

Article

Identifying Source-Language Dialects in Translation

Sergiu Nisioi ^{*,†} , Ana Sabina Uban ^{*,†}  and Liviu P. Dinu 

Human Language Technologies Center, Faculty of Mathematics and Computer Science, University of Bucharest, Academiei 14, 010014 Bucharest, Romania; ldinu@fmi.unibuc.ro

* Correspondence: sergiu.nisioi@unibuc.ro (S.N.); ana-sabina.uban@unibuc.ro (A.S.U.)

† These authors contributed equally to this work.

Abstract: In this paper, we aim to explore the degree to which translated texts preserve linguistic features of dialectal varieties. We release a dataset of augmented annotations to the Proceedings of the European Parliament that cover dialectal speaker information, and we analyze different classes of written English covering native varieties from the British Isles. Our analyses aim to discuss the discriminatory features between the different classes and to reveal words whose usage differs between varieties of the same language. We perform classification experiments and show that automatically distinguishing between the dialectal varieties is possible with high accuracy, even after translation, and propose a new explainability method based on embedding alignments in order to reveal specific differences between dialects at the level of the vocabulary.

Keywords: translationese identification; dialectal varieties; machine translation; feature analysis



Citation: Nisioi, S.; Uban, A.S.; Dinu, L.P. Identifying Source-Language Dialects in Translation. *Mathematics* **2022**, *10*, 1431. <https://doi.org/10.3390/math10091431>

Academic Editors: Florentina Hristea, Cornelia Caragea and Jakub Nalepa

Received: 31 December 2021

Accepted: 24 March 2022

Published: 24 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Computational approaches in Translation studies enforced the idea that translated texts (regarding translations, we will use the abbreviation SL to define source language and TL for target language) have specific linguistic characteristics that make them structurally different from other types of language production that take place directly in the target language. Translations are considered a sub-language (*translationese*) of the target language [1–3] and studies [4–9] imply that translated texts have similar characteristics irrespective of the target language of translation (*translation universals*). Universals emerge from psycholinguistic phenomena such as *simplification*—“the tendency to make do with less words” in the target language [10,11], *standardization*—the tendency for translators to choose more “habitual options offered by a target repertoire” instead of reconstructing the original textual relations [5], or *explicitation*—the tendency to produce more redundant constructs in the target language in order to explain the source language structures [12,13].

In addition, translated texts also exhibit patterns of *language transfer* or *interference* [14]—a phenomenon inspired by second-language acquisition, indicating certain source-language structures that get transferred into the target text. Using text mining and statistical analysis, researchers were able to identify such features [3,15,16] up to the point of reconstructing phylogenetic trees from translated texts [17,18].

Investigations with respect to translationese identification have strong potential for improving machine translation, as [19,20] pointed out for statistical machine translation, and more recently [21] showed that the effect of translationese can impact the system rankings of submissions made to the yearly shared tasks organized by The Conference of Machine Translation [22]. Ref. [23] show that a transformer-based neural machine translation (NMT) system can obtain better fluency and adequacy scores in terms of human evaluation, when the model accounts for the impact of translationese.

While the majority of translation research has been focused on how different source languages impact translations, to our knowledge, little research has addressed the properties of the source language that stem from dialectal or non-native varieties, how and to what degree they are preserved in translated texts.

In our work, we intend to bring this research question forward and investigate whether dialectal varieties produce different types of translationese and whether this hypothesis holds for machine-translated texts. We construct a selection of dialectal varieties based on the Proceedings of the European Parliament, covering utterances of speakers from the British Isles and equivalent sentence-aligned translations into French. Our results imply that *interference* in translated texts does not depend solely on the source language (SL), rather, different language varieties of the same SL can affect the final translated text. Translations exhibit different characteristics depending on whether the original text was produced by speakers of different regional varieties of the same language.

To our knowledge, this is the first result of its kind extracted from a stylistically uniform multi-author corpus using principles of statistical learning and our contribution can be summarized as follows:

1. We build and release an augmented version of the EuroParl [24] corpus that contains information about speakers' language and place of birth.
2. We investigate whether the dialectal information extracted is machine-learnable, considering that the texts in the European Parliament go through a thorough process of editing before being published,
3. Using sentence-aligned equivalent documents in French, we analyze to what degree dialectal features of the SL are preserved in the translated texts. Additionally, we employ a transformer architecture to generate English to French translations [25] and investigate whether dialectal varieties impact the machine-translated texts.
4. For each dialectal variety we fine-tune monolingual embeddings and align them to extract words whose usage differs between varieties of the same language. We analyze and interpret the classification results given the sets of aligned word pairs between different classes of speakers.

We perform a series of experiments in order to achieve our research goals. In order to observe the differences between our classes and to gain additional insights based on the obtained classification performance we compare several different solutions: we use a variety of linguistic features as well as several model architectures, including logistic regression log-entropy-based methods and neural networks. For the second stage of our experiments, we choose state-of-the-art methods used in lexical replacement tasks (including lexical semantic change detection [26], and bilingual lexicon induction [27]), based on non-contextual word embeddings and vector space alignment algorithms, in order to produce a shared embedding space which allows us to more closely compare word usage across the different varieties. We publicly share our code and detailed results (https://github.com/senisioi/dialectal_varieties, accessed on 30 November 2021), as well as the produced dataset.

