Found in Translation: Reconstructing Phylogenetic Language Trees from Translations

Ella Rabinovich[△⋆] **Noam Ordan**[†] **Shuly Wintner**[⋆]
[△]IBM Research Haifa, Israel
[⋆]Department of Computer Science, University of Haifa, Israel
[†]The Arab College for Education, Haifa, Israel
{ellarabi,noam.ordan}@gmail.com, shuly@cs.haifa.ac.il

Abstract

Translation has played an important role in trade, law, commerce, politics, and literature for thousands of years. Translators have always tried to be invisible; ideal translations should look as if they were written originally in the target language. We show that traces of the source language remain in the translation product to the extent that it is possible to uncover the history of the source language by looking only at the translation. Specifically, we automatically reconstruct phylogenetic language trees from *monolingual* texts (translated from several source languages). The signal of the source language is so powerful that it is retained even after two phases of translation. This strongly indicates that source language interference is the most dominant characteristic of translated texts, overshadowing the more subtle signals of universal properties of translation.

1 Introduction

Translation has played a major role in human civilization since the rise of law, religion, and trade in multilingual societies. Evidence of scribe translations goes as far back as four millennia ago, to the time of Hammurabi; this practice is also mentioned in the Bible (Esther 1:22; 8:9). For thousands of years, translators have tried to remain invisible, setting a standard according to which the act of translation should be seamless, and its product should look as if it were written originally in the target language. Cicero (106-43 BC) commented on his translation ethics, "I did not hold it necessary to render word for word, but I preserved the general style and force of the language." These words were echoed 500 years later by St. Jerome (347-420 CE), also known as the patron saint of translators, who wrote, "I render, not word for word, but sense for sense." Translator tendency for invisibility has peaked in the past 150 years in the English speaking world (Venuti, 2008), in spite of some calls for "foreignization" in translations, e.g., the German Romanticists, especially the translations from Greek by Friedrich Hölderlin (Steiner, 1975) and Nabokov's translation of Eugene Onegin. These, however, as both Steiner (1975) and Venuti (2008) argue, are the exception to the rule. In fact, in recent years, the quality of translations has been standardized (ISO 17100). Importantly, the translations we studied in our work conform to this standard.

Despite the continuous efforts of translators, translations are known to feature unique characteristics that set them apart from non-translated texts, referred to as *originals* here (Toury, 1980, 1995; Frawley, 1984; Baker, 1993). This is not the result of poor translation, but rather a statistical phenomenon: various features distribute differently in originals than in translations (Gellerstam, 1986).

Several factors may account for the differences between originals and translations; many are classified as *universal* features of translation. Cognitively speaking, all translations, regardless of the source and target language, are susceptible to the same constraints. Therefore, translation products are expected to share similar artifacts. Such universals include *simplification*: the tendency to make complex source structures simpler in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985); *standardization*: the tendency to over-conform to target language standards (Toury, 1995); and *explicitation*: the tendency to render implicit source structures more explicit in the target language (Blum-Kulka, 1986; Øverås, 1998).

In contrast to translation universals, *interference* reflects the "fingerprints" of the source lan-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 530–540
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-1049

guage on the translation product. Toury (1995) defines interference as “phenomena pertaining to the make-up of the source text tend to be transferred to the target text”. Interference, by definition, is a language-pair specific phenomenon; isomorphic structures shared by the source and target languages can easily replace one another, thereby manifesting the underlying process of cross-linguistic influence of the source language on the translation outcome. Pym (2008) points out that interference is a set of both *segmentational and macrostructural features*.

Our main hypothesis is that, due to interference, languages with shared isomorphic structures are likely to share more features in the target language of a translation. Consequently, the distance between two languages, when assessed using such features, can be retained to some extent in translations from these two languages to a third one. Furthermore, we hypothesize that by extracting structures from translated texts, we can generate a phylogenetic tree that reflects the “true” distances among the source languages. Finally, we conjecture that the quality of such trees will improve when constructed using features that better correspond to interference phenomena, and will deteriorate using more universal features of translation.

