the ordered set T of consecutive time periods t_1, ... t_n. C consists of an ordered set of temporally defined sub-corpora c_{t_1} , ... c_{t_n}. In the present experiments, $T = \langle 2000, \ldots, 2022 \rangle$. Consequently $C = \langle c_{2000}, \ldots, c_{2022} \rangle$. A SGNS model is trained for each sub-corpus in C, in the sorted order of T, from first to last. The vocabulary is restricted by a minimum frequency of 10. The weights of the model for the first time period, M_{2000}, are randomly initialized. For every other model, M_{t_i}, where $t_i > 2000$, the weights of M_{t_i} are initialized with the trained weights of $M_{t_{i-1}}$. For every consecutive pair in T, i.e. the set of transitions $R = \langle\langle t_1, t_2\rangle, \ldots \langle t_{n\text{-}1}, t_n\rangle\rangle = \langle\langle 2000, 2001\rangle, \ldots \langle 2021, 2022\rangle\rangle$, and for every word w existing in both models M_{t_i} and $M_{t_{i+1}}$ the vectors $\overrightarrow{w_{t_i}}$ and $\overrightarrow{w}_{t_{i+1}}$ are compared for two measures: (i) naive cosine change, and (ii) rectified

[4] Code for running experiments can be found at https://github.com/mboholm/dogwhistle-community-divergence.

change.

Naive cosine change for a word w in transition from t_i to t_j, i.e. $\Delta_{t_i,t_j}(w)$, is defined as the angular distance between $\overrightarrow{w}_{t_i}$ and $\overrightarrow{w}_{t_j}$ (Kim et al., 2014; Noble et al., 2021):

$$\Delta_{t_i,t_j}(w) = \frac{\arccos(cossim(\overrightarrow{w_{t_i}}, \overrightarrow{w_{t_j}}))}{\pi}$$

As argued by Dubossarsky et al. (2017), vectors of the same word w derived from different samples are expected to be different. Therefore when studying meaning change this general variation expected for w's vectors from different samples should be controlled for (Dubossarsky et al., 2017). To do so, we use a measure of *rectified change* (Noble et al., 2021). For another approach, see Liu et al. (2021). To measure rectified change we perform $n_Q = 10$ controls for every transition $\langle t_i, t_{i+1}\rangle$ (in R) such that: (1) c_{t_i} and $c_{t_{i+1}}$ are concatenated and then the combined list is shuffled; call this list of (shuffled) sentences $Q^{t_i,t_{i+1}}$. (2) $Q^{t_i,t_{i+1}}$ is split in half, resulting in subsets q_1 and q_2. (3) A SGNS model is trained for q_1 and q_2: M_1^Q and M_2^Q. (4) For every word w in both M_1^Q and M_2^Q, the angular distance of w's vectors in M_1^Q and M_2^Q are recorded. Next, *rectified change* is calculated as the t-statistic of the naive cosine change given the estimated noise distribution from the controls, with Bessel's correction (Noble et al., 2021). That is, for a given word w and a temporal transition from t_i to t_j, *rectified change* is defined as:

$$\Delta^*_{t_i,t_j}(w) = \frac{\Delta_{t_i,t_j}(w) - \bar{x}_{Q,w}}{s_{Q,w}\sqrt{1 + 1/n_Q}}$$

where $\bar{x}_{Q,w}$ and $s_{Q,w}$ are the mean and standard deviation of the naive cosine change measures of the controls (Δ_i^Q , . . . , $\Delta_{n_Q}^Q$). Rectified change can be interpreted as "a measure of how much higher (or lower) the measured naive cosine change is than would be expected if the word's underlying context distribution hadn't changed at all. In other words, it quantifies the strength of the evidence that the word has changed" (Noble et al., 2021). Put differently, rectified change quantifies the evidence that the observed change is a genuine one. As with any statistical test of significance, a significant (genuine) change can be small or large; significance is distinct from effect size.

