Can political dogwhistles be predicted by distributional methods for analysis of lexical semantic change?

Max Boholm,[1] **Björn Rönnerstrand,**[2] **Ellen Breitholtz,**[3]
Robin Cooper[3] **Elina Lindgren,**[2] **Gregor Rettenegger,**[2] and **Asad Sayeed**[3]
[1]Gothenburg Research Institute (GRI),
[2]Journalism Media and Communication (JMG),
[3]Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{max.boholm, asad.sayeed}@gu.se

Abstract

We show that meaning shifts in political dogwhistle expressions (DWEs) are explained by the expressions changing with regard to their "hidden" (in-group) and "public" (out-group) dimensions. We study the association between computational measures of Lexical Semantic Change (LSC) and the In-group/Out-group Ratio (IOR) of four Swedish DWEs. We use a combination of distributional modeling of DWEs in the online discussion forum *Flashback* and data collected from a lexical replacement survey of Swedish residents. We explore several vector-space meaning representation approaches and demonstrate that distributional methods can be used to identify semantic shifts relevant to dogwhistle development, particularly contextual representations from Swedish BERT, SBERT, and multilingual T5.

1 Introduction

Online media is important for political communication, but its fast pace makes it very susceptible to meaning manipulation and deceptive communication strategies. Analyzing communicative patterns in such large quantities of data requires computational methods (Theocharis and Jungherr, 2021). In the context of political discourse, this form of data analysis has been used to combat hate speech and related problems for online moderation.

A key challenge for automated analysis of text is identifying implicit meanings (Magu and Luo, 2018). In this work, we explore computational approaches for modeling the temporal dynamics of political dogwhistles. Following Lo Guercio and Caso (2022, p. 203), political dogwhistles can be defined as "speech acts that explicitly convey a certain content to an audience, while simultaneously sending a different, concealed message to a specific subset of that audience" (Saul, 2018; Howdle, 2023; Witten, 2023). Henceforth, we refer to the explicit meaning of dogwhistles as their *out-group meaning,* and the concealed meaning as their *in-group meaning*. We define a *dogwhistle expression* (DWE) as a linguistic form that encodes this dual function and carries both in-group and out-group meanings (Henderson and McCready, 2018).

Dogwhistles that secretly convey racist or otherwise derogatory attitudes are ethical problems for democratic society (Åkerlund, 2022; Lindgren et al., 2023; Bhat and Klein, 2020; Saul, 2018; Stanley, 2015; Haney-López, 2014). Independent of content, dogwhistles have been discussed as problems for democracy by obscuring political mandate and democratic legitimacy (Goodin and Saward, 2005; Howdle, 2023).

Previous work includes theoretical accounts of how dogwhistles work semantically (Breitholtz and Cooper, 2021; Henderson and McCready, 2018; Stanley, 2015; Khoo, 2017; Lo Guercio and Caso, 2022), experiments that test the consequence of dogwhistle communication for the acceptance of policies and attitudes (White, 2007; Wetts and Willer, 2019), and content analyses of how dogwhistles are used online (Bhat and Klein, 2020; Åkerlund, 2022). Less attention has been devoted to the distributional modeling of dogwhistle meaning (but see, e.g., Hertzberg et al., 2022; Mendelsohn et al., 2023; Boholm and Sayeed, 2023; Xu et al., 2021). In particular, while *semantic change* is essential to the concept of dogwhistle, it has only recently been systematically addressed (Boholm and Sayeed, 2023; Sayeed et al., 2024).

Our aim is to combine established methods of lexical semantic change (LSC) detection (Kutuzov et al., 2018; Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021; Tang, 2018) and survey data from linguistic replacement tests (Arefyev et al., 2022; Lindgren et al., 2023) to model the temporal dynamics of dogwhistle meaning over time. The research questions are **(1)** to what extent are computational measures of LSC associated with shifts in the in-group and out-group meanings of DWEs.

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 144–157

Moreover, we ask **(2)** how different approaches to modeling meaning compare with respect to the relationship between LSC and shifts in in-group and out-group meaning over time.

We analyse the relationship between rate of LSC and the in-group–out-group dynamics of dogwhistles through four ways of modeling meaning: (i) skip-gram with negative sampling (**SGNS**) (Mikolov et al., 2013), (ii) Bidirectional Encoder Representations from Transformer **BERT** (Devlin et al., 2019), (iii) Sentence-BERT (**SBERT**) (Reimers and Gurevych, 2019), and (iv) massively multilingual Text-to-Text Transfer Transformer (**mT5**) (Raffel et al., 2020; Xue et al., 2021). These methods *are* sensitive to the dynamic meaning changes of DWEs, suggesting that they can be developed for the detection and analysis of dogwhistle communication online. We also show that the pipelines with the large language models (LLMs) are better at predicting dogwhistle meaning shifts than the SGNS-based pipelines.

2 Related work

Methods of distributional semantics have recently been applied to the long-standing study of semantic change (Bréal, 1904). Advances include the development and validation of approaches for studying when, how, and how much words change. (Kutuzov et al., 2018; Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021; Tang, 2018). To study *how much* and *when* words change, the features of the vector representations can be compared. Formally, the semantic change of a word w in a transition from t_i to t_j can be defined as the distance of w's vector at t_i ($\overrightarrow{w}_{t_i}$) and its vector at t_j ($\overrightarrow{w}_{t_j}$):

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

Diachronic word embeddings have been built as static word embeddings trained at time periods $t_1, \ldots t_n$ (Hamilton et al., 2016b; Kim et al., 2014), such as SGNS (Mikolov et al., 2013), PPMI (Levy et al., 2015) and GloVe (Pennington et al., 2014); by averaging over contextualized token embeddings at $t_1, \ldots t_n$ (Martinc et al., 2020a; Kutuzov and Giulianelli, 2020), using, for example, BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018); and as probability distributions over clusters of contextualised token embeddings at $t_1, \ldots t_n$ (Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020; Martinc et al., 2020b; Vani et al., 2020).