The main contribution of this paper is thus the demonstration that interference phenomena in translation are powerful to an extent that facilitates clustering source languages into families and (partially) reconstructing intra-families ties; so much so, that these results hold even after two rounds of translation. Moreover, we perform analysis of various linguistic phenomena in the source languages, laying out quantitative grounds for the language typology reconstruction results.

2 Related work

A number of works in historical linguistics have applied methods from the field of bioinformatics, in particular algorithms for generating phylogenetic trees (Ringe et al., 2002; Nakhleh et al., 2005a,b; Ellison and Kirby, 2006; Boc et al., 2010). Most of them rely on lists of *cognates*, words in multiple languages with a common origin that share a similar meaning and a similar pronunciation (Dyen et al., 1992; Rexová et al., 2003). These works all rely on multilingual data, whereas we construct phylogenetic trees from texts in a single language.

The claim that translations exhibit unique properties is well established in translation studies literature (Toury, 1980; Frawley, 1984; Baker, 1993; Toury, 1995). Based on this assumption, several works use text classification techniques employing supervised, and recently also unsupervised, machine learning approaches, to distinguish between originals and translations (Baroni and Bernardini, 2006; Ilisei et al., 2010; Koppel and Ordan, 2011; Volansky et al., 2015; Rabinovich and Wintner, 2015; Avner et al., 2016). The features used in these studies reflect both universal and interference-related traits. Along the way, interference was proven to be a robust phenomenon, operating in every single sentence, even on the morpheme level (Avner et al., 2016). Interference can also be studied on pairs of source- and target languages and focus, for example, on word order (Eetemadi and Toutanova, 2014).

The powerful signal of interference is evident, e.g., by the finding that a classifier trained to distinguish between originals and translations from one language, exhibits lower accuracy when tested on translations from another language, and this accuracy deteriorates proportionally to the distance between the source and target languages (Koppel and Ordan, 2011). Consequently, it is possible to accurately distinguish among translations from various source languages (van Halteren, 2008).

A related task, identifying the native tongue of English language students based only on their writing in English, has been the subject of recent interest (Tetreault et al., 2013). The relations between this task and identification of the source language of translation has been emphazied, e.g., by Tsvetkov et al. (2013). English texts produced by native speakers of a variety of languages have been used to reconstruct phylogenetic trees, with varying degrees of success (Nagata and Whittaker, 2013; Berzak et al., 2014). In contrast to language learners, however, translators translate into their mother tongue, so the texts we studied were written by highly competent native speakers. Our work is the first to construct phylogenetic trees from translations.

3 Methodology

3.1 Dataset

This corpus-based study uses Europarl (Koehn, 2005), the proceedings of the European Parliament and their translations into all the official European Union (EU) languages. Europarl is one of the most popular parallel resources in natural language processing, and has been used extensively in machine translation. We use a version of Europarl spanning the years 1999 through 2011, in which the direction of translation has been established through a comprehensive cross-lingual validation of the speakers' original language (Rabinovich et al., 2015).

All parliament speeches were translated[1] from the original language into all other EU languages (21 at the time) using English as an intermediate, *pivot* language. We thus refer to translations into English as *direct*, while translations into all other languages, via English as a third language, are *indirect*. We hypothesize that indirect translation will obscure the markers of the original language in the final translation. Nevertheless, we expect (weakened) fingerprints of the source language to be identifiable in the target despite the pivot, presumably resulting in somewhat poorer phylogenetic trees.

We focus on 17 source languages, grouped into 3 language families: Germanic, Romance, and Balto-Slavic.[2] These include translations to English and to French from Bulgarian (BG), Czech (CS), Danish (DA), Dutch (NL), English (EN), French (FR), German (DE), Italian (IT), Latvian (LV), Lithuanian (LT), Polish (PL), Portuguese (PT), Romanian (RO), Slovak (SK), Slovenian (SL), Spanish (ES), and Swedish (SV). We also included texts written originally in English and French.