4.2 The SBERT-PRT approach

The second and third approach use SBERT (Reimers and Gurevych, 2019), which is BERT (Devlin et al., 2019) fine-tuned for predicting the semantic similarity of two sentences (Reimers and Gurevych, 2019). SBERT uses a bi-encoder architecture to solve a problem with computational cost in the sentence pair-regression in original BERT, more precisely its cross-encoder architecture. Reimers and Gurevych (2019) show that a bi-encoder with fine-tuning reaches state-of the art performance on sentence similarity, while using the `[CLS]` token or averaging over tokens without fine-tuning does not. We use SBERT to represent DWEs. Thereby this work contrasts with previous work who uses (simple) BERT for LSC detection. The reason for using SBERT instead of BERT is (i) to give more prominence to the full context of DWEs in representing them, and (ii) to be able to represent words not in the vocabulary of BERT.

The implementation of SBERT-PRT approach is in many respects similar to the implementation of SGNS approach. However, a key difference is that in SBERT-PRT, word vectors are only build for the terms in *S-DWE*, not for the complete vocabulary of C as in SGNS. Thus for SBERT-PRT, let B be a diachronic corpus that covers the same consecutive time periods as in SGNS, i.e. T, but where every sub-corpus b_{t_i} in B is a subset of c_{t_i} such that b_{t_i} = sentence s: s is in c_{t_i} $\wedge$ at least one term from *S-DWE* is in s. Sentences in B are encoded by Swedish SBERT (Rekathati, 2021), resulting in 768-dimensional token vectors.

Swedish SBERT is trained using the method for transfer learning in Reimers and Gurevych (2020) where the objective is to make a student model (of an under-resources language, e.g., Swedish) match the sentence embeddings of a high performing teacher model (developed for a well-resourced language, mostly English) in a parallel corpus. Swedish SBERT is trained with the sentence transformer `paraphrase-mpnet-base-v2` hosted on Hugging Face[5] functioning as teacher model and Swedish BERT (Malmsten et al., 2020) functioning as a student model, using several parallel corpora (Rekathati, 2021).

For every term w in *S-DWE* and for every t_i in T, the mean vector (centroid) of the token vectors for w in t_i constitutes $\overrightarrow{w}_{t_i}$. *Naive cosine changes* for the terms in *S-DWE* are then calculated the same

[5] `https://huggingface.co/`

way as for SGNS (see equation above). Similar to the SGNS approach, controls for calculation of rectified change are construed as follows in the SBERT-PRT: for every transition $\langle t_i, t_{i+1} \rangle$ (in R) and for every term w in *S-DWE*: (1) token vectors from sentences in b_{t_i} and $b_{t_{i+1}}$ (which both contain w) are concatenated and then shuffled; the result being $Q^{t_i,t_{i+1}}$; (2) $Q^{t_i,t_{i+1}}$ is split in half, resulting in subsets q_1 and q_2; (3) the mean vectors (centroids) of the token vectors in q_1 and q_2 are calculated, being $\overrightarrow{w}_{q_1}$ and $\overrightarrow{w}_{q_2}$; (4) the angular distance (naive cosine change) of $\overrightarrow{w}_{q_1}$ and $\overrightarrow{w}_{q_2}$ is calculated and recorded. The calculations of rectified change change then follow the same procedure as in the SGNS approach.

4.3 The SBERT-CLT approach

Every sentence in B (defined above) is independently of its time stamp assigned a label from $l_1, \ldots, l_k$ through k-Means clustering where the value of k is determined by the silhouette method (Rousseeuw, 1987), where k is the number of clusters. After this atemporal labeling, labels are counted per time period. Next, the proportion of labels for a time period t is calculated relative the total counts of labels in t. That is, for every term w in *S-DWE* and every time period t (in T), the proportion of each label is calculated.