To investigate *how* words change, we can analyze how words' positions change in the vector space (Hamilton et al., 2016a,b). By measuring the distance between the vector of a word w and those of other words, the nearest neighbors of w at time t_i can be compared with its neighbors at t_j (Charlesworth et al., 2022; Vylomova and Haslam, 2021; Tripodi et al., 2019). With predefined concepts (or dimensions) of interest (Caliskan et al., 2017), w's distance to those "concepts" can be tracked over time (Mendelsohn et al., 2020). This latter approach enables exploration of conceptual shifts in large datasets, possibly over long time spans (Garg et al., 2018). For example, Mendelsohn et al. (2020) studied the dehumanization of LGBTQ people in US media by tracking over time the distance between the words for these groups and the vocabulary relevant for the analytical dimensions investigated (e.g., disgust and power). Other work has tested the theory of "concept creep" (Haslam, 2016) by analyzing the semantic shift of harm-related (Vylomova and Haslam, 2021) and health-related concepts (Baes et al., 2023).

The present work analyses dogwhistles and how their in-group and out-group dimensions of meaning change over time. Previously, philosophers of language and linguists have tried to explain the dual meanings of dogwhistles (Breitholtz and Cooper, 2021; Henderson and McCready, 2018). The role of convention versus pragmatic inference is one of the main theoretical issues addressed in this discussion (Breitholtz and Cooper, 2021; Henderson and McCready, 2018; Stanley, 2015; Khoo, 2017; Lo Guercio and Caso, 2022). Few attempts have been made to use distributional semantics to study dogwhistles, but notable exceptions exist. Hertzberg et al. (2022) partitioned in-group and out-group interpretations of DWEs in a word replacement experiment, using SBERT. Xu et al. (2021) built an annotated data set for Chinese dogwhistles. Similarly, Mendelsohn et al. (2023) presented an extensive database of dogwhistle definitions in a US context. In addition, they illustrated the ability of GPT-3 to identify dogwhistles, based on prompts with definitions from their database.

We expand on these efforts to study dogwhistles by combining LSC techniques and suvey data for modeling in-group–out-group dynamics of DWEs. Although time is essential for dogwhistles, since the in-group meaning evolves in parallel to an existing (out-group) meaning (Sayeed et al., 2024), only recently have the temporal aspects of dogwhistles

been systematically studied. Boholm and Sayeed (2023) used computational methods of LSC analysis to model the rate of change of DWEs in different online discussion forums and found that the rate of semantic change of DWEs observed in the highly politically polarized online community diverged from the rate of semantic change of the same terms (at the same period of time) in the less polarized community, suggesting that dogwhistle evolution is community dependent (Quaranto, 2022; Clark, 1996). However, they did not systematically test whether the rate of change observed for the DWEs was explained by systematic variation in the in-group and out-group meaning of the expressions.

3 Data

3.1 Replacement survey

We use data from a word replacement test implemented via a survey of Swedish residents. The aim of this test was to quantify variability in how individuals understand the meaning of dogwhistles. In the first step, we collected potential dogwhistle words from political messaging in Swedish media. Twelve words were included in the replacement test (February and March 2021). The sample (n=1780) consisted of self-recruited panelists, pre-stratified to reflect the Swedish population in terms of age, gender and education.

Panelists were asked to read sentences and instructed to replace a potential DWE in each sentence with one or more words so that the meaning of the sentence remains largely the same. The replacement test was completed by 1,045 panelists, with a participation rate of 51%.

The test was followed by manual coding of responses. A coding manual was drafted and refined by the research group. Coders classified the replacement words into three categories: 1) the implicit dogwhistle meaning, 2) the explicit literal meaning, or 3) word(s) that could not be coded as 1 or 2. In this study, we take DWEs that had high inter-annotator agreement (Krippendorff's $\alpha > 0.6$) and acceptable corpus frequency (at least 10 instances per year when mentioned). We discuss these in the next section.

3.2 Four Swedish DWEs

The in-group meanings of the DWEs analyzed can be listed at a general level. With the out-group meaning of 'suburban gang', the in-group meaning of the dogwhistle ***förortsgäng*** is that of 'immigrant gang'. As such, this DWE works by a biased place-for-person metonymy, similar to *inner city* discussed in US context (Saul, 2018). The DWE ***återvandring*** ('re-migration') has 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*** ('enrich') is the result of malevolent irony, in response to positive opinions on multiculturalism, where the in-group meaning is the opposite of enrichment, namely criminal and destructive activities (by immigrants). In a Swedish context and elsewhere, ***globalist*** is used with several different in-group meanings, including an anti-Semitic reference to Jews, a nationalistic reference to anti-nationalists (i.e., opponents of nationalism), and a populist reference to elitism.

3.3 Corpus

Flashback is a discussion forum with over 1.5 million users and more than 80 million posts, as of 13 March, 2024 (according to the website's own claim). The topics of discussions are organized in "threads" under 15 general sections (e.g., drugs, economy, lifestyle and politics). With anonymous users, *Flashback* is known for discussion of controversial topics and the expression of controversial opinions, including discrimination and racism (Åkerlund, 2021; Blomberg and Stier, 2019; Malmqvist, 2015). Although hate speech and threats are not allowed by the rules, the website clearly contains offensive language. We here analyze *Flashback* data from 2000 to 2022, on the topic of politics. The corpus, which in total contains 49M sentences (posts) and 785M words, was collected from the Swedish national language data processing infrastructure Språkbanken Text.[1] On average, there are 2.1M sentences (SD = 1.4M) and 34.1M words (SD = 21.7M) per year.

There is considerable variation in frequency of the four DWEs analyzed in the corpus (Table 1). In particular, *förortsgäng* is much less frequent than the other terms. Moreover, term frequencies are very different in different years, which is reflected in the high values of the standard deviation.