2. A Corpus of Translated Dialectal Varieties

Our corpus is extracted from the collection of the proceedings of the European Parliament, which contains edited transcriptions of member's speeches together with their equivalent translations (The standard work-flow in EuroParl is to transcribe and edit the speech, and then to send the texts for translation [28]) made by native speakers into French. The core is based on a combination between multilingual sentence-aligned language annotated corpus released by [29,30]. To further extract the dialectal varieties, we had to re-crawl the EuroParl website, match each session to the existing corpus, and to disambiguate and match the utterances with the correct speaker (There is no convention on how the speaker names are written on the EuroParl website, so we had to disambiguate them using a semi-manual process). We could only do this process for sessions after 1999 that are crawl-able on the current website. After matching each utterance to the correct speaker, we crawled the speaker place of birth from their personal page and traced it using geotagging to the actual state, country or region. At the same time, we annotated the equivalent French translations with the metadata extracted for the source language. We are aware, however, that the place of birth does not necessarily imply dialectal information. The same can be said about

the representative country where there could be multiple official languages. We ignore speakers for whom the location is missing or who were invited as guests in the Parliament (e.g., speeches by the 14th Dalai Lama). We also acknowledge that speakers sometimes employ external teams to write their official speeches and that EuroParl transcriptions are strongly edited before being published in their final form.

Statistics regarding the corpus are rendered in Table 1 where we notice the group of speakers from Wales and the ones with Unknown source are underrepresented with a small amount of data, therefore we decide to ignore these categories from our experiments.

Table 1. Extracted statistics: mean and standard deviation sentence length, and type-token ratio (TTR). Both TTR and average sentence length are statistically significant under a permutation test, with p -value < 0.01 for original English documents from Scotland, pair-wise for: England vs. Scotland and Ireland vs. Scotland.

Regional Variety	Sentences	English Originals			French Translations		
		Mean	std	TTR	Mean	std	TTR
Scotland	15,646	26.02	13.59	3.87	29.99	16.14	4.27
England	60,179	26.40	13.82	1.76	30.01	16.07	1.95
Ireland	31,443	26.09	13.06	2.44	29.72	15.26	2.73
Wales	2466	25.76	12.30	8.92	29.33	14.71	10.04
Unknown	6607	25.84	13.19	5.79	29.19	15.26	6.59

We render the sentence length mean and standard deviation, and the overall type/token ratio to highlight shallow information with respect to the lexical variety of texts. At a first glance, original English texts appear to have shorter sentences and smaller type-token ratios (smaller lexical variety) compared to their French counterparts. Rich lexical variety in translated texts has been previously linked [31,32] to the *explicitation* phenomenon.

In addition, we construct a machine translated corpus of French sentences using a transformer-based [33] neural machine translation trained in a distributed fashion using the *fairseq-py* library (<https://github.com/pytorch/fairseq>, accessed on 30 November 2021). Ref. [25] report state-of-the-art results on English-to-French translation for the WMT'14 dataset [34]. We acknowledge that the parallel data used for training the transformer contains also the EuroParl v7 (<http://statmt.org/europarl/>, accessed on 30 November 2021) [24] among Common Crawl, French-English 10⁹, News Commentary, and the United Nations Parallel Corpora. It is likely that the model has already “seen” similar data during its training which could probably lead to more fluent automatic translations. In our work we aim to see whether the dialectal information influences the machine-translated generated output.

3. Experimental Setup

In our experiments, we use statistical learning tools to observe the structural differences between our classes. Our aim is to minimize any type of classification bias that could appear because unbalanced classes, topic, parliamentary sessions, and specific user utterances, with the purpose of exposing grammatical structures that shape the dialectal varieties. To minimize the effect of uniform parliamentary sessions, we shuffle all the sentences for each dialectal variety. The data are split into equally-sized documents of approximately 2000 tokens to ensure the features are well represented in each document, following previous work on translationese identification [3,15,35,36]. Splitting is done by preserving the sentence boundary, each document consisting of approximately 66 sentences. Larger classes are downsampled multiple times and evaluation scores are reported as an average across all samples of equally-sized classes. To compare the classification of the same documents across different languages, we construct a test set of 40 sentence-aligned chunks. When not mentioned otherwise, we report the average 10-fold cross-validation scores across multiple down-samplings. We illustrate the stages performed from data collection to pre-processing and classification in Figure 1.

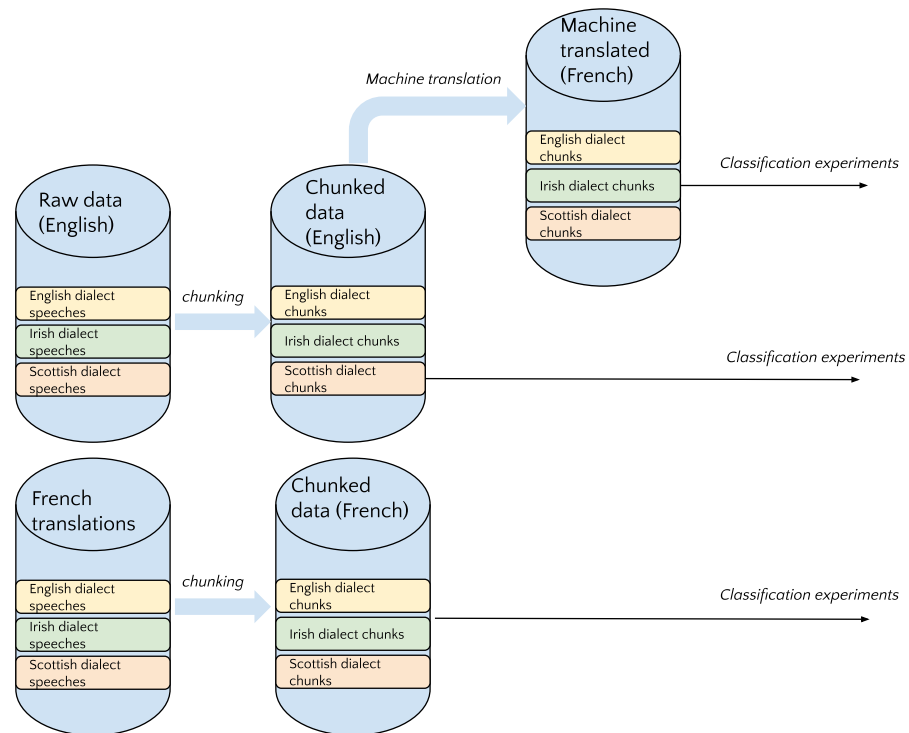


Figure 1. Data collection and pre-processing.