All datasets were split on sentence boundary, cleaned (empty lines removed), tokenized, and annotated for part-of-speech (POS) using the Stanford tools (Manning et al., 2014). In all the tree reconstruction experiments, we sampled equal-sized chunks from each source language, using as much data as available for all languages. This yielded $27,000$ tokens from translations to English, and $30,000$ tokens from translations into French.

3.2 Features

Following standard practice (Volansky et al., 2015; Rabinovich and Wintner, 2015), we represented both original and translated texts as feature vectors, where the choice of features determines the extent to which we expect source-language interference to be present in the translation product. Crucially, the features abstract away from the contents of the texts and focus on their structure, reflecting, among other things, morphological and syntactic patterns. We use the following feature sets: 1. The top-1,000 most frequent POS trigrams, reflecting shallow syntactic structure. 2. Function words (FW), words known to reflect grammar of texts in numerous classification tasks, as they include non-content words such as articles, prepositions, etc. (Koppel and Ordan, 2011).[3] 3. Cohesive markers (Hinkel, 2001); these words and phrases are assumed to be over-represented in translated texts, where, for example, an implicit contrast in the original is made explicit in the target text with words such as *'but'* or *'however'*.[4] Note that the first two feature sets are strongly associated with interference, whereas the third is assumed to be universal and an instance of explicitation. We therefore expect trees based on the first two feature sets to be much better than those based on the third.

3.3 The Indo-European phylogenetic tree

The last few decades produced a large body of research on the evolution of individual languages and language families. While the existence of the Indo-European (IE) family of languages is an established fact, its history and origins are still a matter of much controversy (Pereltsvaig and Lewis, 2015). Furthermore, the actual subgroupings of languages within this family are not clear-cut (Ringe et al., 2002). Consequently, algorithms that attempt to reconstruct the IE languages tree face a serious evaluation challenge (Ringe et al., 2002; Rexová et al., 2003; Nakhleh et al., 2005a).

To evaluate the quality of the reconstructed trees, we define a metric to accurately assess their distance from the "true" tree. The tree that we use as ground truth (Serva and Petroni, 2008) has

[1] The common practice is that one translates into one's native language; in particular, this practice is strictly imposed in the EU parliament where a translator must have perfect proficiency in the target language, meeting very high standards of accuracy.

[2] We excluded source languages with insufficient amounts of data, along with Greek, which is the only representative of the Hellenic family.

[3] For French we used the list of FW available at `https://code.google.com/archive/p/stop-words/`.

[4] For French we used `http://utilisateurs.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml`.

several advantages. First, it is similar to a well-accepted tree (Gray and Atkinson, 2003) (which is not insusceptible to criticism (Pereltsvaig and Lewis, 2015)). The differences between the two are mostly irrelevant for the group of languages that we address in this research. Second, it is a binary tree, facilitating comparison with the trees we produce, which are also binary branching. Third, its branches are decorated with the approximate year in which splitting occurred. This provides a way to induce the distance between two languages, modeled as lengths of paths in the tree, based on chronological information.

We projected the gold tree (Serva and Petroni, 2008) onto the set of 17 languages we considered in this work, preserving branch lengths. Figure 1 depicts the resulting gold-standard subtree.