The corpus has been preprocessed for all pipelines (SGNS, BERT, SBERT and mT5) by lower-casing and removing URLs and emojis. Corpus data for the SGNS approach have been further processed by removal of numbers and punctuation;

[1] https://spraakbanken.gu.se/en/resources/flashback-politik

DWE	Total	M	SD
berika	20936	27.92	12.18
förortsgäng	227	0.23	0.26
globalist	31156	32.07	39.62
återvandring	12999	13.19	22.20

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

separation of compounds that contain the DWEs under analysis as their left-hand element, e.g., "globalistelit" is replaced by "globalist elit" (with space); and lemmatization of the DWEs analyzed, for example, "globalisten" (definite form of *globalist*) is replaced by "globalist" (lemma form). Regular expressions were used for lemmatization and splitting of compounds. For the other approaches, there was no additional step of preprocessing to the steps listed above, but some minor changes were made to facilitate mapping of input words and tokenisation for BERT and mT5.

4 Semantic modeling

Below we introduce four pipelines to test the relationship between the LSC and the in-group/out-group dynamics of DWEs. The piplelines have two basic steps: (a) modeling of the rate of semantic change of DWEs in the corpus; and (b) modeling of the degree of in-group vs. out-group meaning of the DWEs based on the replacements observed in the survey. The key difference between the four pipelines is the algorithm used for modeling meaning: SGNS, BERT, SBERT and mT5.[2]

4.1 LSC modeling

The semantic change of a word w in a transition from t_i to t_j, i.e., $\Delta_{t_i,t_j}(w)$, is defined as the angular distance of w's vector at t_i (i.e., $\overrightarrow{w}_{t_i}$) and its vector at t_j (i.e., $\overrightarrow{w}_{t_j}$) (Kim et al., 2014; Noble et al., 2021):

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

We apply four approaches to build time-indexed word vectors in the diachronic corpus C, which is a collection of sentences from the consecutive set of time periods, $T = \langle 2000, \ldots, 2022 \rangle$. Thus, $C = \langle c_{2000}, \ldots, c_{2022} \rangle$. Vectors are trained only for words at t with a minimum frequency of 10.

[2]Code for running experiments can be found at https://github.com/mboholm/dogwhistle-lsc-predicition.

4.1.1 The SGNS approach

A SGNS model is trained for each sub-corpus in C, in the sorted order of T, from first to last. The weights of the model are randomly initialized for the first time period, M_{2000}, but 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 $R = \langle \langle t_1, t_2 \rangle, \ldots \langle t_{n-1}, t_n \rangle \rangle = \langle \langle 2000, 2001 \rangle, \ldots \langle 2021, 2022 \rangle \rangle$, 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 $\Delta_{t_i,t_j}(w)$. We train six SGNS variants for 100 and 200 dimensions and window sizes of 5, 10, and 15.

4.1.2 The BERT approach

The diachronic corpus B is a subset of C, such that it covers the same consecutive time periods in T, but where every sub-corpus b_t = {sentence s: s is in $c_t \wedge$ at least one the analyzed DWEs is in s}. Sentences in B are encoded by Swedish BERT (Malmsten et al., 2020).[3] A word vectors of a DWE w at t is built in two steps: first, contextualised token embeddings of w in sentences from b_t, are built by averaging over the token embeddings of the last hidden layer of BERT that correspond to w in the input. Next, the mean vector of the contextualized token embeddings for w in t constitutes $\overrightarrow{w}_{t_i}$.[4]

4.1.3 The mT5 approach

The third approach uses the mT5 model (Xue et al., 2021), a multilingual variant of T5 (Raffel et al., 2020) trained on the multilingual extension of the Colossal Clean Crawled Corpus (C4), mC4, which in total contains 6.3T tokens. Swedish is among the 101 languages in mC4. With T5, every NLP task is generalized as text-to-text problem. The model is similar to the original transformer model in Vaswani et al. (2017), with some alternations of, for example, normalization of layers and position embeddings (Raffel et al., 2020; Xue et al., 2021). T5 was originally developed to test, in a unified and controlled way, the effectiveness of transfer learning on a variety of NLP tasks (Raffel et al., 2020). However, our implementation does not fine-tune the pre-trained model. Rather, our main motive for

[3]https://huggingface.co/KB/bert-base-swedish-cased

[4]For some compound words, the tokenization for BERT or mT5 does not perfectly match the DWE part of the compound. We then use the embeddings of tokens that maximize the similarity of the two strings by the Ratcliff et al. (1988) algorithm implemented as `SequenceMatcher` in Python.

using mT5 is to test a recent large-scale transformer. Here we use the 3.7 billion parameter version of the model, named XL.[5]

We build word vectors at *t* as in the BERT approach: contextualized token embeddings are built by averaging token embeddings of the last hidden layer, corresponding to *w* in the input sentence; $\overrightarrow{w}_{t_i}$ is the mean vector of the contextualised token embeddings at *t*.

4.1.4 The SBERT approach

The fourth and final approach uses Swedish SBERT (Rekathati, 2021).[6] SBERT (Reimers and Gurevych, 2019) is BERT (Devlin et al., 2019) fine-tuned for predicting the semantic similarity of two sentences. SBERT has a bi-encoder architecture to reduce the computational cost of sentence pair-regression in original BERT. Reimers and Gurevych (2019) show that a bi-encoder with fine-tuning reaches state-of-the-art performance on sentence similarity. Swedish SBERT is trained with transfer learning in Reimers and Gurevych (2020), where the objective is to make a student model[7] (of an under-resources language, here: Swedish) match the sentence embeddings of a high-performing teacher model[8] (developed for a well-resourced language, here: English) in a parallel corpus.

