Grammatical Profiling for Semantic Change Detection

Mario Giulianelli[*]
ILLC, University of Amsterdam
m.giulianelli@uva.nl

Andrey Kutuzov[*]
University of Oslo
andreku@ifi.uio.no

Lidia Pivovarova[*]
University of Helsinki
first.last@helsinki.fi

Abstract

Semantics, morphology and syntax are strongly interdependent. However, the majority of computational methods for semantic change detection use distributional word representations which encode mostly semantics. We investigate an alternative method, grammatical profiling, based entirely on changes in the morphosyntactic behaviour of words. We demonstrate that it can be used for semantic change detection and even outperforms some distributional semantic methods. We present an in-depth qualitative and quantitative analysis of the predictions made by our grammatical profiling system, showing that they are plausible and interpretable.

1 Introduction

Lexical semantic change detection has recently become a well-represented field in NLP, with several shared tasks conducted for English, German, Latin and Swedish (Schlechtweg et al., 2020), Italian (Basile et al., 2020) and Russian (Kutuzov and Pivovarova, 2021a). The overwhelming majority of solutions employ either static word embeddings like word2vec (Mikolov et al., 2013) or more recent contextualised language models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These models build upon the distributional semantics hypothesis and can capture lexical meaning, at least to some extent (e.g., Iacobacci et al., 2016; Pilehvar and Camacho-Collados, 2019; Yenicelik et al., 2020). Thus, they are naturally equipped to model semantic change.

Yet it has long been known for linguists that semantics, morphology and syntax are strongly interrelated (Langacker, 1987; Hock and Joseph, 2019). Semantic change is consequently often accompanied by morphosyntactic drifts. Consider the English noun '*lass*': in the 20th century, its 'SWEETHEART' meaning became more dominant over the older sense of 'YOUNG WOMAN'. This was accompanied by a sharp decrease in plural usages ('*lasses*'), as shown in Figure 1.

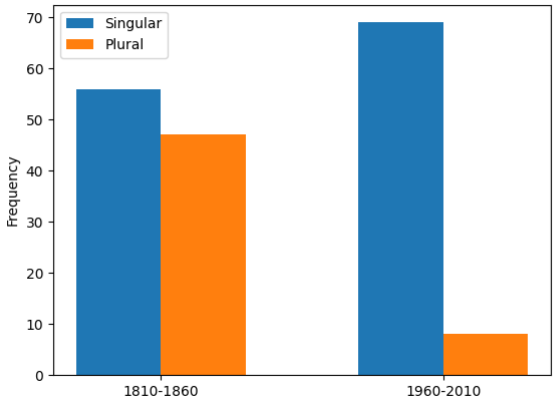

Figure 1: Changes in the number category distribution for the English noun '*lass*' over time, calculated on the English corpora of the SemEval 2020 shared task 1 (Schlechtweg et al., 2020). '*Lass*' is annotated as semantically changed in the SemEval dataset.

Exploiting distributions of *grammatical profiles*—i.e., morphological and syntactic features—to detect lexical semantic change is the focus of this paper. We investigate to what extent lexical semantic change can be detected using *only* morphosyntax. Our main hypothesis is that significant changes in the distribution of morphosyntactic categories can reveal useful information on the degree of the word's semantic change, even without help from any lexical or explicitly semantic features.

Due to the interdependence of semantics and morphosyntax, it is often difficult to determine which type of change occurred first, and whether it triggered the other. Establishing the correct causal direction is outside the scope of this study; it is sufficient for us to know that semantic and morphosyntactic changes often co-occur.

By proposing this functionalist approach to lexical semantic change detection, we are not aiming at establishing a new state-of-the-art. This

[*]Equal contribution, the authors listed alphabetically.

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 423–434
November 10–11, 2021.

is hardly possible without taking semantics into account. But what exactly *is* possible in such a functionalist setup?

We investigate this question experimentally[1] using standard semantic change datasets for English, German, Swedish, Latin, Italian and Russian. Our main findings are the following:

1. Tracing the changes in the distribution of dependency labels, number, case, tense and other morphosyntactic categories outperforms count-based distributional models. In many cases, prediction-based distributional models (static word embeddings) are outperformed as well. This holds across six languages and three different datasets.
2. Morphological and syntactic categories are complementary: combining them improves semantic change detection performance.
3. The categories most correlated with semantic change are language-dependent, with number being a good predictor cross-linguistically.
4. The predictions derived from grammatical profiling are usually interpretable (as in the '*lass*' example above), which is not always the case for methods from prior work based on word embeddings, either static or contextualised. This makes our method suitable for linguistic studies that require qualitative explanations.