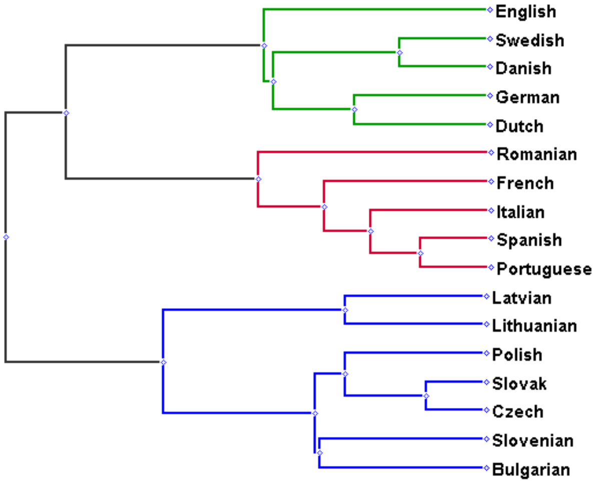


Figure 1: Gold standard tree, pruned

We reconstructed phylogenetic language trees by performing agglomerative (hierarchical) clustering of feature vectors extracted separately from English and French translations. We performed clustering using the variance minimization algorithm (Ward Jr, 1963) with Euclidean distance (the implementation available in the Python SciPy library). All feature values were normalized to a zero-one scale prior to clustering.

3.4 Evaluation methodology

To evaluate the quality of the trees we generate, we compute their similarity to the gold standard via two metrics: *unweighted*, assessing only structural (topological) similarity, and *weighted*, estimating similarity based on both structure and branching length.

Several methods have been proposed for evaluating the quality of phylogenetic language trees (Pompei et al., 2011; Wichmann and Grant, 2012; Nouri and Yangarber, 2016). A popular metric is the Robinson-Foulds (RF) methodology (Robinson and Foulds, 1981), which is based on the symmetric difference in the number of *bi-partitions*, the ways in which an edge can split the leaves of a tree into two sets. The distance between two trees is then defined as the number of splits induced by one of the trees, but not the other. Despite its popularity, the RF metric has well-known shortcomings; for example, relocating a single leaf can result in a tree maximally distant from the original one (Böcker et al., 2013). Additional methodologies for evaluating phylogenetic trees include *branch score distance* (Kuhner and Felsenstein, 1994), enhancing RF with branch lengths, *purity score* (Heller and Ghahramani, 2005), and *subtree score* (Teh et al., 2009). The latter two ignore branch lengths and only consider structural similarities for evaluation.

We opted for a simple yet powerful adaptation of the L2-norm to leaf-pair distance, inherently suitable for both unweighted and weighted evaluation. Given a tree of N leaves, l_i, $i \in [1..N]$, the *weighted distance* between two leaves l_i, l_j in a tree τ, denoted $D_\tau(l_i, l_j)$, is the sum of the weights of all edges on the shortest path between l_i and l_j. The *unweighted distance* sums up the *number* of the edges in this path (i.e., all weights are equal to 1). The distance $Dist(\tau, g)$ between a generated tree τ and the gold tree g is then calculated by summing the square differences between all leaf-pair distances (whether weighted or unweighted) in the two trees:

$$Dist(\tau, g) = \sum_{i,j \in [1..N]; i \neq j} (D_\tau(l_i, l_j) - D_g(l_i, l_j))^2$$

4 Detection of Translations and their Source Language

4.1 Identification of translation

We first reconfirmed that originals and translations are easily separable, extending results of supervised classification of *O* vs. *T* (where O refers to original English texts, and T to translated English) (Baroni and Bernardini, 2006; van Halteren, 2008; Volansky et al., 2015) to the 16 original languages considered in this work. We also conducted similar experiments with French originals and translations. We used 200 chunks of approximately 2K tokens (respecting sentence boundaries) from both O and T, and normalized the values of lexical features by the number of tokens in each chunk. For classification, we used Platt's sequential minimal optimization algorithm (Keerthi et al., 2001; Hall et al., 2009) to train support vector machine classifiers with the default linear kernel. We evaluated the results with 10-fold cross-validation.

Table 1 presents the classification accuracy of (English and French) O vs. T using each feature set. In line with previous works (Ilisei et al., 2010; Volansky et al., 2015; Rabinovich and Wintner, 2015), the binary classification results are highly accurate, achieving over 95% accuracy using POS-trigrams and function words for both English and French, and above 85% using cohesive markers.