We adopt the log-entropy weighting scheme to vectorize documents, since log-entropy has been extensively used in information retrieval [37] and its purpose is to reduce the importance of high frequency features, and increase the weight for the ones that are good discriminants between documents. From our observations, this type of weighing scheme achieved the best results and it has been previously used to improve classification scores for medium-sized documents [38].

We compute the entropy for a feature i by the following formula:

$$g_i = 1 + \sum_{j=1}^{\mathcal{N}} \frac{p_{ij} \log 1 + p_{ij}}{\log \mathcal{N}} \quad (1)$$

where \mathcal{N} is the number of documents in the corpus and p_{ij} is defined by the normalized frequency of term i in document j .

To normalize the p_{ij} values, we divide by the global frequency in the corpus:

$$p_{ij} = \text{tf}_{ij} / \left(\sum_{j=1}^{\mathcal{N}} \text{tf}_{ij} \right)$$

The final weight of a feature is computed by multiplying the entropy with the log weight:

$$\text{logent}_{ij} = g_i \log(\text{tf}_{ij} + 1) \quad (2)$$

We apply this feature weighting in combination with a logistic regression classifier with liblinear optimizer [39] and l_2 penalty. Similar models based on BoW representations have been successfully used in tasks with small amounts of data for detecting the dialectal or native language variety of a speaker [40,41].

Features

Function words (FW) consist of conjunctions, preposition, adverbs, determiners, auxiliary and modal verbs, pronouns, qualifiers, and question words. Some function words are also part of the *closed class* because languages rarely introduce changes (historically) in this vocabulary subset [42,43]. They possess primarily a grammatical meaning and their frequency in a document reflects syntactical constructions that are particular to style. This word category has a long history of usage, being the primary features of analysis for the identification of authorship, translationese, or dialectal varieties [15,36,44,45] since they tend to be less biased by the topics or content covered in the texts. Ref. [46] argue that different brain functions are used to process the closed class and the open class of words.

Pronouns are a subclass of function words that have been previously tied to *explicitation* [12,47–49], translators showing an increased usage of personal pronouns. In our experiments, we observed that these features play a more important role in distinguishing human- and machine- translated dialectal varieties than original English texts.

Part of Speech n-grams are useful for capturing shallow syntactic constructs. We extract PoS bigrams and trigrams from our texts using the latest version of spaCy 3.2 [50] transformer models for English based on RoBERTa [51] and the French model based on CamemBERT [52] (Transformer-based models latest release <https://spacy.io/usage/v3-2> accessed on 30 November 2021). We insert an additional token in the PoS list (SNTSEP) that indicates whether the next token is sentence end, in this way we hope to cover syntactic constructions that are typical to start/end the sentences. Unlike the previous features and due to the sheer size of possible combinations, PoS tag n-grams have a tendency to be sparsely represented in documents. This may lead to accurate classifications without exposing an underlying linguistic difference between the classes. To alleviate this, we have capped the total number of n-grams to 2000 and further conducted experiments with a list of 100 PoS n-grams curated using a Recursive Feature Elimination method [53] for both English and French corpora.

Word n-grams including function and content words. For each text replace all the entities discovered by spaCy with the corresponding entity type, including proper nouns that could potentially cover places, locations, nationality, and countries, but also numeric entities such as percentages and dates which could bias the classification. Using this feature set, we hope to understand how much the semantics of the texts, the usages and choices of content words, and potentially the topics addressed by different speakers contribute to separating between language varieties. This feature set is biased by the topics that are repeatedly addressed by different groups, given their regional interests in Scotland, Ireland or England. Furthermore, the feature set can potentially introduce sparsity and in order to alleviate this, we enforce a strict limit and cap the total number of allowed n-grams to the most frequent 300. We experimented with smaller numbers of word n-grams (100, 200) and observed that the majority of features were comprised of function word combinations and expressions such as: “I would like”, “member states”, “SNTSEP However”, “we should”, “the commissioner”, “mr president I”. We also experimented with larger numbers of n-grams: 400, 500 which easily achieved perfect accuracy due to the topic information embedded in the higher dimension.

Convolutional Neural Networks (CNN) are able to extract relevant semantic contexts from texts by learning filters over sequences of representations. We apply convolutions over word sequences (including all words in the text) using one 1-dimensional convolutional layer of 10 filters, with filter size 3, followed by a max pooling and an output layer.

4. Results

Table 2 contains the full set of classification results that compare the logistic regression log-entropy-based method with different features and the convolutional neural networks' results.

Content-dependent methods based on word n-grams and convolutional networks stand out as having the largest scores (above 0.9 for English and French) even though

proper nouns and entities have been removed beforehand. This is an indicator that speakers belonging to these regions are classified based on different specific topics addressed in the European Parliament. Translated dialects appear to be easier to separate with word n-grams. Manually analysing the entity tagging and removing process for French, we could observe that markers of location (*irlandais, britannique*) were not completely removed from the texts. Content words as features for text classification are less relevant than content-independent features to test linguistic hypotheses. We can also observe here that CNNs obtain slightly lower scores for this task, possibly due to the small size of the dataset and the 2000-word length of the input classification documents.

Table 2. Average F_1 scores for distinguishing dialectal varieties and their translations into French. Values in bold indicate the best accuracy obtained using topic-independent features. The feature set 100 PoS En are the most representative n-grams for classifying original English documents and similarly 100 PoS Fr, for the human-translated French documents. Word n-grams (limited to a maximum of 300 most frequent) and convolutional neural networks (CNN) are covering content words and are biased by topic, therefore we do not highlight the classification scores of the two methods.