The implementation of the SBERT approach is in most respects similar to the implementation of the other transformer models, but does not require mapping between token embeddings and the DWE of the input, nor selection of layer, since SBERT output 1 × 768-dimensional vectors that serve as the contextualized token embedding. The mean vector of the contextualized embeddings for *w* at *t* constitutes $\overrightarrow{w}_t$.

4.2 In-group and out-group modeling

We modeled the semantic dimensions of in-group and out-group meaning of a DWE *w* at time *t* by measuring the similarity between (a) the embedding for *w* at *t* trained on online community data (as defined above, sect. 4.1) and (b) the (averaged) embedding for text replacements $R^w = \{r_1^w, ..., r_n^w\}$ for *w* in the replacement survey, annotated as "in-group" (I^w) or "out-group" (O^w). Details on how the in-group and out-group embeddings, $\overrightarrow{I^w}$ and $\overrightarrow{O^w}$, are built from I^w and O^w are presented in the following (sect. 4.2.1 - 4.2.2); each approach parallels those defined above for the analysis of LSC.

Once in-group and out-group embeddings for DWE *w* are derived, we use cosine similarity to calculate an in-group score (*IS*) and an out-group score (*OS*) at each time *t*:

$$IS_t(w) = cossim(\overrightarrow{w_t}, \overrightarrow{I^w})$$

$$OS_t(w) = cossim(\overrightarrow{w_t}, \overrightarrow{O_w})$$

Next, we define the In-group/Out-group Ratio (*IOR*) of DWE *w*, reflecting a normalized measure of *w*'s in-group meaning relative to its out-group meaning (Kapron-King and Xu, 2021):

$$IOR_t(w) = \frac{IS_t(w)}{IS_t(w) + OS_t(w)}$$

To measure the change in *IOR* for *w* over time, we define the absolute difference in *IOR* as:

$$\Delta_{t_i,t_j}^{IOR}(w) = abs(IOR_{t_j}(w) - IOR_{t_i}(w))$$

This study uses linear regression to test whether the difference in IOR (i.e., $\Delta_{t_i,t_j}^{IOR}(w)$) is a predictor of the LSC of DWEs (i.e., $\Delta_{t_i,t_j}(w)$). Regression models are described in more detail below (sect. 5.1), but first vectorization of in-group and out-group dimensions is addressed.

4.2.1 SGNS

For the SGNS approach, vectorization of in-group and out-group dimensions is based on the word embeddings trained for the diachronic corpus data (sect. 4.1.1). A bag-of-words (BOW) approach was implemented to build in-group and out-group embeddings from the SGNS models.[9]

The steps for building in-group and out-group embeddings from the BOW-sets are as follows:[10] first, stopwords were removed. Second, the top 3 words of each BOW set were selected based on

[5] https://huggingface.co/google/mt5-xl

[6] https://huggingface.co/KBLab/sentence-bert-swedish-cased

[7] https://huggingface.co/KB/bert-base-swedish-cased

[8] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[9] An alternative approach could have been to build token vectors of multi-word inputs (replacements) by pooling SGNS word vectors of the input and then averaging over those token vectors (similar to the approaches described below). The BOW approach implemented here has lower computational cost than a pooling of multi-word inputs would have had.

[10] We have tested different strategies for selection. Other examples of strategies include selecting the top 3 most frequent words in I^w and O^w without overlap.

their keyness (Gabrielatos, 2018), using the odds ratio, which is an asymmetric measure of the probability of a word in a target corpus relative to a reference corpus (e.g., the probability of finding a word *x* in the in-group replacements relative to the out-group replacements). Third, we generalize from the selected words, by *adding* related word forms of the same lexeme, using existing resources for Swedish morphology (Borin and Forsberg, 2009). For example, in the replacement survey participants were asked to replace the plural form of the DWE *globalist*, i.e., "globalist*er*" (plural). Consequently, the replacements for *globalist* are dominated by plural forms of nouns, e.g., "elitist*er*" (plural). However, embeddings for other wordforms than the exact ones used in the replacement survey might be relevant for modeling the in-group and out-group dimension. Once the word forms of each lexeme were identified, to minimize the influence of infrequent words, word forms that were not frequent enough to account for at least 20% of the frequency of the lexeme were removed. Finally, after selection and expansion, for each word in a remaining set, its SGNS embedding was collected. The in-group and out-group embeddings are defined as the average vector of the collected embeddings for each set (see Appendix A for examples of words).

4.2.2 BERT, mT5, and SBERT

For BERT and mT5, we represent each replacement r by the average embedding of the last hidden layer (Ni et al., 2021).[11] Since SBERT is designed to represent sentences, there is no need for (additional) pooling of token embeddings. Replacements are represented by sentence embeddings. We define the in-group ($\overrightarrow{I_w}$) and out-group embeddings ($\overrightarrow{O_w}$) as the mean vectors of the contextualized token embeddings for the replacements in I^w and O^w.

For examples of sentences from the *Flashback* training data, with high and low scores of IOR, see Appendix B.

5 Analysis

5.1 Regression models

The relationship between IOR and LSC is modeled by linear regressions (OLS), implemented in Python through `statsmodels` package. We try to predict the rate of semantic change of DWEs ($\Delta_{t_i,t_j}(w)$) from their change in IOR ($\Delta^{IOR}_{t_i,t_j}(w)$). If the coefficient for the (independent) variable is significant, the semantic change observed for the DWEs is explained by their shifting meaning with regard to in-group and out-group meanings. In addition to the significance of the coefficent for Δ^{IOR}, the pipelines defined above can be compared with respect to the total variance explained (R^2). In total, there are 64 DWE-time pairs in the data.

Previous research has shown that semantic change is strongly correlated with term frequency (Dubossarsky et al., 2017; Hamilton et al., 2016b). To avoid having term frequency as a confounding factor between $\Delta_{t_i,t_j}(w)$ and $\Delta^{IOR}_{t_i,t_j}(w)$, we control for the effect of term frequency by having term frequency per million (FPM) (at t_i, log_2-transformed) and proportional change in FPM from t_i to t_j as predictors (control variables).