Feature	English	French
POS-trigrams	97.60	98.40
Function words	96.45	95.15
Cohesive markers	86.50	85.25

Table 1: Classification accuracy (%) of English and French O vs. T

4.2 Identification of source language

Identifying the source language of translated texts is a task in which machines clearly outperform humans (Baroni and Bernardini, 2006). Koppel and Ordan (2011) performed 5-way classification of texts translated from Italian, French, Spanish, German, and Finnish, achieving an accuracy of 92.7%. Furthermore, misclassified instances were more frequently assigned to genetically related languages.

We extended this experiment to 14 languages representing 3 language families (the number of languages was limited by the amount of data available). We extracted 100 chunks of 1,000 tokens each from each source language and classified the translated English (and, separately, French) texts into 14 classes using the best performing POS-trigrams feature set. Cross-validation evaluation yielded an accuracy of 75.61% on English translations (note that the baseline is $100/14 = 7.14\%$).

The corresponding confusion matrix, presented in Figure 2 (left), reveals interesting phenomena: much of the confusion resides within language families, framed by the bold line in the figure. For example, instances of Germanic languages are almost perfectly classified as Germanic, with only a few chunks assigned to other language families. The evident intra-family linguistic ties exposed by this experiment support the intuition that cross-linguistic transfer in translation is governed by typological properties of the source language. That is, translations from *related* sources tend to resemble each other to a greater extent than translations from more *distant* languages.

This observation is further supported by the evaluation of a three-way classification task, where the goal is to only identify the language family (Germanic, Romance, or Balto-Slavic): the accuracy of this task is 90.62%. Note also that the mis-classified instances of both Romance and Germanic languages are nearly never attributed to Balto-Slavic languages, since Germanic and Romance are much closer to each other than to Balto-Slavic.

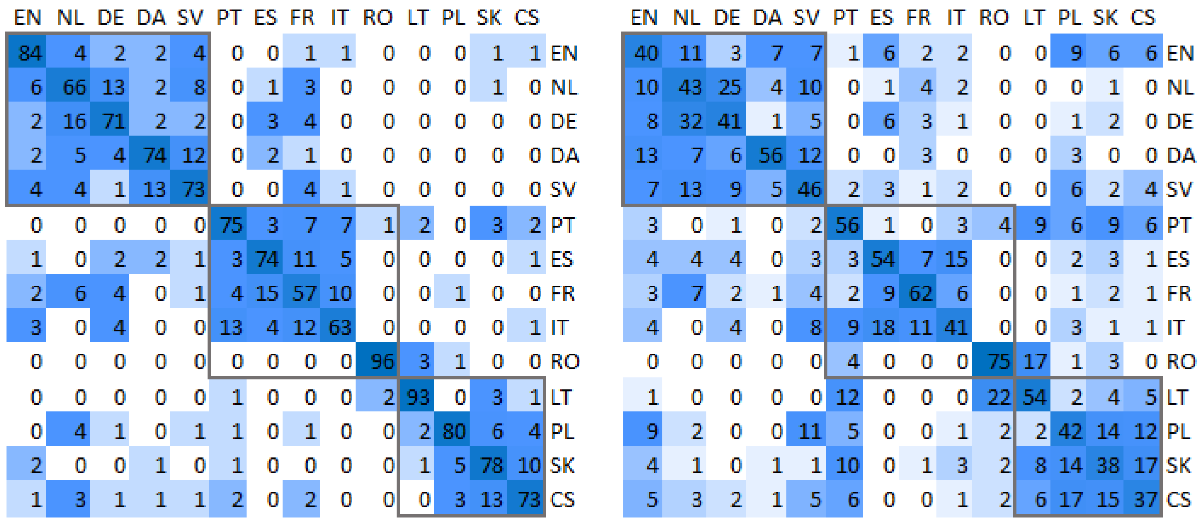
Figure 2 (right) displays a similar confusion matrix, the only difference being that *French* translations are classified. We attribute the lower cross-validation accuracy (48.92%, reflected also by the lower number of correctly assigned instances on the matrix diagonal, compared to English) to the intervention of the pivot language in the translation process. Nevertheless, the confusion is still mainly constrained to intra-family boundaries.