The proportions of $l_1, \ldots\ l_k$ at t, call it $L_{w,\,t}$, sums to 1 and can be treated as a probability distribution over labels. In SBERT-CLT, w at t is vectorized as $L_{w,\,t}$, i.e. $\overrightarrow{w}_t = L_{w,\,t}$. Next, in SBERT-CLT, w's change in meaning from t_i to t_{i+1} is measured by through *Jensen-Shannon distance* (JSD), which measures the similarity (difference) between two (or more) probability distributions. JSD is defined as the square root of the symmetrical and smoothed variant of Kullback–Leibler divergence (D_{KL}) of two probability distributions P and Q; see Appendix B.[6] The JSD-based measure of w's semantic change from t_i to t_{i+1}, is defined as follows:

$$\Delta^{JSD}_{t_i,t_{i+1}}(w) = JSD(L_{w,t_i} \parallel L_{w,t_{i+1}})$$

For SBERT-CLT there is no parallel to the shuffled controls to calculate rectified change as in the other two approaches described above.

Approach	Measure	D^{KS}	p
SGNS	naive	0.568	<0.001
SGNS	rectified	0.500	<0.001
SBERT-PRT	naive	0.750	<0.001
SBERT-PRT	rectified	0.318	<0.05
SBERT-CLT	JSD	0.636	<0.001

Table 2: Results of KS-tests (N = 44).

5 Results

For an approach A and a corpus Ω, let $S_{A,\Omega}$ be the series of measures of change at each word–transition combination, $\Delta_1, \ldots\ \Delta_N$, where N is the total number of combinations such that the frequency of w at t_i *and* t_{i+1} is at least 10 (minimum frequency).

H1 has multiple variants depending on which approach that is considered. Moreover, for the SGNS and SBERT-PRT approaches, variants are defined for both naive and rectified change. For SBERT-CLT, only the JSD measure of semantic change is tested. These combinations result in five variants of H1 being tested, one for each of: (1) SGNS with naive change, (2) SGNS with rectified change, (3) SBERT-PRT with naive change, (4) SBERT-PRT with rectified change, and (5) SBERT-CLT with JSD change.

To clarify, for each hypothesis, two series of change measures are defined by the same approach and the same change metric, but for data from different communities, i.e. *Flashback* and *Familjeliv*. Note that for every version of H1 there is a corresponding null hypothesis H0, that the two samples are equal.

Statistically, all variants of H1 are tested through the two-sample Kolmogorov–Smirnov test (KS-test), see Appendix C. The test-statistic D^{KS} of a KS-test provides a measure of the likelihood that two samples derive from the same distribution. Like other statistical testing, if D^{KS} reaches the critical value at the decided alpha-level (α = 0.05), H0 is considered unlikely and is rejected, in support of H1. The KS-tests are only based on transitions which fulfill the minimum frequency criterion in both samples (N= 44).

All versions of H1 are supported (Table 2). For each variant of hypothesis H1, a KS-test supports that the scores of semantic change measured in the *Flashback* data are different from those in the

[6] Here we compare the probability distributon over clusters by Jensen-Shannon *distance* implemented through the Python package *SciPy* (`scipy.spatial.distance.jensenshannon`). This diverges from others who compare probability distributions over clusters by Jensen-Shannon *divergence*, which is the square root of JSD, as defined here. For present purposes, the implementation of Jensen-Shannon divergence or distance does not really matter for the analysis.

Familjeliv data. Thus, semantic change of terms in *S-DWE* is community-dependent. The semantic changes of the terms observed in one community are significantly different from those observed in another community. This observation gives provisional support for the isolated change of DWEs hypothesis.[7]

5.1 Correlation of models

An auxiliary question is the extent to which the different modeling approaches are correlated with one another, which we test here on the *Flashback* data. If they are correlated, then it is more likely that all these measures are capturing the same generalizations about semantic change in this setting. If they are not correlated, then it suggests that they are capturing different aspects of semantic change, which could then motivate future work in determining which components of semantic change are captured by which method.