Thus, we model the following relationship:

$$\Delta_{t_i,t_j}(w) = \beta_0 + \beta_1 \times \Delta^{IOR}_{t_i,t_j}(w) + \beta_2 \times log_2(FPM_{t_i}(w)) + \beta_3 \times \Delta^{FPM}_{t_i,t_j}(w)$$

For comparability, the model variables are normalized by z-scores. We assess there being no problem with multicollinearity, since the variance inflation factor (VIF) for independent variables is close to 1 (below 2) in all models. For all regression models, except the ones based on the pipeline for 200-dimensional SGNS models, the residuals are not normally distributed, as measured by the Jarque-Bera test. Under the assumption of the central limit theorem, we proceed with the regression model proposed above, despite nonnormal residuals, relying on our sample size being sufficiently large (N = 64, with three predictors) (Weisberg, 2013; Schmidt and Finan, 2018). However, we did test transformations of variables to meet the assumption of normal residuals, see Appendix C. The overall patterns are the same.

5.2 Results

For most models, shifts in IOR ($\Delta^{IOR}_{t_i,t_j}(w)$) is a significant predictor of rate of semantic change ($\Delta_{t_i,t_j}(w)$). That is, the rate of change observed for DWEs using common methods for LSC-modeling is related to shifts in in-group and out-group meaning. Overall, these findings suggest that the established computational methods of LSC detection are, in fact, sensitive to the emergence and decline

[11] For BERT we also tested the embedding of the `CLS` token, which resulted in slightly higher R^2 scores. Here we focus on the mean pooling approach for comparability with the pipeline for mT5-XL, which lacks a `CLS` token (Ni et al., 2021).

	Dependent variable: Δ_{t_i,t_j}								
	SBERT	BERT	mT5-XL	SGNS-*w*5-*d*100	SGNS-*w*10-*d*100	SGNS-*w*15-*d*100	SGNS-*w*5-*d*200	SGNS-*w*10-*d*200	SGNS-*w*15-*d*200
$\Delta^{IOR}_{t_i,t_j}$	0.794***	0.546***	0.555***	0.250*	0.129	0.246*	0.273*	0.184	0.361**
	(0.059)	(0.103)	(0.102)	(0.121)	(0.130)	(0.122)	(0.112)	(0.123)	(0.114)
$\Delta^{FPM}_{t_i,t_j}$	-0.022	-0.078	-0.220*	0.256*	0.236	0.198	0.265*	0.202	0.206
	(0.057)	(0.099)	(0.103)	(0.122)	(0.126)	(0.123)	(0.113)	(0.120)	(0.114)
FPM (log)	-0.265***	-0.236*	-0.451***	0.049	-0.032	-0.078	0.347**	0.258*	0.223
	(0.059)	(0.103)	(0.099)	(0.122)	(0.130)	(0.123)	(0.113)	(0.125)	(0.115)
Const.	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.057)	(0.098)	(0.097)	(0.120)	(0.125)	(0.122)	(0.112)	(0.119)	(0.113)
R^2	0.807	0.426	0.430	0.129	0.067	0.113	0.248	0.149	0.240
Adj. R^2	0.798	0.398	0.401	0.086	0.021	0.069	0.210	0.106	0.202
Resid. Std. Error	0.453 (df=60)	0.782 (df=60)	0.780 (df=60)	0.964 (df=60)	0.997 (df=60)	0.973 (df=60)	0.896 (df=60)	0.953 (df=60)	0.901 (df=60)
F Stat.	83.866*** (df=3; 60)	14.862*** (df=3; 60)	15.079*** (df=3; 60)	2.971* (df=3; 60)	1.447 (df=3; 60)	2.545 (df=3; 60)	6.589*** (df=3; 60)	3.495* (df=3; 60)	6.302*** (df=3; 60)
Note:									$N = 64$; *p<0.05; **p<0.01; ***p<0.001

Table 2: Explaining semantic change of DWEs (standardized coefficients)

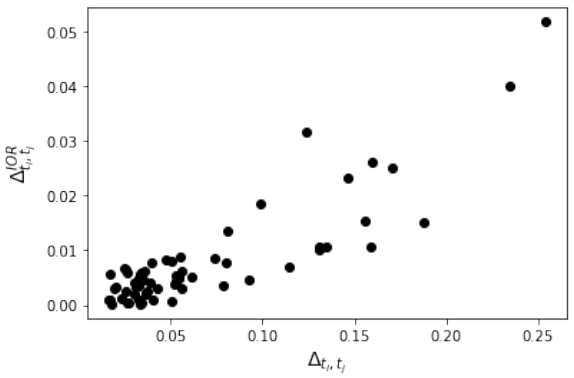

Figure 1: Relationship between LSC and IOR in SBERT pipeline

of dogwhistles. Exceptions to this general observation are found among variants of the SGNS models, where the coefficient for $\Delta^{IOR}_{t_i,t_j}(w)$ is not significant (at $\alpha = 0.05$); namely those with window size = 10. We return to this pattern below.

Out of the pipelines, the LLM-based pipelines explain more variability of the data and have larger coefficients for $\Delta^{IOR}_{t_i,t_j}$, than the SGNS-based models. Thus, in predicting semantic change, these models rely more on the semantic variability related to the IOR, than the SGNS models do. When comparing LLM-based pipelines, the SBERT-based approach shows higher R^2 and a stronger effect of Δ^{IOR} than the BERT and mT5 approaches. For SBERT, the strong correlation between LSC and IOR is illustrated in Figure 1. These observations suggest that sentence embeddings are beneficial for explaining the semantic change of dogwhistles (SBERT), compared with averaging over the embeddings of input tokens mapping to the DWE (BERT, mT5). Note that these findings derive from *pipelines* that contain both the rate of change *and* the IOR. Thus, the different observed can be a consequence of how replacements are represented, how LSC is modeled, or both.