5 Reconstruction of Phylogenetic Language Trees

5.1 Reconstructing language typology

Inspired by the results reported in Section 4.2, we generated phylogenetic language trees from both English and French texts translated from the other European languages. We hypothesized that interference from the source language was present in the translation product to an extent that would facilitate the construction of a tree sufficiently similar to the gold IE tree (Figure 1).

The best trees, those closest to the gold standard, were generated using POS-trigrams: these are the features that are most closely associated with source-language interference (see Section 3.2). Figure 3 depicts the trees produced from English and French translations using POS-trigrams. Both trees reasonably group individual languages into three language-family branches. In particular, they cluster the Germanic and Romance languages closer than the Balto-Slavic. Capturing the more subtle intra-family ties turned out to be



	EN	NL	DE	DA	SV	PT	ES	FR	IT	RO	LT	PL	SK	CS
EN	84	4	2	2	4	0	0	1	1	0	0	0	1	1
NL	6	66	13	2	8	0	1	3	0	0	0	0	1	0
DE	2	16	71	2	2	0	3	4	0	0	0	0	0	0
DA	2	5	4	74	12	0	2	1	0	0	0	0	0	0
SV	4	4	1	13	73	0	0	4	1	0	0	0	0	0
PT	0	0	0	0	0	75	3	7	7	1	2	0	3	2
ES	1	0	2	2	1	3	74	11	5	0	0	0	0	1
FR	2	6	4	0	1	4	15	57	10	0	0	1	0	0
IT	3	0	4	0	0	13	4	12	63	0	0	0	0	1
RO	0	0	0	0	0	0	0	0	0	96	3	1	0	0
LT	0	0	0	0	0	1	0	0	0	2	93	0	3	1
PL	0	4	1	0	1	1	0	1	0	0	2	80	6	4
SK	2	0	0	1	0	1	0	0	0	0	1	5	78	10
CS	1	3	1	1	1	2	0	2	0	0	0	3	13	73

	EN	NL	DE	DA	SV	PT	ES	FR	IT	RO	LT	PL	SK	CS
EN	40	11	3	7	7	1	6	2	2	0	0	9	6	6
NL	10	43	25	4	10	0	1	4	2	0	0	0	1	0
DE	8	32	41	1	5	0	6	3	1	0	0	1	2	0
DA	13	7	6	56	12	0	0	3	0	0	0	3	0	0
SV	7	13	9	5	46	2	3	1	2	0	0	6	2	4
PT	3	0	1	0	2	56	1	0	3	4	9	6	9	6
ES	4	4	4	0	3	3	54	7	15	0	0	2	3	1
FR	3	7	2	1	4	2	9	62	6	0	0	1	2	1
IT	4	0	4	0	8	9	18	11	41	0	0	3	1	1
RO	0	0	0	0	0	4	0	0	0	75	17	1	3	0
LT	1	0	0	0	0	12	0	0	0	22	54	2	4	5
PL	9	2	0	0	11	5	0	0	1	2	2	42	14	12
SK	4	1	0	1	1	10	0	1	3	2	8	14	38	17
CS	5	3	2	1	5	6	0	0	1	2	6	17	15	37

Figure 2: Confusion matrix of 14-way classification of English (left) and French (right) translations. The actual class is represented by rows and the predicted one by columns.