2 Related work

Behavioural profiles were introduced in corpus linguistics by Hanks (1996) as the set of syntactic and lexical preferences of a word, revealed by studying a large concordance extracted from a corpus. The behavioural profile of a word consists of corpus counts of various linguistic properties, including morphological features, preferred types of clauses and phrases, collocates and their semantic types (Gries and Otani, 2010). Subtle distinctions in word meaning are reflected in behavioural profiles. Indeed this technique, which combines lexical and grammatical criteria for word sense distinction, was used to study synonymy and polysemy (Divjak and Gries, 2006; Gries and Divjak, 2009) as well as antonymy (Gries and Otani, 2010).

One of the theoretical roots for profiling is the theory of *lexical priming* (Hoey, 2005). According to this theory, words trigger a set of grammatical and lexical constraints, referred to as *primings* and stored in a mental concordance. The theory states that 'Drifts in priming ... provide a mechanism for temporary or permanent language change' (Hoey, 2005, p. 9), and since primings are thought to be organised in the mental concordance in the form of behavioural profiles (Gries and Otani, 2010), it is theoretically plausible that diachronic word meaning change is reflected in a change of behavioural profiles. As far as we are aware, this idea has not been further developed in corpus linguistics.

In spite of its theoretical validity, behavioural profiling as a practical data analysis technique has serious limitations. Profiles include a large variety of word properties and some of them, especially those related to semantics, cannot be easily extracted from a corpus automatically. Usually, a particular subset of word properties is selected based on researchers' intuition and background knowledge, and statistical tests are sometimes used for feature selection at later stages of the analysis (Divjak and Gries, 2006). Moreover, the variety of properties comprised in a behavioural profile makes statistical analysis difficult due to correlations between language phenomena of different levels and sparsity of the data (Kuznetsova, 2015, section 2.2.2). For these reasons, some studies (Janda and Lyashevskaya, 2011; Eckhoff and Janda, 2014) reduce a word's possibly very broad behavioural profile to a more compact *grammatical profile*, i.e. a set of preferred morphological forms for the word. These studies too, however, rely on an *a priori* selection of relevant morphological tags.

These technical difficulties may explain why profiling has not been used in computational approaches to lexical semantic change detection. Most attempts to tackle word meaning change in NLP are based on distributional patterns of *lexical* co-occurrences, starting from early count-based approaches (Juola, 2003; Hilpert and Gries, 2008), continuing with dimensionality reduction techniques (Gulordava and Baroni, 2011), and later accelerated by embeddings-based models (Kutuzov et al., 2018). More recently, contextualised embeddings were also applied to this task (Giulianelli et al., 2020; Montariol et al., 2021).

As far as we are aware, there is one exception to this trend: Ryzhova et al. (2021) employed grammatical profiles to detect the semantic change of Russian nouns. In their work, a profile of case and number frequency distributions is collected separately for each time period, and the degree of

[1] Our code is available at https://github.com/glnmario/semchange-profiling

semantic change is measured as the cosine distance between the two distributions. The results obtained with this method are close to the results yielded by word2vec embeddings, but lower than those of contextualised embeddings. Inspired by Ryzhova et al. (2021), we further investigate the ability of grammatical profiles to capture word meaning change. We propose a number of improvements and evaluate them on datasets in six different languages. Most importantly, we use *all* available morphological tags, without any manual pre-selection, and we conduct an in-depth analysis of our results to understand why grammatical profiling works for this task and what are its limitations.

3 Data and tasks

Following the standard evaluation approach adopted for automatic lexical semantic change detection, we cast the problem as either binary classification (Subtask 1, using the terminology of the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task (Schlechtweg et al., 2020)) or as a ranking task (Subtask 2). In Subtask 1, given a set of target words, a system must determine whether the words lost or gained any senses between two time periods. In Subtask 2, a system has to rank a set of target words according to the degree of their semantic change.

Annotating data for word meaning change detection is a non-trivial process because it requires taking into account numerous word occurrences from every time period of interest. The current practice adopted in the community is to annotate pairs of sentences containing a target word used either in the same or in a different sense; then pairwise scores are aggregated to obtain a final measure of change, either binary or continuous (Schlechtweg et al., 2018). This procedure has been used by organizers of three recent shared tasks: the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task (Schlechtweg et al., 2020), EvaLita (Basile et al., 2020) and RuShiftEval (Kutuzov and Pivovarova, 2021a). We use the data from these three shared tasks, allowing to compare our approach with the state-of-the-art results obtained by distributional models.