An explanation for why SBERT explains more variability in the data might be that SBERT is fine-tuned for a task that has a similar structure as the one implemented in our pipeline for modeling in-group and out-group scores, namely to predict the similarity of embeddings (Reimers and Gurevych, 2019). It might also be the case that in-group and out-group meanings of DWEs are best captured holistically by sentence representations that give more prominence to the full context of DWEs.

The pipelines with BERT and mT5 are very similar in terms of R^2 and effect of Δ^{IOR}. On the one hand, the large computational overhead of mT5-XL compared to BERT does not result in stronger predictions, as modeled in the present context. On the other hand, the multilingual transformer performs on par with the language-specific one.

For the SGNS models, both the window size and the number of dimensions of the vectors matter. With higher dimensionality of the vectors, more variation in $\Delta_{t_i,t_j}(w)$ is explained. When different window sizes are compared, a U-shaped pattern emerges. For both 100- and 200-dimensional models, the strongest effect of Δ^{IOR} and the highest values of R^2 are observed for window size = 5. However, almost as strong effects are found for

window size = 15, but smaller effect sizes for window size = 10. These observations indicate that words used both in close proximity and far away from the DWE are relevant to communicate in-group messages. This U-shaped pattern may be related to the fact that we model different DWEs. That is, for some DWEs, words in close context may be central to the in-group meaning, but for other DWEs, a wider context is important.

As in previous studies, term frequency (at t_i) explains the rate of semantic change (Hamilton et al., 2016b; Dubossarsky et al., 2017). For LLMs, the relationship is negative: the more frequent a word is, the less it changes, which is in line with "law of conformity" (Hamilton et al., 2016b). However, for the SGNS models, the relation between term frequency and semantic change is in most cases not significant; and when significant, the relationship is, unlike for the LLM pipelines, positive.The *change* in frequency from t_i to t_j have no effect on $\Delta_{t_i,t_j}(w)$, besides the case of SGNS, where window size = 5 and d = 100.[12] In both models for the BERT-pipelines, $\Delta^{IOR}_{t_i,t_j}(w)$ has a stronger effect on Δ_{t_i,t_j} than term frequency, while for mT5 pipeline, the effect of IOR and term frequency are in the same magnitude (though the latter is negative).

6 Discussion

We find that the observed meaning shifts for DWEs using distributional methods are explained by their in-group and out-group dimensions. That is, the methods for detecting LSC *are* sensitive to the dynamic meaning of DWE, suggesting that the measures of LSC could be used to detect dogwhistles online. However, it *could* have been the case that LSC measures did pick up on contextual drifts of DWEs, which were *not* directly related to their function as dogwhistles. After all, as an implementation of the distributional hypothesis, meaning is in LSC detection modeled as statistical correlation over context words.

But context can vary for various reasons, not all of which are straightforward cases of change in meaning (Bender and Koller, 2020). Words can be used in the *same* sense in relation to different topics of discussion at different times, which poses challenges for modeling meaning change (Hengchen et al., 2021; Tang, 2018). For example, previous work has showed that distributional methods for LSC sometimes overgeneralizes due to "referential effects", i.e., the observed change of word usage is explained by reference to different persons or events at different times (Del Tredici et al., 2018). In such cases, "the meaning of the word stays the same, despite the change in context" (Del Tredici et al., 2018, 2073). These types of "semantic" (or contextual) shifts are not clear examples of meaning change or differentiation of senses that have been mainly discussed in theoretical linguistics (Traugott and Dasher, 2002). But from the point of view of distributional semantics, it is difficult to distinguish these different aspects of variable usage (Geeraerts et al., 2024). Given a strict interpretation of the distributional thesis, a change in context *is* a change of meaning.

In the context of these potential challenges that have been raised for the interpretation of distributional LSC detection results, our results are notably interpretable. The rate of change of the DWEs is, in fact, related to changes in the in-group vs. out-group "senses" of these words. From the geometric viewpoint that defines distributional modeling of meaning, the shifting positions of DWEs in semantic space over time (as identified by LSC) are repositioning along the in-group vs. out-group axes (as identified by Δ^{IOR}). Given the high values of R^2, for many of the pipelines tested here, the IOR of the DWEs is a key factor in explaining their semantic variability over time.

The above findings suggest that the pipelines with LLMs are better than the SGNS models at the relevant meaning variation of DWEs. This finding is in line with the general trend, with transformer models having substantially improved the state-of-the-art for NLU tasks. The nuanced semantic representation enabled by these models seems to be important also for the related challenge of modeling dogwhistle meaning.

Future research should attempt to scale up the present approach for the analysis of a wider range of DWEs. A key challenge in doing so is inferring and representing the in-group and out-group dimensions of the DWEs. This study used a survey methodology to develop an independent basis for defining the in-group and out-group dimensions of DWEs, but such an approach is costly, especially at a large scale. Another possibility for future research is using definitions of dogwhistles in existing online databases to represent in-group and out-group embeddings (Mendelsohn et al., 2023).

[12] Other operationalisation of change in term frequency (than percental difference) were tested: (non-proportional) raw change in frequency and absolute difference of frequency, but the overall pattern persists: no effect for predicting $\Delta_{t_i,t_j}(w)$.

Limitations

This work applies to the political media context of Sweden. Although we believe that the general methodologies developed should also apply to other national, linguistic, and political contexts, this must be tested in other work.

Since DWEs emerge and disappear on the basis of politically relevant current affairs, it is not possible to develop a sample of relevant DWEs that allows analysis of DWEs themselves as a general category. As a result, our work shows our hypothesis for an admittedly limited set of dogwhistles from which we cannot make global generalizations. However, the fact that the effects are strong is a contribution that calls for future testing of the methodology at a larger scale, with additional terms, and in other national contexts.