Correlation of models is measured by Spearman's correlation coefficient ρ of the series of semantic change values. For example, the *correlation*($S^*_{A1,Flashb.}$, $S^*_{A2,Flashb.}$) is measured to test the correlation of SGNS and SBERT-PRT with respect to rectified change, with data from *Flashback*.

Results are shown in Table 3. There are two general observations here. First, the three approaches often disagree. With naive change, the SGNS, SBERT-PRT, and SBERT-CLT are mostly non-correlated or even negatively correlated with each other (Table 3). The first two approaches' relationship with the third approach is weak with rectified change as well (Table 3). Moreover, while the stronger correlations in Table 3 are in the range of 0.4 to 0.6, there is still a large proportion of the variance of the relationships that is not explained. The deeper insight here is that, deciding how to computationally model the semantic change of terms in *S-DWE* is far from trivial. In particular, SBERT-CLT does not have much in common with SBERT-PRT, despite that both approaches are based on Sentence-BERT. Clustering of data and differing distance metrics seem to have an effect, which is an observation in line with previous research.

Second, rectification clearly has an effect. The relationship between the SGNS approach and the SBERT-PRT approach goes from being negatively correlated when considering naive cosine change to being clearly positively correlated when considering rectified change. However, rectification does not have any effect on the first two approaches' relationship with SBERT-CLT. Remember that there was no control for noise in the third approach, but given the convergence of SGNS and SBERT-PRT when considering rectified change, the cluster based method (SBERT-CLT) is clearly "the odd one out". That is, by clustering token embeddings and using another distance measure (JSD instead of angular distance), quite different conclusions about the data seem to emerge.

6 Discussion

This study finds support for the isolated change of DWEs hypothesis. There is a detectable difference in the rate of semantic change of DWEs between the more politically polarized community and the less polarized community. It could have been possible that DWEs change to the **same** degree in the community more representative of the in-group and the community more representative of the out-group, even if they meant different things to the community participants. In that case, our measures would not have detected a difference. But there is a difference in degree likely driven by the communicative needs of the in-group community.

As such, this paper corroborates previous work that has emphasized the role of community in accounting for dogwhistle meanings (Henderson and McCready, 2018; Quaranto, 2022), but this finding must also be seen in the light of a previous emphasis on the importance of community for word meaning in general (Clark, 1996). Following Lewis (1969)'s notion of convention, Clark (1996) writes "conventional meaning hold not for a word *simpliciter*, but for a word *in a particular community*. You can't talk about conventional word meaning without saying what community it is conventional in" (p. 107, emphasis in original). Clark (1996) continues by defining a "communal lexicon" as the set of word conventions of an individual community and notes that such communal lexicons sometimes contain unique word forms (e.g., *quark* in the community of modern physicists), but more often the same word form is shared among different

[7] Correlation measures confirm this. Spearman's correlation (ρ) of S from *Flashback* and *Familjeliv* are close to zero and non-significant (N = 44): $\rho(S_{SGNS,Fla.}, S_{SGNS,Fam.})$ = 0.120 , p = 0.443; $\rho(S^*_{SGNS,Fla.}, S^*_{SGNS,Fam.})$ = 0.120 , p = 0.439; $\rho(S_{SBERT-PRT,Fla.}, S_{SBERT-PRT,Fam.})$ = –0.074 , p = 0.635; $\rho(S^*_{SBERT-PRT,Fla.}, S^*_{SBERT-PRT,Fam.})$ = 0.265 , p = 0.080; and $\rho(S^{JSD}_{SBERT-CLT,Fla.}, S^{JSD}_{SBERT-PRT,Fam.})$ = 0.134 , p = 0.386.