The SemEval dataset consists of target words in four languages—37 English, 48 German, 40 Latin, and 32 Swedish—that are manually annotated for both subtasks. The EvaLita dataset consists of 18 Italian words annotated for Subtask 1 only. Finally, the RuShiftEval dataset consists of 99 Russian nouns annotated for Subtask 2. All datasets are accompanied by diachronic corpora. Most of the corpora are split in two time periods, except for the RuShiftEval corpus, which is separated into three time bins: *Russian1* and *Russian2* are annotated with semantic shifts between the pre-Soviet and Soviet periods, and between the Soviet and post-Soviet periods respectively; *Russian3* is annotated with semantic shifts between the pre-Soviet and post-Soviet periods (Kutuzov and Pivovarova, 2021b).

In sum, we have at our disposal several dozens words from three Indo-European language groups: Italic, Germanic and Slavic. Though our results may not generalize to other language families or to other languages within the families analysed, these are the most diverse data that are currently available for this kind of study.

4 Methods

4.1 Basic procedure

To obtain grammatical profiles, the target historical corpora are first tagged and parsed with UDPipe (Straka and Straková, 2017).[2] Then we count the frequency of morphological and syntactic categories for each target word in both corpora. More precisely, we count the FEATS values of a corpus's CONLLU file and store the frequencies in two data structures—one for each time period. For example, `{'Number=Sing': 338, 'Number=Plur': 114}` is the morphological dictionary obtained for an English noun in a single time period. We store syntactic features in an additional dictionary, where keys correspond to the labels of the dependency arc from the target word to its syntactic head (as found in the DEPREL field of a CONLLU-formatted corpus).

For each target word and for both morphological and syntactic dictionaries, we create a list of features by taking the union of keys in the corresponding dictionaries for the two time bins. The feature list will be `['Number=Sing', 'Number=Plur']` for the example above. Then, we create feature vectors $\vec{x}_1$ and $\vec{x}_2$, where each dimension represents a grammatical category and the value it takes is the frequency of that category in the corresponding time period. If a feature does

[2] We use the following models: *english-lines-ud-2.5*, *german-gsd-ud-2.5*, *latin-proiel-ud-2.5*, *swedish-lines-ud-2.5*, *russian-syntagrus-ud-2.5*, *italian-isdt-ud-2.5*.

not occur in a time period, its value is set to 0. The resulting feature vectors represent grammatical profiles for a word in the corresponding periods. Since the feature list is produced separately for each word, the size of the vectors varies across words.

Finally, we compute the cosine distance $cos(\vec{x}_1, \vec{x}_2)$ between the vectors to quantify the change in the grammatical profiles of the target word. This is done separately for morphological and syntactic categories, yielding two distance scores d_{morph} and d_{synt}. They are used directly to rank words in Subtask 2: the larger is the distance, the stronger is the semantic change. To solve the binary classification task (Subtask 1), we classify the top n target words in the ranking as 'changed' (1) and the rest of the list as 'stable' (0). The value of n can be either set manually or inferred from the ranking using off-the-shelf algorithms of change point detection (Truong et al., 2020).

We also combine the scores obtained separately for morphological and syntactic tags by averaging d_{morph} and d_{synt} for each target word (rounding to the nearest integer in the case of binary classification) and then re-rank the words according to the resulting values. In the end, we have three solutions for each task: 'morphology', 'syntax' and 'averaged'. In the next subsections, we describe a number of improvements that we use to amend this basic procedure.

4.2 Filtering

To reduce noise that could be introduced due to rare word forms and possible tagging errors, we exclude rare grammatical categories from the analysis. A feature is filtered out from a feature vector $\vec{x}$ if the sum of the feature occurrences in the two time slices amounts to less than five percent of the total word usages. It is possible to optimise this threshold, but we do not tune any numerical parameters to avoid over-fitting to the target datasets.

4.3 Category separation

In the basic procedure described above, we extract exactly one morphological feature for each word occurrence; this type of morphological feature is a combination of morphological categories that exhaustively describes a word form. For example, this is an excerpt from a grammatical profile of the English verb '*circle*' in the 1810-1860 time period:

```
Tense=Pres|VerbForm=Part : 50
Mood=Ind|Tense=Past|VerbForm=Fin : 24
Tense=Past|VerbForm=Part|Voice=Pass : 17
VerbForm=Inf : 9
Mood=Ind|Tense=Pres|VerbForm=Fin : 1
Tense=Past|VerbForm=Part : 1
```

This representation is very sparse—some features appear only once in the corpus—and it conflates categories of different nature, such as verb form and tense. We therefore introduce a category separation step, where feature vectors are created separately for each morphological category. Thus, we transform a distribution of *word forms* into a distribution of *morphological categories* and obtain a denser and more meaningful representation:

```
Tense : {Past 42, Pres 51}
VerbForm : {Part 68, Fin 25, Inf 9}
Mood : {Ind 25}
Voice : {Pass : 17}
```

Then cosine distance is computed for each category separately. In the example above, we obtain separate distance values for `Tense`, `VerbForm`, `Mood`, and `Voice`; the number of distances differs across words and languages. We take the maximum distance value as the final change score, assuming that a significant change in the distribution of a single category indicates semantic change, regardless of the other categories.[3]

When separation is combined with filtering, filtering is performed *after* feature separation to preserve maximum information. Continuing with the previous example: in the basic procedure, the word form `Tense=Past|VerbForm=Part` is filtered out, as it appears once in the first corpus and it is rare in the second corpus as well. In the category separation strategy this form is taken into account, separately contributing to the `Tense` and `VerbForm` distances.