	Feature	En vs. Ir	En vs. Sc	Sc vs. Ir	3-Way
French translations	function words	0.84	0.87	0.78	0.71
	pronouns	0.82	0.80	0.72	0.66
	PoS n-grams	0.91	0.87	0.81	0.76
	100 PoS En	0.8	0.76	0.71	0.59
	100 PoS Fr	0.78	0.76	0.62	0.59
	Word n-grams	0.95	0.89	0.89	0.84
	CNN	0.95	0.8	0.95	0.84
French machine transl.	function words	0.88	0.84	0.81	0.72
	pronouns	0.85	0.85	0.74	0.71
	PoS n-grams	0.96	0.92	0.87	0.85
	100 PoS En	0.78	0.79	0.72	0.62
	100 PoS Fr	0.83	0.73	0.77	0.66
	Word n-grams	0.96	0.91	0.87	0.85
	CNN	0.94	0.9	0.91	0.89
English originals	function words	0.9	0.91	0.85	0.8
	pronouns	0.63	0.76	0.69	0.57
	PoS n-grams	0.91	0.87	0.91	0.83
	100 PoS En	0.88	0.85	0.86	0.78
	100 PoS Fr	0.82	0.71	0.77	0.64
	Word n-grams	0.91	0.89	0.92	0.83
	CNN	0.94	0.91	0.93	0.95

The magnitude of logistic regression coefficients can give an estimate of feature importance for classification corresponding to each class. A manual inspection (The supplementary material contains the full set of features ordered by importance.) of the most important classification features from Table 3 shows that certain debates have (key)words acting as good discriminators between our classes. The topics hint towards political tensions with respect to Northern Ireland, fishing rights, and state-policies in the region.

Table 3. The top most relevant word n-grams in binary text classification scenarios.

Experiment	Function Words with High Discriminatory Value
En-Ir (en)	energy, regard, must, sure, research, policy, nuclear, food, s, recent, children, peace, farmers
En-Sc (en)	fisheries, we have, in writing, i voted, your, aid, human rights, want, research, policy
Ir-Sc (en)	he, fisheries, s, regard, people, want to, treaty, sure, report, peace, want, hope
En-Ir (fr)	nord, regions, secteur, amendement, trois, de l, son, enfants, traite, industrie, donc, nous
En-Sc (fr)	ai vote, regions, escrit, votre, vote, processus, tres, secteur, mesures, mais, est un, reforme
Ir-Sc (fr)	ecrit, vote, traite, ont, de m, programme, ait, deja, du nord, rapport, certains, assemblee

From the total most frequent words in the corpus, several function words appear to have a high importance for English: *regard, must, we, to, s, you, he, sure*; and French: *de l, son, donc, nous, votre, tres, mais, ont, de m, deja*. Table 3 contains several words marked as important in separating different classes, where we can observe that dialectal varieties are potentially influenced by function word and more specifically pronoun usage.

Topic-independent features that include function words, pronouns, and PoS n-grams yield relatively high scores for original English texts, indicating that *the place of birth is a valid indicator of dialectal information* for our particular dataset. The translations show significantly lower scores, but still above 0.8 for 3-way classification on both human and machine-translated versions. These features are an indicator of the grammatical structures that transfer from source to target language and we highlight in boldface the highest scores for each. PoS n-grams tend to achieve the highest classification scores among all experiments, when taking into account the sparse high dimensionality of each classification example. When restricting the overall set to the 100 most frequently occurring PoS n-grams, unsurprisingly, the overall scores drop by 10%. While taking into account this drop, we can still observe a fair degree of pair-wise separation between the classes. Furthermore, the curated list of PoS n-grams is language independent and we used the list extracted from the French corpus to classify the annotated documents in English and vice-versa. Original English can be separated with an F_1 score ranging from 0.71 to 0.82 when using PoS n-gram features extracted from the French data. A similar phenomenon occurs for translated French documents can be separated with an F_1 score ranging from 0.71 to 0.8 using PoS n-grams extracted from the English data. This is a clear indicator that shallow syntactic constructs that are specific to each class are transferred during translation into the resulting documents.

With respect to machine-translated texts, it appears that all the classification experiments achieve slightly higher scores than the equivalent human-translated data. Since machine-generated translations are more rudimentary, it could very well be that the original dialectal patterns are simply amplified or mistranslated into the target language, thus generating the proper conditions to achieve statistical separation between the classes.

Pronouns show the opposite result on both machine and human translation outputs. Original English dialectal varieties are weakly classifiable using pronouns - England vs. Scotland achieving at best a 0.74 score. Pronouns appear to be better markers of separation in translated texts, these words being markers of *explicitation*, as previous research hypothesised [12,47,48]. The results show that pronoun distribution in translation accentuates original patterns of the texts, mainly due to explicitation, a phenomenon that appears to be mimicked by machine-translation systems trained on human translations. For example, the most important pronouns in English classification are: *we, this, anyone, anybody, several, everyone, what*. For French we observe several different personal pronouns of high importance: human translations: *nous, l, j, les, la, m, je*; and machine translation: *nous, j, la, en, l, qui, m, quoi que, celles*.

5. Classification Analysis

Given the high accuracy obtained on both English and French version of the corpora using PoS n-grams, we render in Table 4 the average confusion matrices across all cross-validation epochs for these results. On the English side of the table, the largest confusion is between speakers from England and Scotland, while on the French translations, the confusions are more uniformly distributed. From this result, it becomes clear that translations preserve certain syntactic aspects of the source-language dialect, although the differences between the classes are slightly lost in the process.

We have also constructed a comparable train/test split designed with the same documents in both English and French classification scenarios. The first four rows of Table 5 render the percentage of documents from the test set classified with the same label in both English and French human-translated versions. The process is similar to computing an accuracy score of the French classifications given the English equivalent as the gold

standard. The result gives us an estimation of the number of test documents that have the same distinguishing pattern w.r.t a feature type. From Table 5 we confirm the fact that pronouns have different roles in translated texts—showing little overlap between the predictions on the French test set vs. the English equivalent. Function words and PoS n-grams have slightly higher overlap percentages, again, proving that certain grammatical patterns transfer from the dialectal varieties onto the French translation. Word n-grams and CNNs share the highest prediction similarities between the two test sets. We believe this is to a lesser degree due to source-language transfer, rather it corroborates that topics addressed in the debates determine similar classification patterns across languages.