	SGNS	SGNS*	SBERT-PRT	SBERT-PRT*	SBERT-CLT
SGNS	1.000	0.721	-0.306	0.385	0.037
SGNS*	0.721	1.000	-0.239	0.601	0.137
SBERT-PRT	-0.306	-0.239	1.000	-0.383	0.290
SBERT-PRT*	0.385	0.601	-0.383	1.000	0.126
SBERT-CLT	0.037	0.137	0.290	0.126	1.000

Table 3: Cross-correlation (Spearman) of the three approaches (N = 117). Asterix (*) for rectified measures; JSD is used for SBERT-CLT; otherwise, naive measure.

communal lexicons, but with different meanings. The latter case of shared form across communities, but with different meanings that evolve in relation to the local needs and interactions of particular communities is an important insight with clear relevance for an account of dogwhistle meaning.

Although Clark (1996) does not discuss his notion of communal lexicon in relation to semantic change, Noble et al. (2021) have expanded on Clark's ideas and did observe that meanings of terms evolve relative to the communities they are used in Noble et al. (2021). Our result is quantitative evidence in the Swedish online context of different communal lexicons evolving in parallel in relation to a political drive regarding messaging on a controversial topic, immigration and refugees.

Dogwhistle meaning can thus be understood partially in relation to some general principles of lexical meaning. However, whether DWEs' dependence on community for semantic change is especially strong in comparison with words not laden with the role of DWE is an interesting question for future research.

Another point should be noted with regard the isolated change of DWEs hypothesis. Its support has implications for the task of automated detection of dogwhistles, which is important to counteract hidden racist language online, by potential disclosure of concealed derogatory messages. The lesson here, from our experimental support for the isolated change of DWEs hypothesis, is that terms that change in one community, but not in another, are possible indicators of emerging dogwhistles. Although such community specific change of meaning is not a sufficient criterion for the identification of dogwhistles, it can be part of a solution to a complex problem of detecting dogwhistles and other concealed code words, which is gaining increasing attention in NLP (Gavidia et al., 2022; Hertzberg, 2022; Hertzberg et al., 2022; Xu et al., 2021; Zhu and Bhat, 2021).

There are a number of avenues for future work on this topic. One of these would be to address *how* the assumed DWEs change. This can include a more detailed qualitative analysis of the linguistic contexts of the dogwhistles in the years that they exhibit greater change difference between the two communities. Future studies can systematically address the extent that semantic change of these terms is related to their potential dogwhistle functions. For example, do changes reflect encoding of in-group meanings or do they rather reflect other forms of semantic drift, for example, with regard to various topics? Another avenue for future work would be an analysis of the differences between the change measurement approaches, since they are often poorly correlated with one another. A further, more ambitious agenda, would be to identify characteristics of DWE-related lexical semantic change that differ from non-DWE community-based semantic change, which would enable their detection and differentiation in large corpora. Part of this agenda, could be a systematic comparison DWEs and other words with regard to their community divergence of semantic change in order to determine the extent that community divergence is a feature of special importance for words functioning as DWEs compared with words in general.

Limitations

Our work applies to the Swedish political and media context. We believe that it should also apply to other languages, national political contexts, and media, but this will have to be tested by other work.

It is impossible to develop a sample of relevant DWEs that allow for a hypothesis to be tested over DWEs themselves as a general category, since DWEs emerge and disappear based on politically relevant current affairs. Consequently, our work demonstrates our hypothesis for the dogwhistles we present, but we cannot generalize to all dogwhistles everywhere. Nevertheless, showing that the effects are possible and strong is a contribution that makes the case for larger scale testing over newly emerging dogwhistles in different national contexts.

There are also significant differences in the frequencies and distributions of the tested expressions in the two communities of interest. Furthermore, we rely on the rectification approach to deal with the fact that we have a low frequency threshold for including a DWE in the analysis.

Ethics Statement

There is always a problem of dual use when creating a system to detect potentially negative social phenomena. Malicious actors can use the same technique to evaluate, e.g., their own attempts at manipulating political discourse. Nevertheless, we believe that such actors are motivated to do this anyway and that the public research should not be fully "disarmed" and have tools available for detecting these phenomena. Furthermore, this work is a part of the groundwork that will contribute to understanding this phenomenon, and not a full detector in itself.