4.4 Combination of morphology and syntax

Category separation opens new possibilities for taking syntactic categories into account. We can average morphological and syntactic distances, as in our basic procedure, or append the syntactic distance value to the array of morphological distances, and then choose the maximum. In the first strategy, morphological and syntactic rankings are weighted equally regardless of the number of morphological categories for a given word. In the second strategy,

[3] We also experimented with averaging category distances. This improves the results compared to using categories without separation, but it is not as effective as taking the maximum.

syntactic labels are weighted down depending on the richness of the morphological profile.

5 Results

We evaluate our method on both subtasks of the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task (Schlechtweg et al., 2020). As described in Section 3, Subtask 1 is a binary classification task, evaluated with accuracy. Subtask 2 is a ranking task, evaluated with Spearman's rank correlation.

Basic procedure Using only morphological features, we obtain an average correlation of 0.181 across the four SemEval languages, as can be seen in Table 1. Syntactic features yield a $+0.017$ increase, and after averaging d_{morph} and d_{synt} (see Section 4.1) we reach a correlation score of 0.208. This is already substantially higher than the SemEval baseline which employed count-based distributional models (see Table 1).

Frequency threshold Filtering out rare features as described in Section 4.2 has a small but positive impact on all three setups: $+0.011$ for morphological features, $+0.033$ for syntactic features, and $+0.065$ for the combination of the two.

Category separation Measuring distance between morphological categories separately (see Section 4.3) produces an additional significant boost: we obtain a correlation score of 0.278 using these refined morphological representations. In combination with syntactic features (Section 4.4), this approach yields an average correlation of 0.369 with human judgements. This is our best result on Subtask 2, more than twice higher than a correlation obtained by the SemEval count-based baseline (see Table 1); for Latin, a language with rich morphology, grammatical profiles actually outperform even the *best* SemEval 2020 submission. These scores are particularly impressive given that, unlike those based on distributional vectors, our method has no access to lexical semantic information.

As can be seen in Table 1, our category separation approach does not extend well to the Russian test sets, obtaining an average correlation score of 0.130.[4] A possible explanation for the lower correlation may be related to smaller distances between Russian time bins as compared to the SemEval setup: *Russian1* and *Russian2* are annotated with semantic shifts between pre-Soviet and Soviet and between Soviet and post-Soviet periods respectively, while *Russian3* measures the change between pre-Soviet and post-Soviet periods, with a significant time gap in between. Indeed we obtain much higher scores on *Russian3*. In addition, the annotation procedures for the RuShiftEval dataset differ in some details from those for SemEval'20.

Another observation is that morphological category separation does not improve results for English. The best method for English relies only on syntactic features. The most plausible explanation is that English morphology is rather poor and it tends to mark grammatical categories with separate words. Our method can be potentially improved by taking into account multi-word forms, e.g. to determine English verb mood.

Subtask 1 Following our basic procedure (Section 4.1), we assign a classification score of 1 to the top 43% of the target words[5] for each language, ranked according to their grammatical profile changes. This yields an accuracy close to that of the SemEval count-based baseline (see Table 2).[6] Filtering rare features hardly yields any improvement here, but once combined with morphological category separation and automatic change point detection it produces an accuracy of 0.603. We also observe that using change point detection with dynamic programming (Truong et al., 2020) does not cause any significant accuracy decrease in comparison to using the hard-coded 43% ratio, showing that our method does not require knowledge of the test data distribution. On the Italian test set, we correctly classify 3 more words (out of 18) than the collocation-based baseline (Basile et al., 2019b), obtaining an accuracy of 0.778.

6 Qualitative analysis

In Section 5, we showed that grammatical profiling alone can detect a word meaning change better than count-based distributional semantic models which exploit lexical co-occurrence statistics. This is a remarkable finding: it confirms that meaning change leaves traces in grammatical profiles and it demonstrates that these traces can be used as effective predictors of a word's meaning stability. In this Section, to better understand when change in grammatical profiles is a good indicator of lexical

[4] At the same time, in the basic procedure, morphological features yield a much higher correlation score of 0.225.