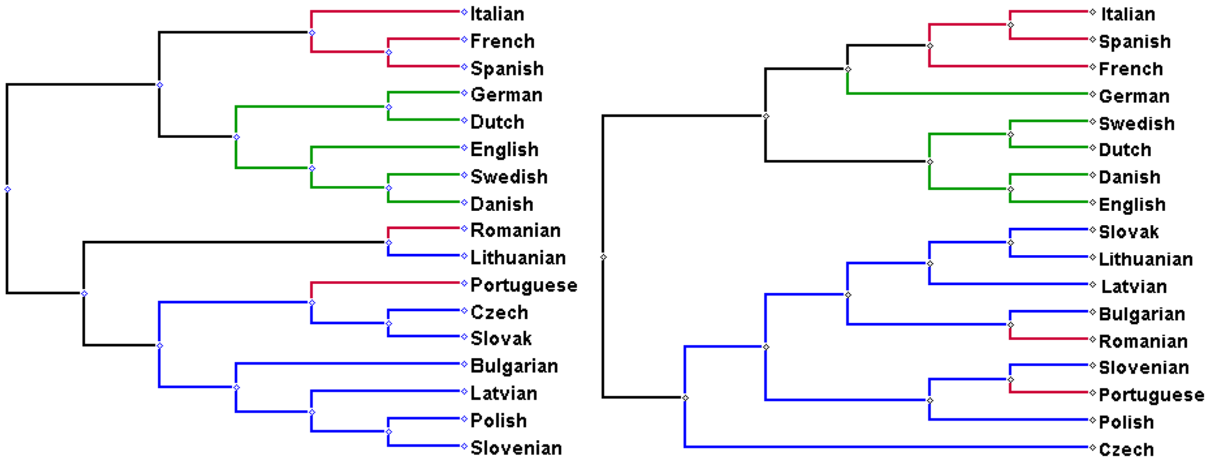


Figure 3: Phylogenetic language trees generated with English (left) and French (right) translations

more challenging, although English outperformed its French counterpart on this task by almost perfectly reconstructing the Germanic sub-tree.

We repeated the clustering experiments with various feature sets. For each feature set, we randomly sampled equally-sized subsets of the dataset (translated from each of the source languages), represented the data as feature vectors, generated a tree by clustering the feature vectors, and then computed the weighted and unweighted distances between the generated tree and the gold standard. We repeated this procedure 50 times for each feature set, and then averaged the resulting distances. We report this average and the standard deviation.[5]

5.2 Evaluation results

The *unweighted* evaluation results are listed in Table 2. For comparison, we also present the distance obtained for a random tree, generated by sampling a random distance matrix from the uniform $(0, 1)$ distribution. The reported random tree evaluation score is averaged over 1000 experiments. Similarly, we present *weighted* evaluation results in Table 3. All distances are normalized to a zero-one scale, where the bounds – zero and one – represent the identical and the most distant tree w.r.t. the gold standard, respectively.

The results reveal several interesting observations. First, as expected, POS-trigrams induce

[5] All the trees, both cladograms (with branches of equal length) and phylograms (with branch lengths proportional to the distance between two nodes), can be found at http://cl.haifa.ac.il/projects/translationese/acl2017_found-in-translation_trees.pdf

Target language	English		French	
Feature	AVG	STD	AVG	STD
POS-trigrams + FW	0.362	0.07	**0.367**	0.06
POS-trigrams	**0.353**	0.06	0.399	0.08
Function words	0.429	0.07	0.450	0.08
Cohesive markers	0.626	0.16	0.678	0.14
Random tree	0.724	0.07	0.724	0.07

Table 2: Unweighted evaluation of generated trees. AVG represents the average distance of a tree from the gold standard. The lowest distance in a column is boldfaced.

Target language	English		French	
Feature	AVG	STD	AVG	STD
POS-trigrams + FW	**0.278**	0.03	**0.348**	0.02
POS-trigrams	0.301	0.03	0.351	0.03
Function words	0.304	0.03	0.376	0.05
Cohesive markers	0.598	0.12	0.636	0.07
Random tree	0.676	0.10	0.676	0.10

Table 3: Weighted evaluation of generated trees. AVG represents the average distance of a tree from the gold standard. The lowest distance in a column is boldfaced.

trees closest to the gold standard among *distinct* feature sets. This corroborates our hypothesis that this feature set carries over interference of the source language to a considerable extent (see Section 1). Furthermore, function words achieve more moderate results, but still much better than random. This reflects the fact that these features carry over some grammatical constructs of the source language into the translation product.