Ethics Statement

When creating a system that detects potentially negative social phenomena, there is always a risk of malicious use of the system. In principle, the developed technology can be used for evaluating, for example, attempts to manipulate political discourse. However, we believe that actors motivated to do so can do so anyway and that public research should not avoid the analysis of harmful communication for this reason. Rather, tools should be developed to detect and combat these harmful phenomena. In addition, this work is part of the foundational work that contributes to understanding dogwhistle communication; it does not enable full detection on its own.

The corpus data used in this project were obtained from a national repository given responsibility for archiving Swedish documents of political and cultural significance. The replacement test survey was approved by the Swedish Ethical Review Authority.

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 as well as a Swedish Research Council (VR) grant (2014-39) for the Centre for Linguistic Theory and Studies in Probability (CLASP). We wish to thank the anonymous reviewers for their constructive 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.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2022. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. *arXiv preprint arXiv:2206.11815*.

Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2023. Semantic shifts in mental health-related concepts. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 119–128.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

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.

Max Boholm and Asad Sayeed. 2023. Political dogwhistles and community divergence in semantic change. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 53–65.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of saldo and wordnet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series*, volume 7.

Michel Bréal. 1904. *Essai de sémantique (science des significations).* Hachette.

Ellen Breitholtz and Robin Cooper. 2021. Dogwhistles as inferences in interaction. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 40–46.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119.

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

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2018. Short-term meaning shift: A distributional exploration. *arXiv preprint arXiv:1809.03169*.

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.

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.

Costas Gabrielatos. 2018. Keyness analysis. In Charlotte Taylor and Anna Marchi, editors, *Corpus Approaches to Discourse: A Critical Review*, pages 225–258. Routledge, London.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2024. *Lexical variation and change: A distributional semantic approach*. Oxford University Press.

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. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. 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.

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

Nick Haslam. 2016. Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological inquiry*, 27(1):1–17.

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.

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

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.

Giles Howdle. 2023. Microtargeting, dogwhistles, and deliberative democracy. *Topoi*, 42(2):445–458.

Anna Kapron-King and Yang Xu. 2021. A diachronic evaluation of gender asymmetry in euphemism. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 28–38, Online. Association for Computational Linguistics.

Justin Khoo. 2017. Code words in political discourse. *Philosophical Topics*, 45(2):33–64.

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.

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.

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.

Nicolás Lo Guercio and Ramiro Caso. 2022. An account of overt intentional dogwhistling. *Synthese*, 200(3):203.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

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, Petra Kralj Novak, and Senja Pollak. 2020a. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

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.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. *Preprint*, arXiv:2305.17174.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, 3.

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.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *Preprint*, arXiv:2108.08877.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

John W Ratcliff, David E Metzener, et al. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46.

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.

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.

Asad Sayeed, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Gregor Rettenegger, and Björn Rönnerstrand. 2024. The utility of (political) dogwhistles–a life cycle perspective. *Journal of Language and Politics*.

Amand F Schmidt and Chris Finan. 2018. Linear regression and the normality assumption. *Journal of clinical epidemiology*, 98:146–151.

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.

Yannis Theocharis and Andreas Jungherr. 2021. Computational social science and the study of political communication. *Political Communication*, 38(1-2):1–22.

"Elizabeth Closs Traugott and Richard B." Dasher. 2002. *Regularity in semantic change*. Cambridge University Press.

Rocco Tripodi, Massimo Warglien, Simon Levis Sullam, and Deborah Paci. 2019. Tracing antisemitic language through diachronic embedding projections: France 1789-1914. *arXiv preprint arXiv:1906.01440*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ekaterina Vylomova and Nick Haslam. 2021. Semantic changes in harm-related concepts in english. *Computational approaches to semantic change*, 6:93.

Sanford Weisberg. 2013. *Applied linear regression*, fourth edition, volume 528. John Wiley & Sons.

Rachel Wetts and Robb Willer. 2019. Who is called by the dog whistle? Experimental evidence that racial resentment and political ideology condition responses to racially encoded messages. *Socius*, 5:2378023119866268.

Ismail K White. 2007. When race matters and when it doesn't: Racial group differences in response to racial cues. *American Political Science Review*, 101(2):339–354.

Kimberly Witten. 2023. The definition and typological model of a dogwhistle. *Manuscrito*, 46:e–2023.

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*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

A In-group and out-group vocabulary for SGNS approach

Table 3 exemplifies the words whose embeddings are used to model the in-group and out-group embeddings of the four DWEs in the SGNS approach.

B Example sentences

Warning: the following examples may be upsetting or offensive to some readers.

Examples are selected from the training corpus (*Flashback*) to illustrate high and low IOR values from years with high and low general IOR values, as measured by BERT. To identify examples, IOR values for individual sentences were computed. That is, we compute $IOR_t(w)$, as defined above, but where $\overrightarrow{w}$ is not the diachronic embedding of *t*, but the embedding of a word instance from the time bin *t*. We show examples from the top ("high IOR") and bottom ("high IOR") five of the sentences of a year, measured by their individual IOR value.

B.1 *berika*

1. (low IOR, 2007) *den typen av invandring är bra och berikande och bör uppmuntras*
   (that kind of immigration is good and enriching and should be encouraged)

2. (high IOR, 2010) *kålsvart hår och mörkt hudpigment, troligen hemmavarande i Iran eller Irak, varför måste vi skandinaver berikas med detta drägg?*
   (coal black hair and dark skin pigment, probably native to Iran or Iraq, why do we Scandinavians have to be enriched with this dreg?)