[5] Average ratio of changed words across SemEval datasets.

[6] Note that the SemEval'20 count baseline also uses a manually defined threshold value in Subtask 1.

Categories	SemEval 2020 languages					Russian			
	English	German	Latin	Swedish	Mean	Russian1	Russian2	Russian3	Mean
	Basic procedure								
Morphology	0.234	0.043	0.241	0.207	0.181	**0.137**	0.210	**0.327**	**0.225**
Syntax	0.319	0.163	0.328	-0.017	0.198	0.060	0.101	0.269	0.143
Average	0.293	0.147	0.304	0.088	0.208	0.101	0.191	0.294	0.195
	5% filtering								
Morphology	0.211	0.080	0.285	0.191	0.192	0.127	0.185	0.264	0.192
Syntax	**0.331**	0.146	0.265	0.184	0.231	0.056	0.111	0.279	0.149
Average	0.315	0.171	0.345	0.263	0.273	0.094	0.183	0.278	0.185
	Category separation and 5% filtering								
Morphology	0.218	0.074	0.519	0.303	0.278	0.028	**0.241**	0.293	0.187
Average	0.321	0.227	0.523	**0.381**	0.363	0.002	0.179	0.278	0.153
Combination	0.320	**0.298**	**0.525**	0.334	**0.369**	0.000	0.149	0.242	0.130
	Prior SemEval results					**Prior RuShiftEval results***			
Count baseline	0.022	0.216	0.359	-0.022	0.144	0.314	0.302	0.381	0.332
Best shared task system (Ryzhova et al., 2021)	0.422	0.725	0.412	0.547	0.527	0.798	0.803	0.822	0.807
	-	-	-	-	-	0.157	0.199	0.343	0.233

Table 1: Performance in graded change detection (SemEval'20 Subtask 2 and RuShiftEval), Spearman rank correlation coefficients. Note that RuShiftEval features three test sets for three different time period pairs.
*The RuShiftEval baseline relies on CBOW word embeddings and their local neighborhood similarity. (Ryzhova et al., 2021) used an ensemble method with much higher performance, we report the results obtained solely with profiling. While SemEval results are fully unsupervised, the best RuShiftEval results are supervised and not directly comparable to our setting.

Categories	English	German	Latin	Swedish	Mean	Italian
	Basic procedure					
Morphology	0.595	0.521	0.525	0.581	0.555	0.722
Syntax	0.541	**0.646**	0.575	0.645	0.602	0.611
Average	0.568	0.583	0.475	**0.710**	0.584	0.722
	Automatic change point detection					
Morphology	**0.622**	0.479	**0.625**	0.548	0.569	0.722
Syntax	0.514	0.625	0.500	0.677	0.579	0.611
Average	0.595	0.542	0.525	0.677	0.585	**0.778**
	Category separation, change point detection and 5% filtering					
Morphology	0.622	0.583	**0.625**	0.581	**0.603**	0.500
Average	0.595	0.625	0.450	**0.710**	0.595	0.667
Combination	0.541	0.583	0.575	0.645	0.586	0.500
	Prior SemEval results					**Prior EvaLita results***
Baseline	0.595	0.688	0.525	0.645	0.613	0.611
Best shared task system	0.622	0.750	0.700	0.677	0.687	0.944

Table 2: Performance in binary change detection (SemEval'20 Subtask 1 and EvaLita), accuracy. Note that in this paper we mostly focus on ranking (Subtask 2). All the binary change detection methods here are entirely based on the scores produces by the ranking methods.
*The Italian baseline relies on collocations (Basile et al., 2019a): for each target word, two vector representations are built, with the Bag-of-Collocations related to the two different time periods. Then, the cosine similarity between them is computed.

semantic change, we analyse the characteristics of the target words to which our method assigns the most and least accurate rankings.

6.1 When is grammatical profiling enough?

We begin by analysing the most accurately ranked words (see Appendix A). The Italian word '*lucciola*', for example, is ranked 1st out of 18 by our method due to the singular usages of the word disappearing after 1990. The singular usage is indeed much more likely for the dying sense of the word (an euphemism for 'PROSTITUTE'), whereas the plural form '*lucciole*' is more likely used for the stable sense of the word ('FIREFLIES') or in the idiomatic expression *prendere lucciole per lanterne* (*getting the wrong end of the stick*), which makes up for most of the occurrences between 1990 and 2014. Another example of correctly identified semantically shifted words is the Latin '*imperator*' (ranked 1st out of 40). In the second time period—ranging from 0 to 2000 A.D.—nominative usages become predominant. A possible explanation for this change is that the more frequent agentive usages of the word correspond to the new role of the 'EMPEROR' in the imperial Rome (27 B.C. to A.D. 476) rather than that of a generic 'COMMANDER'—the older sense of the word.[7]

For English, the noun '*stab*' is ranked 4th out of 37, mostly because of syntactic changes: 27% of its occurrences in the 20th century are used as oblique arguments, compared to only 13% in the 19th century. This is arguably associated with the emergent sense of 'SUDDEN SHARP FEELING' ('*...left me with a sharp stab of sadness*'). The German word '*artikulieren*' correctly receives a high rank (9th out of 48): it occurs only 3 times in the 19th century and 210 times between 1946 and 1990, shifting towards a much richer grammatical profile. Sharp changes in frequency are reflected in the diversity of grammatical profiles and can also help detect lexical semantic change.