Table 4. Comparison of average confusion matrices for original English and French classification experiments using PoS n-grams features.

		En	Ir	Sc
Fr	En	80	10	10
	Ir	7.5	80	12.5
	Sc	5	5	90
En	En	85	5	10
	Ir	4.5	91	4.5
	Sc	0	5	95

Table 5. The percentage of documents from the test set classified with the same label in both English and French translated versions. The last row compares the 3-way classification similarities between dialectal classification of documents from human and machine translated output.

	Function Wds.	Pronouns	PoS n-Grams	wd. n-Grams	CNN
3-way	64.2%	44.2%	75%	82.5%	79%
England vs. Ireland	81.3%	58.8%	88.75%	95%	90%
England vs. Scotland	80%	65%	87.5%	91%	88.2%
Ireland vs. Scotland	73.8%	65%	76.3%	91%	87.5%
3-way Human vs. MT	67%	70%	78%	84%	85%

For French human vs. machine translation, we present only the 3-way classification similarities (last row in Table 5), since the pair-wise versions have similar values. In this case we observe the divergence between human- and machine- generated translations in terms of different features. The output produced by the transformer-based NMT system does not resemble typical human language in terms of function words distribution, as seen in the low amount of classification overlap (67%). However, the machine appears to do better at imitating *translationese explicitaion*, given the higher importance of pronouns in classification (0.71 F_1 score and 70% overlap between human and machine translation classifications). Similarly, the ability of PoS n-grams to distinguish English varieties with a 0.83 F_1 score and with 78% similarity to classification human translation, indicates that dialectal syntactic structures are reasonably preserved in both human- and machine- translation. Content-wise, both CNNs and word n-grams lead to similar classification patterns on the test set (84% overlap and 0.95 avg. F_1 score). Overall, the dialectal markers yield prediction correlations between machine- and human- generated translations.

6. Words in Context

Using weights learned by a linear model to infer feature importance can be useful for interpreting the behavior of the classifier and explain some of the underlying linguistic mechanisms that distinguish between the classes considered, in our case - language varieties. Nevertheless, this method has its limits: based on feature importance in our classifier we are essentially only able to find differences between the word distributions of two corpora, in terms of frequencies. In order to gain more insight into the phenomena behind the aspects

of language that make different varieties of English distinguishable with such high accuracy, we propose a novel method for feature analysis based aligned word embedding spaces in order to identify word pairs which are used differently in two corpora to be compared.

Aligned embedding spaces have previously been exploited in various tasks in computational linguistics, from bilingual lexicon induction [27], to tracking word sense evolution [54], and identifying lexical replacements [55,56]. We also propose using word embedding spaces for finding lexical replacements, this time across dialectal varieties. By training word embeddings on two different corpora, and comparing their structural differences, we can go beyond word frequency distribution, and look into the specific lexical differences that distinguish word usage between texts in the two classes. If the feature weight method could tell us, for example, that English speakers use *maybe* more than the Irish do, the embedding alignment based method should be able to show exactly how that word is used differently, what word Irish speakers use instead in their speech, in the form of a word analogy: where English speakers say *maybe*, Irish speakers say *X*.

Word Embedding Alignment

The algorithm for identifying pairs of words which are used differently in the two corpora (in our case, corresponding to different dialectal varieties) consists of the following steps:

Separately train word embeddings for each of the two corpora. We train word2vec on each of our datasets, using standard hyperparameters and embedding dimension 100. We use Wikipedia pre-trained embeddings to initialize the weights which we further fine-tune on our data.

Obtain a shared embedding space, common to the two corpora. Vectors in two separately trained embedding spaces are not directly comparable, so an alignment algorithm is necessary to obtain a shared embedding space. To align the embedding spaces, a linear transformation is applied to one of the spaces, such that the distance between a few seed word pairs is minimized (which are assumed to have the same meaning in both spaces/corpora). We use a linear regression method, and a random sample of the first 60% most frequent words from our vocabulary as seed—minimizing their pairwise distance will constitute the objective of training the linear regression model to obtain a transformation matrix.

Identify misaligned word pairs, where the nearest neighbor (based on cosine distance) of a word in the first corpus is not the same word in the second corpus, and extract the actual nearest neighbor. The resulted misaligned word pairs constitute words which are used differently in the two corpora. We also define a score of the misalignment, measuring strength of the difference in usage in different corpora for a word pair. The higher this score, the most significant the usage difference between the two corpora. Scores can range from 0 to 1, where scores of zero would be assigned to word pairs that show an identical usage pattern across the two corpora. Details on how this score is computed are described in Algorithm 1.

Algorithm 1 Detection of misaligned word pairs between two corpora.

- 1: Given a word w_1 and its corresponding embedding $emb(w_1, C_1)$ in the embedding space of corpus C_1 :
 - 2: Find the word w_2 with embedding representation $emb(w_2, C_2)$ in the embedding space of corpus C_2 such that for any w_i in C_2 , $\text{dist}(emb(w_1, C_1), emb(w_2, C_2)) < \text{dist}(emb(w_1, C_1), emb(w_i, C_2))$
 - 3: Extract (w_1, w_2) as pair with unmatched usage, with the property that $w_1 \neq w_2$
 - 4: $\text{Score}(w_1) = \text{dist}(emb(w_1, C_1), emb(w_1, C_2)) - \text{dist}(emb(w_1, C_1), emb(w_2, C_2))$
-

For each pair of dialects, we train embeddings, perform embedding space alignments and extract nearest neighbors for misaligned word pairs. Table 6 shows some descriptive statistics of the distribution of misalignment scores for all misaligned words in each corpus pair. A higher average misalignment score should point to a bigger difference in patterns

of word usage between two corpora. The ranking of dialectal “similarities” inferred in this way is still maintained after translation into French, although, differences in language usage seem to be reduced after translation.