The community corpus data used in this project was collected from a national repository charged with archiving Swedish political and cultural discourse. The DWE selection was motivated by published experiments conducted by other researchers under the supervision of an ethics review board.

Acknowledgements

Funding for this work was provided by the Gothenburg Research Initiative for Politically Emergent Systems (GRIPES) supported by the Marianne and Marcus Wallenberg Foundation grant 2019.0214. We wish to thank the anonymous reviewers for their constructive criticism and comments.

References

Mathilda Åkerlund. 2021. Influence Without Metrics: Analyzing the Impact of Far-Right Users in an Online Discussion Forum. *Social Media + Society*, 7(2):20563051211008831.

Mathilda Åkerlund. 2022. Dog whistling far-right code words: The case of 'culture enricher'on the Swedish web. *Information, Communication & Society*, 25(12):1808–1825.

Prashanth Bhat and Ofra Klein. 2020. Covert hate speech: White nationalists and dog whistle communication on twitter. *Twitter, the public sphere, and the chaos of online deliberation*, pages 151–172.

Helena Blomberg and Jonas Stier. 2019. Flashback as a rhetorical online battleground: Debating the (dis) guise of the Nordic Resistance Movement. *Social Media+ Society*, 5(1):2056305118823336.

Herbert H. Clark. 1996. *Using Language*. Cambridge university press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yadolah Dodge. 2008. Kolmogorov–Smirnov Test. *The Concise Encyclopedia of Statistics*, pages 283–287.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

J. R. Firth. 1957. A Synopsis of Linguistic Theory, 1930-1955.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms. *arXiv preprint arXiv:2205.02728*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Robert E Goodin and Michael Saward. 2005. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Linnea Hanell and Linus Salö. 2017. Nine months of entextualizations: Discourse and knowledge in an online discussion forum thread for expectant parents. In *Entangled Discourses: South-North Orders of Visibility*, pages 154–170. Routledge, New York.

Ian Haney-López. 2014. *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. Oxford University Press.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10:146–162.

Robert Henderson and Elin McCready. 2018. How dogwhistles work. In *New Frontiers in Artificial Intelligence: JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Tsukuba, Tokyo, November 13-15, 2017, Revised Selected Papers 9*, pages 231–240. Springer.

Niclas Hertzberg. 2022. Semantic modeling of Swedish dog whistles. Master's thesis, University of Gothenburg.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175.

Internetstiftelsen. 2021. Svenskarna och Internet 2021. Technical report.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. *arXiv preprint arXiv:2005.00050*.

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.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. *arXiv preprint arXiv:2103.07259*.

David Lewis. 1969. *Convention: A philosophical study*. Harvard University Press.

Elina Lindgren, Björn Rönnerstrand, Ellen Breitholtz, Robin Cooper, Gregor Rettenegger, and Asad Sayeed. 2023. Can Politicians Broaden Their Support by Using Dog Whistle Communication? In *119th APSA Annual Meeting & Exhibition, August 31 – September 3, 2023, Held in Los Angeles, California*, Los Angeles, California.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karl Malmqvist. 2015. Satire, racist humour and the power of (un) laughter: On the restrained nature of Swedish online racist discourse targeting EU-migrants begging for money. *Discourse & Society*, 26(6):733–753.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden–Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020a. Capturing Evolution in Word Usage: Just Add More Clusters? In *Companion Proceedings of the Web Conference 2020*. ACM.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings Not Always Better than Static for Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73, Barcelona (online). International Committee for Computational Linguistics.

Tali Mendelberg. 1997. Executing Hortons: Racial crime in the 1988 presidential campaign. *The Public Opinion Quarterly*, 61(1):134–157.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings Of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Anne Quaranto. 2022. Dog whistles, covertly coded speech, and the practices that enable them. *Synthese*, 200(4):330.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Faton Rekathati. 2021. The KBLab Blog: Introducing a Swedish Sentence Transformer.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Magnus Sahlgren. 2008. The Distributional Hypothesis. *The Italian Journal of Linguistics*, 20:33–54.