Our qualitative analysis reveals that the successful examples are often cases of broadening and narrowing of word meaning. These kinds of semantic change seem to be easily picked with profiling. However, some examples of broadening and narrowing fail to be detected, as will be shown in Section 6.2, especially if they involve metaphorical extensions of word meaning. A consistent characterisation of the kinds of semantic change detected and overlooked by our method would require diachronic corpora where both the degree and the type of semantic changes are annotated.

6.2 When it is not enough?

Although it largely outperforms simple distributional semantic models, our grammatical profiling approach is still not on par with state-of-the-art semantics-based algorithms. To find out when changes in morphosyntactic profiles are not sufficient to detect a word's meaning change, we analyse *false positives* and *false negatives*: i.e., target words that are assigned an erroneously high or low semantic change score, respectively.

False positives are words whose change in grammatical profile does not correspond to semantic change. An example of a false positive is the Italian word '*cappuccio*' ('HOOD'). The increase from 9% to 41% of plural usages causes our method to assign this word a relatively high change score—6th out of 18 (6 words are annotated as changing in the Italian dataset). Inspecting the Italian corpora, we notice that between 1945 and 1970 the word is mainly used to describe the pointed hood of the robes typically worn by Ku Klux Klan members; after 1990, the word's context of usage becomes much less narrow. The meaning of the word, however, does not change. This type of errors is, at least to a certain extent, an artifact of the source data: grammatical profiles are less accurate when the set of domains covered by a corpus is limited.

Another type of false positives is also partially related to corpus imbalance. We have seen in the previous section that sharp frequency increases correspond to significant changes in grammatical profiles, and that this information can be exploited by our method to detect changing words. However, frequency change can be an unfaithful indicator of meaning change. This is the case, for example, for the German words '*Lyzeum*' ('LYCEUM'; ranked 1st out of 48), and '*Truppenteil*' (a 'UNIT OF TROOPS'; ranked 11th), and for the Latin word '*jus*' (a 'RIGHT', the 'LAW'; ranked 4th out of 40).

False negatives, on the other hand, are words whose semantic change is not reflected in changes in grammatical profile. The German word '*ausspannen*' ('TO REMOVE', 'TO UNCLAMP') is used across the 19th and 20th century only in its infini-

[7] We are aware that the current separation of the Latin corpus into two time periods can be controversial. Still, we follow the splits defined by the SemEval 2020 organisers (Schlechtweg et al., 2020) for consistency and comparability with prior work.

tive form, so our method assigns it a relatively low change score (23rd out of 48). Most of the occurrences in the 19th century, however, are literal usages of the word (e.g., *die Pferde ausspannen*, *to unhitch the horses*), whereas in the (second part of the) 20th century the novel metaphorical usage of the word (e.g., *für fünf Minuten ausspannen*, *to relax for five minutes*) is the most frequent one. Another example of a German word whose novel metaphorical sense remains undetected (ranked 31st) is '*Ohrwurm*' ('EARWORM'): the grammatical profile of this word remains stable (except for the accusative case becoming slighlty more frequent), but the word acquires the meaning of *catchy song*, or *haunting melody*. Similarly, the singular usages of the Latin word '*pontifex*' increase from 63% to 83%, signalling the semantic narrowing of the word occurred in medieval Latin (from a 'BISHOP' to the 'POPE'), but the case distribution remains similar; this results in a rather low change score (ranked 22nd out of 40). The last two examples show that taking the maximum distance across categories (see 4.3) is a correct strategy, yet sometimes the changes in that grammatical category are still insufficient for our method to detect change.

7 Category importance

In this Section, we conduct an additional experiment to find out which grammatical categories are most related to semantic change. To this end, we train logistic regression classifiers for binary classification using English, German, Latin, Swedish and Italian data. The classifier features are cosine distances between frequency vectors of each particular category from different time bins. Before fitting the classifier, each feature is independently zero-centered and scaled to the unit variance. Then, regression coefficients are estimated for each feature: we consider positive weights as an indication of usefulness of a feature for classification. The outcome of this analysis is shown in Table 3. We list English nouns and verbs separately since the SemEval'20 dataset explicitly annotates part-of-speech tags for the English target words. This is not the case for the other languages in this dataset.