B.2 *globalist*

3. (low IOR, 2006) *jag är alltså globalist, frihandelsförespråkare, demokrat och kapitalist för att detta är det bästa sättet att göra fattiga*

DWE	In-group	Out-group
berika	förstöra (destroy, inf.), förstör (destroy, pres.), utnyttjar (exploit, pres.), utnyttja (exploit, inf.), negativ (negative), negativa (negative, pl.), negativt (negative, neut.)	positiv (positive), positiva (positive, pl.), positivt (positive, neut.), ger (give, pres.), ge (give, inf.), gynna (benefit, inf.), gynnar (benefit, pres.)
globalist	judar (jews), eliten (elite, def.), elit (elite, indef.)	världsmedborgare (world citizen), internationellt (international, neut.), internationell (international), internationella (international, pl.)
återvandring	utvisning, skickar (send, pres.), skicka (send, inf.)	flytta (move, inf.), återvänder (return, pres.), återvända (return, inf.), hemland (home country, indef.), hemlandet (home country, def.)
förortsgäng	invandrargäng (immigrant gang, indef.), invandrare (immigrant, indef.), invandrarungdomar (immigrant youths)	utsatt (exposed), utsatta (exposed, neut./pl.), förorten (suburb, def.), förorter (suburbs, indef.), förorterna (suburbs, def.), ungdomsgäng (youth gangs, indef.)

Note: def. = definite; indef. = indefinite; inf. = infinitive; neut = neuter; pres. = present; pl. = plural

Table 3: Vocabulary for in-group and out-group

människor rikare och utvecklar alla länder som ingår i handelsutbytet
(so I am a globalist, free trade advocate, democrat and capitalist because this is the best way to make poor people richer and develop all countries that are part of the trade exchange)

4. (high IOR, 2008) *globalist-maffian med judarna i spetsen har ju mer eller mindre full kontroll över Amerika, och därmed har dom tillgång till världens starkaste armé*
(the globalist mafia with the Jews at the head has more or less full control over America, and thus they have access to the world's strongest army)

B.3 *återvandring*

5. (low IOR, 2011) *de flesta invandrar p.g.a studier, arbete, återvandring eller för att de har anhöriga i Sverige*
(most people immigrate due to studies, work, re-migration or because they have relatives in Sweden)

6. (high IOR, 2018) *återvandring och utvisning nu, det är enda lösningen*
(re-migration and deportation now, that is the only solution)

B.4 *förortsgäng*

7. (low IOR, 2014) *kan tillägga att vi var ett helsvenskt förortsgäng med 50 % skinnskallar och 50 % fotbollshuliganer*
(can add that we were an all-Swedish suburban gang with 50 % skinheads and 50 % football hooligans)

8. (high IOR, 2015) *ett passivt / slappt invandrarflöde orsakar sånt, och man måste aktivt minska folkvandringen som bosätter sig i förorterna om man vill bli av med förortsgäng*
(a passive / slack immigrant flow causes that, and you have to actively reduce the migration of people settling in the suburbs if you want to get rid of suburban gangs)

C Transformations

To maximize the number of models having normal distribution of residuals, we tested combinations of log transformation of variables. The log transformation of the dependent variable and of the $\Delta^{IOR}_{t_i,t_j}$ resulted in normally distributed residuals for all models but BERT and mT5-XL. No combination of transformed variables was found that makes the error term normally distributed for all models. The regression models for the transformed data are shown in Table 4.

	Dependent variable: $log_2(\Delta_{t_i,t_j})$								
	SBERT	BERT	mT5-XL	SGNS-*w*5-*d*100	SGNS-*w*10-*d*100	SGNS-*w*15-*d*100	SGNS-*w*5-*d*200	SGNS-*w*10-*d*200	SGNS-*w*15-*d*200
Δ^{IOR}	0.640***	0.424***	0.464***	0.286*	0.190	0.272*	0.285*	0.209	0.351**
	(0.066)	(0.092)	(0.091)	(0.119)	(0.128)	(0.121)	(0.109)	(0.120)	(0.113)
Δ^{FPM}	-0.006	-0.043	-0.159	0.259*	0.249	0.205	0.257*	0.207	0.205
	(0.065)	(0.088)	(0.092)	(0.120)	(0.125)	(0.122)	(0.110)	(0.117)	(0.113)
FPM (log)	-0.451***	-0.493***	-0.647***	0.097	-0.033	-0.071	0.400***	0.300*	0.268*
	(0.067)	(0.092)	(0.088)	(0.120)	(0.129)	(0.122)	(0.110)	(0.122)	(0.114)
const	0.000	0.000	0.000	0.000	0.000	0.000	-0.000	0.000	0.000
	(0.064)	(0.087)	(0.087)	(0.119)	(0.123)	(0.120)	(0.109)	(0.116)	(0.111)
R^2	0.757	0.541	0.549	0.156	0.089	0.129	0.289	0.189	0.256
Adj. R^2	0.745	0.518	0.527	0.114	0.043	0.085	0.253	0.148	0.219
Resid. Std. Error	0.509	0.699	0.693	0.949	0.986	0.964	0.871	0.930	0.891
	(df=60)	(df=60)	(df=60)	(df=60)	(df=60)	(df=60)	(df=60)	(df=60)	(df=60)
F Stat.	62.391***	23.607***	24.378***	3.709*	1.949	2.958*	8.128***	4.659**	6.889***
	(df=3; 60)	(df=3; 60)	(df=3; 60)	(df=3; 60)	(df=3; 60)	(df=3; 60)	(df=3; 60)	(df=3; 60)	(df=3; 60)

Notes: $N = 64$; *p<0.05; **p<0.01; ***p<0.001
Relationship estimated: $log_2(\Delta_{t_i,t_j}(w)) = \beta_0 + \beta_1 \times \Delta^{IOR}_{t_i,t_j}(w) + \beta_2 \times log_2(FPM_{t_i}(w)) + \beta_3 \times \Delta^{FPM}_{t_i,t_j}(w)$

Table 4: Explaining semantic change of DWEs (standardized coefficients), log-transformed data