In Figure 2 we plot the distribution of misalignment scores for all words (including non-misaligned ones) and all dialect pairs, along with French translations. The distribution is skewed, with most words having a score in the vicinity of 0, showing similar usage patterns in the two corpora.

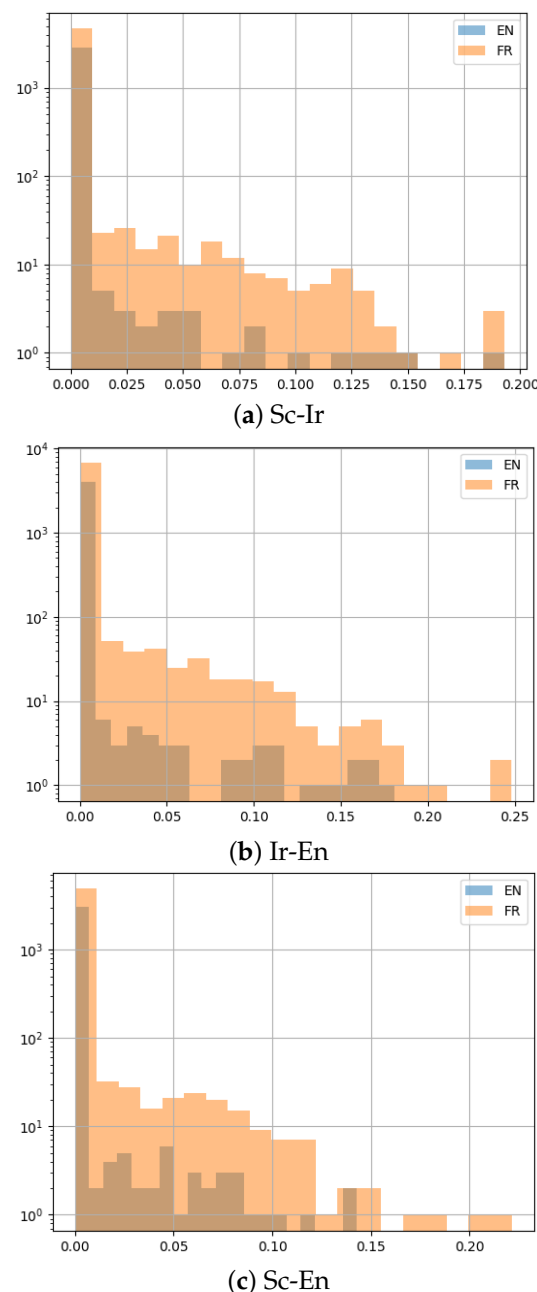


Figure 2. Distribution of misalignment scores across dialect pairs for English and French data sets.

Table 6. Average and standard deviation of misalignment scores for all corpus pairs, in original and translated versions.

Varieties	English		French	
	Mean	Std	Mean	Std
En-Sc	0.049	0.040	0.048	0.040
En-Ir	0.063	0.053	0.058	0.048
Ir-Sc	0.053	0.048	0.050	0.040

We take a closer look at some examples of misaligned words (The full set of aligned words is available in the supplementary material along with their corresponding similarity scores.) that are used differently across corpora in Table 7. The method unveils word pairs that capture differences in topic content between the two corpora, further corroborating that topic contributes to distinguishing between texts written by speakers of different English varieties. Such an example is the pair *Scotland/England*, which captures mentions of proper nouns: in contexts where Scottish speakers say *Scotland*, the English say *England*. The same occurs in the case of *irlandais* and *écossais* for the French translations of Irish and Scottish texts.

More interestingly, the method helps capture an underlying stylistic dimension of content word usage as well, by identifying misaligned pairs of words with the same meaning (synonyms). Content-independent features are traditionally employed in stylistic analyses in order to remove bias from topic. Our analysis shows content words can encapsulate a stylistic dimension as well, and should not be ignored when considering aspects of the language independent from topic.

Table 7. Examples of unmatched embeddings.

Corpora	Word	Nearest Neighbor
En-Sc	England	Scotland
	reply	answer
	but	however
	extremely	very
En-Sc (fr)	aspiration recommandation	ambition proposition
Ir-Sc	plan	program
	she	he
Ir-Sc (fr)	plan irlandais	programme écossais
Ir-En	absolutely keep	perfectly hold
Ir-En (fr)	absolument comprendre	vraiment croire

To express the same concept, the Irish tend to use *plan* where the Scottish say *program*, and the same pattern can be observed in the translated versions of the texts: the French words *plan* and *programme* are nearest neighbors. The same is true for the Irish *absolutely* versus and the English *perfectly*, translated as *absolument* and *vraiment* in French. In addition, several example pairs may still yield an unwanted nearest neighbor, as it is the case for *unless* vs. *if*, *indeed* vs. *nevertheless*. These examples show that a certain threshold must be enforced in order to filter them out. A few examples of function words also stand out, such as *very* and *extremely* that distinguishes Scottish from English speakers. This last pair is also consistent with the feature importance analysis from our logistic regression results.

7. Conclusions

We construct an augmented version of the English-French parallel EuroParl corpus that contains additional speaker information pointing to native regional dialects from the British Isles (We will open-source the data and code for reproducing the experiments). The corpus has several properties useful for the joint investigation of dialectal and translated varieties: it is stylistically uniform, the speeches are transcribed and normalized by the same editing process, there are multiple professional translators and speakers, and the translators always translate into their mother tongue.