In line with the results presented in Section 5, Swedish and Italian classifiers yield the highest accuracy and F-score. Latin, a highly inflectional language, has by far the largest set of categories contributing positively to semantic change detection (interestingly, excluding syntax). English, a highly analytical language, is on the other end of the spectrum.

Language	Top categories	Accur.	F1
English nouns	number	0.576	0.523
English verbs	verb form, syntax	0.750	0.733
German	number, syntax, gender	0.542	0.541
Swedish	syntax, mood, voice, definiteness, number	0.839	0.797
Latin	voice, number, degree, case, gender, mood, aspect, person, tense	0.650	0.649
Italian	number, tense, syntax	0.778	0.723

Table 3: Categories with positive weights in binary classifiers of semantic change (logistic regression). 'Syntax' stands for dependency relation to the syntactic head of the word. Evaluation scores are calculated on the train data, F1 is macro-averaged.

Additionally, we estimate the relative importance of morphosyntactic categories by calculating the Spearman's rank-correlation of their respective cosine distance values (across all target words) with the gold semantic change rankings. In other words, we single out each category, e.g. verbal mood, and test whether diachronic change in its frequency distribution is correlated with manually annotated semantic change scores.

In Table 4, we show the categories with statistically significant ($p < 0.05$) correlations for each language and dataset. In English, as expected given its analytical nature, only changes in syntactic roles yield such a correlation; other categories are either non-existent in this language, or are not linked to semantic change strongly enough. For an inflection language such as Latin, number and adjectival degree are highly predictive (the latter is arguably because Latin has the highest ratio of adjectives among all SemEval 2020 Task 1 datasets: about 20%). Not surprisingly for a synthetic language, the morphological categories of number and case show strong correlations for Russian. In the case of the larger time gap between pre-Soviet and post-Soviet periods (Russian 3), syntactic relationships also become a good predictor.

What *is* surprising, however, is that changes in gender are also correlated with semantic change in the Russian case. This result is hard to inter-

	Number	Mood	Degree	Gender	Case	Syntax
English	-	-	-	-	-	0.331
German	-	-	-	-	-	-
Latin	0.304	-	0.301	-	-	-
Swedish	0.402	0.397	-	-	-	-
Russian 1	-	-	-	0.218	0.196	-
Russian 2	-	-	-	0.231	0.324	-
Russian 3	0.246	-	-	0.218	0.327	0.279

Table 4: Spearman rank correlations between diachronic grammatical profile distances for different categories and manually annotated semantic change estimations. '-' stands for no significant correlation.

pret, since grammatical gender is a lexical feature of Russian nouns and does not change from occurrence to occurrence; even diachronically, such cases are quite rare. The reason for this is slightly erroneous morphological tagging: our tagger mixes up homographic inflected forms, which abound in Russian, and assigns feminine gender to masculine nouns, and vice versa. The reliance on the tagger performance can be seen as a limitation of our grammatical profiling approach. However, the existence of the correlation hints that these errors are not entirely random, and their frequency is influenced by word usage: gender is ambiguous only in certain case and number combinations, and the frequency of these combinations seems to change diachronically. For example, for the form '*cheki*' ('cheques/grenade pin'), the masculine lemma licenses the accusative plural reading, while the feminine lemma licenses the genitive singular reading. Thus, even the tagger errors are in fact informative.

Interestingly, for German, no single category changes are significantly correlated with semantic change. This is in line with our weak—although still higher than the count-based baseline—results for German described above, but is somewhat surprising, given the fusional nature of the language, with its rich spectrum of inflected word forms.[8] Some peculiarities of the employed tagger model might be responsible for this finding, which should be further tested and explained in future work.

8 Conclusion

Semantic change is inextricably tied to changes in the distribution of morphosyntactic properties of words, i.e. their grammatical profiles. In this paper, we showed that tracking these changes is enough to build a semantic change detection system which, without access to any lexical semantic information, consistently outperforms count-based distributional semantic approaches to the task. Grammatical profiling yields surprisingly good evaluation scores across different languages and datasets, without any language-specific tuning. For Latin, a language with rich morphology, our methods even establish a new SOTA in Subtask 2 of SemEval'20 Task 1.

These results indicate that grammatical profiling cannot compete with state-of-the-art methods based on large pre-trained language models, since they have the potential to encode both semantics and grammar. Yet reaching the highest possible scores on the task was not our goal. Instead, the aim of our study was to demonstrate that more attention should be paid to the relation between morphosyntax and semantic change. Whether morphosyntactic and semantic features are complementary and can be successfully combined is a interesting question to be addressed in future work.