Jennifer Saul. 2018. Dogwhistles, political manipulation, and philosophy of language. In Daniel Fogal, Daniel Harris, and Matt Moss, editors, *New Work on Speech Acts*, pages 360–383. Oxford University Press, Oxford.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Jason Stanley. 2015. *How Propaganda Works*. Princeton University Press.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. Language Science Press Berlin.

Nina Tahmasebi and Haim Dubossarsky. 2023. Computational modeling of semantic change. In Claire Bowern and Bethwyn Evans, editors, *Routledge Handbook of Historical Linguistics*, 2nd edition. Routledge.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

K. Vani, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. *ArXiv*, abs/2010.00857.

Thomas Viehmann. 2021. Numerically more stable computation of the p-values for the two-sample Kolmogorov-Smirnov test.

Wikipedia contributors. 2023. Kolmogorov–Smirnov test — Wikipedia, The Free Encyclopedia.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the dog whistle: A Chinese dataset for cant understanding with common sense and world knowledge. *arXiv preprint arXiv:2104.02704*.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. *arXiv preprint arXiv:2109.04666*.

A Correlation of LSC and word frequency

Table 4 shows correlation of semantic change and word frequency at the first year of transitions (only *Flashback* data).

B Jensen-Shannon distance (JSD)

For two probability distributions P and Q, Jensen-Shannon distance (JSD) is defined as follows:

$$JSD(P \parallel Q) = \sqrt{\frac{D_{KL}(P \parallel \frac{P+Q}{2}) + D_{KL}(Q \parallel \frac{P+Q}{2})}{2}}$$

where D_{KL} can be defined as follows:

$$D_{KL} = \sum_{x \in X} P(x) \log(\frac{P(x)}{Q(x)})$$

where X is the sample space (the labels in the present case).

C Kolmogorov–Smirnov test (KS-test)

Let $F_n(x)$ and $G_m(x)$ be the the empirical cumulative distribution function (ECDF) of two samples X and Y, then:

$$D^{KS}_{n,m} = \text{sup}_x |F_n(x) - G_m(x)|$$

where *sup* is the supremum function, which for present purposes can be approximated by the *max* function (Viehmann, 2021). The null hypothesis ($X = Y$) is rejected at level α, if $D^{KS}_{n,m} > D^{KS}_{n,m,\alpha}$, where:

$$D^{KS}_{n,m,\alpha} = c(\alpha)\sqrt{\frac{n+m}{n \cdot m}}$$

Here $c(\alpha)$ is the inverse of the Kolmogorov distribution at α. For $\alpha = 0.05$, $c(\alpha) \approx 1.358$ (Wikipedia contributors, 2023).

The Mann-Whitney/Wilcoxon rank-sum test (MWW test) is another common non-parametric

	SGNS		SBERT-PRT		SBERT-CLT
DWE	Naive	Rect.	Naive	Rect.	JSD
berika	0.043	0.014	0.278	-0.048	0.386
globalist	-0.767***	-0.11	-0.615**	0.647**	-0.037
hjälpa på plats	-0.253	-0.571*	-0.692**	0.253	-0.275
kulturberika	0.579*	0.524*	-0.844***	0.103	-0.215
kulturberikare	0.279	0.372	-0.16	0.496*	-0.293
återvandra	0.532*	0.257	-0.796***	0.279	-0.386
återvandring	0.05	0.207	-0.638**	0.253	-0.571*

Table 4: Correlation (Spearman's rho) between semantic change (naive, rectified and JSD) and log-transformed fpm (at first year of transition) in the *Flashback* data. Statistical significance is denoted by *p<0.05, **p<0.01, ***p<0.001.

.

tests, which like the KS test, tests the null hypothesis that the underlying distributions of the two samples are equal. However, the MWW test detects a difference between the medians of the samples, while KS test considers the distribution functions collectively not restricted to differences in the central values of the samples (Dodge, 2008).