Our experimental setup brings forward the first translation-related result (to the best of our knowledge) showing that translated texts depend not only on the source language, but also on the dialectal varieties of the source language. In addition, we show that machine translation is impacted by the dialectal varieties, since the output of a state-of-the-art transformer-based system preserves (or exacerbates, see Table 2) syntactic and topic-independent information specific to these language varieties. W.r.t pronouns, we show that these are discriminating markers for dialectal varieties in both human- and machine-translations (as a source of explicitation), being less effective on original English texts.

We provide a computational framework to understand the lexical choices made by speakers from different groups and we release the pairs of extracted content words in the supplementary material. The embeddings-based method offers promising insights into the word choices and usages in different contexts and we are currently working on filtering aligned pairs and adapting it to phrases.

Author Contributions: Investigation, S.N. and A.S.U.; Methodology, S.N. and A.S.U.; Supervision, L.P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by two grants of the Ministry of Research, Innovation and Digitization, Unitatea Executiva pentru Finantarea Invatamantului Superior, a Cercetarii, Dezvoltarii si Inovarii-CNCS/CCCDI—UEFISCDI, CoToHiLi project, project number 108, within PNCDI III, and CCCDI—UEFISCDI, INTEREST project, project number 411PED/2020, code PN-III-P2-2.1-PED-2019-2271, within PNCDI III.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Ethical Considerations: The data we release with this paper, including speaker information, is publicly available in an electronic format on the European Parliament Website at <https://www.europarl.europa.eu/> (accessed on 30 November 2021).

References

1. Toury, G. *Search of a Theory of Translation*; The Porter Institute for Poetics and Semiotics; Tel Aviv University: Tel Aviv, Israel, 1980.
2. Gellerstam, M. Translationese in Swedish novels translated from English. In *Translation Studies in Scandinavia*; Wollin, L., Lindquist, H., Eds.; CWK Gleerup: Lund, Sweden, 1986; pp. 88–95.
3. Baroni, M.; Bernardini, S. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Lit. Linguist. Comput.* **2006**, *21*, 259–274. [\[CrossRef\]](#)
4. Baker, M. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and Technology: In Honour of John Sinclair*; Baker, M., Francis, G., Tognini-Bonelli, E., Eds.; John Benjamins: Amsterdam, The Netherlands, 1993; pp. 233–252.
5. Toury, G. *Descriptive Translation Studies and beyond*; John Benjamins: Amsterdam, PA, USA, 1995.
6. Mauranen, A.; Kuusimäki, P. (Eds.) *Translation Universals: Do They Exist?* John Benjamins: Amsterdam, The Netherlands, 2004.
7. Laviosa, S. Universals. In *Routledge Encyclopedia of Translation Studies*, 2nd ed.; Baker, M., Saldanha, G., Eds.; Routledge: New York, NY, USA, 2008; pp. 288–292.
8. Xiao, R.; Dai, G. Lexical and grammatical properties of Translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. *Corpus Linguist. Linguist. Theory* **2014**, *10*, 11–55. [\[CrossRef\]](#)
9. Bernardini, S.; Ferraresi, A.; Miličević, M. From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective. *Target. Int. J. Transl. Stud.* **2016**, *28*, 61–86. [\[CrossRef\]](#)

10. Blum-Kulka, S.; Levenston, E.A. Universals of lexical simplification. In *Strategies in Interlanguage Communication*; Faerch, C., Kasper, G., Eds.; Longman: London, UK, 1983; pp. 119–139.
11. Vanderauwera, R. *Dutch Novels Translated into English: The Transformation of a “Minority” Literature*; Rodopi: Amsterdam, The Netherlands, 1985.
12. Blum-Kulka, S. Shifts of Cohesion and Coherence in Translation. In *Interlingual and Intercultural Communication Discourse and Cognition in Translation and Second Language Acquisition Studies*; House, J., Blum-Kulka, S., Eds.; Gunter Narr Verlag: Tübingen, Germany, 1986; Volume 35, pp. 17–35.
13. Øverås, L. In Search of the Third Code: An Investigation of Norms in Literary Translation. *Meta* **1998**, *43*, 557–570. [\[CrossRef\]](#)
14. Toury, G. Interlanguage and its Manifestations in Translation. *Meta* **1979**, *24*, 223–231. [\[CrossRef\]](#)
15. Koppel, M.; Ordan, N. Translationese and Its Dialects. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA, 19–24 June 2011; Association for Computational Linguistics: Portland, OR, USA, 2011; pp. 1318–1326.
16. Rabinovich, E.; Wintner, S. Unsupervised Identification of Translationese. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 419–432. [\[CrossRef\]](#)
17. Rabinovich, E.; Ordan, N.; Wintner, S. Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. *arXiv* **2017**, arXiv:1704.07146.
18. Chowdhury, K.D.; Española-Bonet, C.; van Genabith, J. Understanding Translationese in Multi-view Embedding Spaces. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6056–6062.
19. Kurokawa, D.; Goutte, C.; Isabelle, P. Automatic Detection of Translated Text and its Impact on Machine Translation. In Proceedings of the MT-Summit XII, Ottawa, ON, Canada, 26–30 August 2009; pp. 81–88.
20. Lembersky, G.; Ordan, N.; Wintner, S. Improving Statistical Machine Translation by Adapting Translation Models to Translationese. *Comput. Linguist.* **2013**, *39*, 999–1023. [\[CrossRef\]](#)
21. Zhang, M.; Toral, A. The Effect of Translationese in Machine Translation Test Sets. *arXiv* **2019**, arXiv:1906.08069.
22. Ondrej, B.; Chatterjee, R.; Christian, F.; Yvette, G.; Barry, H.; Matthias, H.; Philipp, K.; Qun, L.; Varvara, L.; Christof, M.; et al. Findings of the 2017 conference on machine translation (wmt17). In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 169–214.
23. Graham, Y.; Haddow, B.; Koehn, P. Statistical Power and Translationese in Machine Translation Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 72–81.
24. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Tenth Machine Translation Summit, AAMT, Phuket, Thailand, 13–15 September 2005; pp. 79–86.
25. Ott, M.; Edunov, S.; Grangier, D.; Auli, M. Scaling Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, 31 October–1 November 2018; pp. 1–9.
26. Schlechtweg, D.; McGillivray, B.; Hengchen, S.; Dubossarsky, H.; Tahmasebi, N. SemEval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv* **2020**, arXiv:2007.11464.
27. Zou, W.Y.; Socher, R.; Cer, D.; Manning, C.D. Bilingual word embeddings for phrase-based machine translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1393–1398.
28. Pym, A.; Grin, F.; Sfreddo, C.; Chan, A.L. *The Status of the Translation Profession in the European Union*; Anthem Press: London, UK, 2013.
29. Rabinovich, E.; Wintner, S.; Lewinsohn, O.L. A Parallel Corpus of Translationese. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, 3–9 April 2016.
30. Nisioi, S.; Rabinovich, E.; Dinu, L.P.; Wintner, S. A Corpus of Native, Non-native and Translated Texts. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), Portorož, Slovenia, 23–28 May 2016.
31. Olohan, M.; Baker, M. Reporting That in Translated English: Evidence for Subconscious Processes of Explicitation? *Across Lang. Cult.* **2000**, *1*, 141–158. [\[CrossRef\]](#)
32. Zufferey, S.; Cartoni, B. A multifactorial analysis of explicitation in translation. *Target* **2014**, *26*, 361–384. [\[CrossRef\]](#)
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
34. Bojar, O.; Buck, C.; Federmann, C.; Haddow, B.; Koehn, P.; Leveling, J.; Monz, C.; Pecina, P.; Post, M.; Saint-Amand, H.; et al. Findings of the 2014 Workshop on Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; Association for Computational Linguistics: Baltimore, MA, USA, 2014; pp. 12–58.
35. Ilisei, I.; Inkpen, D.; Pastor, G.C.; Mitkov, R. Identification of Translationese: A Machine Learning Approach. In Proceedings of the CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing, Iași, Romania, 21–27 March 2010; Gelbukh, A.F., Ed.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6008, pp. 503–511.
36. Rabinovich, E.; Nisioi, S.; Ordan, N.; Wintner, S. On the Similarities Between Native, Non-native and Translated Texts. *arXiv* **2016**, arXiv:1609.03204.
37. Dumais, S. Improving the retrieval of information from external sources. *Behav. Res. Methods Instruments Comput.* **1991**, *23*, 229–236. [\[CrossRef\]](#)