We performed an extensive quantitative and qualitative analysis of our semantic change detection methods, showing that profiling yields interpretable results across several languages. Nevertheless, we still lack an understanding of some aspects of the interaction between semantics and morphosyntax. Finding the reasons behind the relatively poor performance on some datasets, e.g. German, is an important direction for future studies.

Another interesting question is how to incorporate full dependency trees into grammatical profiles, rather than only dependency relations to the syntactic head of a word. This is particularly important for analytical languages, where grammatical markers are presented in more than one word, such as with English verb mood and aspect. Moreover, dependency structure can be crucial for languages from families other than the Indo-European, e.g. to take into account detached counters in Japanese or plural markers in Yoruba.

In light of our experimental results, we argue that grammatical profiling should become one of the standard baselines for semantic change detection.

Acknowledgements

We thank the anonymous CoNLL-2021 reviewers for their helpful comments. This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grants 770299 (NewsEye), 825153 (EMBEDDIA), and 819455 (DREAM).

[8] We computed correlations for German nouns and verbs separately, but did not find any significant correlation either.

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020) CEUR Workshop Proceedings (CEUR-WS.org)*.

Pierpaolo Basile, Annalina Caputo, Seamus Lawless, and Giovanni Semeraro. 2019a. Diachronic analysis of entities by exploiting Wikipedia page revisions. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 84–91, Varna, Bulgaria. INCOMA Ltd.

Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2019b. Kronos-it: a Dataset for the Italian Semantic Change Detection Task. In *CLiC-it*.

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.

Dagmar Divjak and Stefan Th Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2:23J60.

Hanne M Eckhoff and Laura A Janda. 2014. Grammatical profiles and aspect in old church slavonic. *Transactions of the Philological Society*, 112(2):231–258.

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.

Stefan Th Gries and Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New directions in cognitive linguistics*, 57:75.

Stefan Th Gries and Naoki Otani. 2010. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, 34:121–150.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Patrick Hanks. 1996. Contextual dependency and lexical sets. *International journal of corpus linguistics*, 1(1):75–98.

Martin Hilpert and Stefan Th Gries. 2008. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

Hans Henrich Hock and Brian D Joseph. 2019. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Walter de Gruyter GmbH & Co KG.

Michael Hoey. 2005. *Lexical Priming: A New Theory of Words and Language*. Routledge.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Laura A Janda and Olga Lyashevskaya. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian. *Cognitive linguistics*, 22(4):719–763.

Patrick Juola. 2003. The time course of language change. *Computers and the Humanities*, 37(1):77–96.

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.

Andrey Kutuzov and Lidia Pivovarova. 2021a. RuShiftEval: a shared task on semantic shift detection for Russian. *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Andrey Kutuzov and Lidia Pivovarova. 2021b. Three-part diachronic semantic change dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Julia Kuznetsova. 2015. *Linguistic profiles: Going from form to meaning via statistics*. Walter de Gruyter GmbH & Co KG.

Ronald W Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Matthew 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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Anastasiia Ryzhova, Daria Ryzhova, and Ilya Sochenkov. 2021. Detection of semantic changes in Russian nouns with distributional models and grammatical features. *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

Appendix

A Model predictions

Table 5 shows the top 10 ranked words for each of the target languages according to the semantic change score of our best model. Because three time bins are available for Russian, we show the change score estimated for the interval between first and second (1-2), second and third (2-3), as well as first and third (1-3) periods.

English	German	Latin	Swedish	Italian	Russian 1-2	Russian 2-3	Russian 1-3
gas	Lyzeum	imperator	bröllop	lucciola	blagodarnost	polosa	ambitsia
chairman	vorweisen	beatus	studie	palmare	vek	vek	nalozhenie
rag	Schmiere	regnum	motiv	tac	sobrat	zhest'	ponedelnik
stab	zersetzen	jus	krita	unico	vyzov	favorit	vyzov
ball	verbauen	adsumo	konduktör	pacchetto	brat	sobrat	blin
lass	Eintagsfliege	potestas	annandag	cappuccio	jubiley	ambitsia	polosa
prop	beimischen	licet	aktiv	egemonizzare	ambitsia	nalozhenie	khren
tip	Engpaß	sensus	granskare	brama	khren	jubiley	uglevodorod
record	artikulieren	nobilitas	bolagsstämma	campanello	uglevodorod	lishenie	lishenie
plane	voranstellen	sacramentum	färg	piovra	ponedelnik	blin	chastitsa

Table 5: The top 10 rankings obtained with our best method for all the target languages. The topmost word is the one with the highest assigned change score. Russian words are transliterated from Cyrillics to Latin script.