38. Jarvis, S.; Bestgen, Y.; Pepper, S. Maximizing Classification Accuracy in Native Language Identification. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, GA, USA, 13 June 2013; Association for Computational Linguistics: Atlanta, Georgia, 2013; pp. 111–118.
39. Fan, R.E.; Chang, K.W.; Hsieh, C.J.; Wang, X.R.; Lin, C.J. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
40. Malmasi, S.; Evanini, K.; Cahill, A.; Tetreault, J.R.; Pugh, R.A.; Hamill, C.; Napolitano, D.; Qian, Y. A Report on the 2017 Native Language Identification Shared Task. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark, 8 September 2017; pp. 62–75.
41. Zampieri, M.; Malmasi, S.; Scherrer, Y.; Samardžić, T.; Tyers, F.; Silfverberg, M.; Klyueva, N.; Pan, T.L.; Huang, C.R.; Ionescu, R.T.; et al. A Report on the Third VarDial Evaluation Campaign. In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, Minneapolis, MN, USA, 7 June 2019; Association for Computational Linguistics: Ann Arbor, MI, USA, 2019; pp. 1–16.
42. Koppel, M.; Akiva, N.; Dagan, I. Feature instability as a criterion for selecting potential style markers. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 1519–1525. [[CrossRef](#)]
43. Dediu, D.; Cysouw, M. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLoS ONE* **2013**, *8*, e55009.
44. Mosteller, F.; Wallace, D.L. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *J. Am. Stat. Assoc.* **1963**, *58*, 275–309.
45. Nisioi, S. Feature Analysis for Native Language Identification. In Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2015), Cairo, Egypt, 14–20 April 2015; Gelbukh, A.F., Ed.; Springer: Berlin/Heidelberg, Germany, 2015.
46. Münte, T.F.; Wieringa, B.M.; Weyerts, H.; Szentkúti, A.; Matzke, M.; Johannes, S. Differences in brain potentials to open and closed class words: Class and frequency effects. *Neuropsychologia* **2001**, *39*, 91–102. [[CrossRef](#)]
47. Olohan, M. Leave it out! Using a Comparable Corpus to Investigate Aspects of Explicitation in Translation. *Cadernos de Tradução* **2002**, *1*, 153–169.
48. Zhang, X.; Kruger, H.K.; Fang, J. Explicitation in children’s literature translated from English to Chinese: A corpus-based study of personal pronouns. *Perspectives* **2020**, *28*, 717–736. [[CrossRef](#)]
49. Volansky, V.; Ordan, N.; Wintner, S. On the Features of Translationese. *Digit. Scholarsh. Humanit.* **2015**, *30*, 98–118. [[CrossRef](#)]
50. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-Strength Natural Language Processing in Python. 2020. Available online: <https://spacy.io/> (accessed on 30 November 2021).
51. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
52. Martin, L.; Muller, B.; Ortiz Suárez, P.J.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; Sagot, B. CamemBERT: A Tasty French Language Model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7203–7219.
53. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [[CrossRef](#)]
54. Hamilton, W.L.; Leskovec, J.; Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv* **2016**, arXiv:1605.09096.
55. Szymanski, T. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 448–453.
56. Uban, A.; Ciobanu, A.M.; Dinu, L.P. Studying Laws of Semantic Divergence across Languages using Cognate Sets. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, Florence, Italy, 2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 161–166.