Finally, in all cases, the least accurate tree, nearly random, is produced by cohesive markers; this is an evidence that this feature is source-language agnostic and reflects the universal effect of explicitation (see Section 3.2). While cohesive markers are a good indicator of translations, they reflect properties that are not indicative of the source language. The combination of POS-trigrams and FW yields the best tree in three out of four cases, implying that these feature sets capture different, complementary aspects of the source-language interference.

Surprisingly, reasonably good trees were also generated from French translations; yet, these trees are systematically worse than their English counterparts. The original signal of the source language is distorted twice: first via a Germanic language (English) and then via a Romance language (French). However, the signal is strong enough to yield a clear phylogenetic tree of the source languages. Interference is thus revealed to be an extremely powerful force, partially resistant to intermediate distortions.

6 Analysis

We demonstrated that source-language traces are dominant in translation products to an extent that facilitates reconstruction of the history of the source languages. We now inspect some of these phenomena in more detail to better understand the prominent characteristics of interference. For each phenomenon, we computed the frequencies of patterns that reflect it in texts translated to English from each individual language, and averaged the measures over each language family (Germanic, Romance, and Balto-Slavic). Figure 4 depicts the results.

6.1 Definite articles

Languages vary greatly in their use of articles. Like other Germanic languages, English has both definite (*'a'*) and indefinite (*'the'*) articles. However, many languages only have definite articles and some only have indefinite articles. Romance languages, and in particular the five Romance languages of our dataset, have definite articles that can sometimes be omitted, but not as commonly as in English. Balto-Slavic languages typically do not have any articles.

Mastering the use of articles in English is notoriously hard, leading to errors in non-native speakers (Han et al., 2006). For example, native speakers of Slavic languages tend to *over*use definite articles in German (Hirschmann et al., 2013). Similarly, we expect translations from Balto-Slavic languages to overuse *'the'*. We computed the frequencies of *'the'* in translations to English from each of the three language families. The results show a significant overuse of *'the'* in translations from Balto-Slavic languages, and some overuse in translations from Romance languages.

6.2 Possessive constructions

Languages also vary in the way they mark possession. English marks it in three ways: with the clitic *''s'* (*'the guest's room'*), with a prepositional phrase containing *'of'* (*'the room of the guest'*), and, like in other Germanic languages, with noun compounds (*'guest room'*). Compounds are considerably less frequent in Romance languages

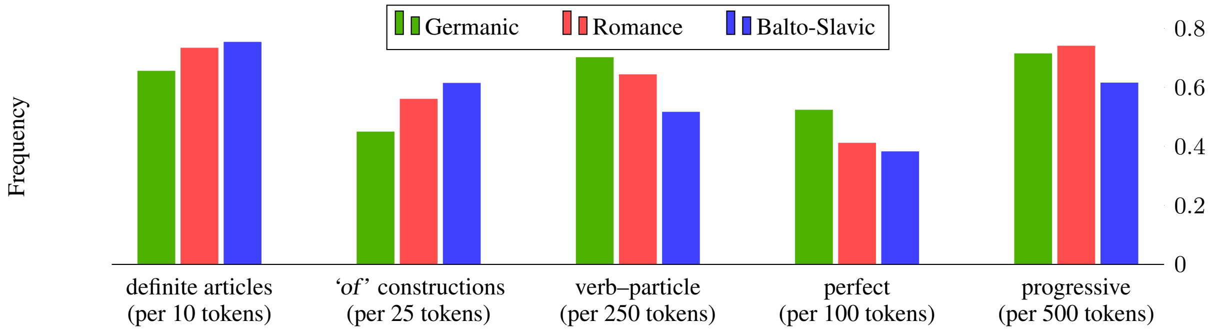


Figure 4: Frequencies reflecting various linguistic phenomena (Sections 6.1– 6.4) in English translations

(Swan and Smith, 2001); Balto-Slavic indicates possession using case-marking. Languages also vary with respect to whether or not possession is head-marked. In Balto-Slavic languages, the genitive case is head-marked, which reverses the order of the two nouns with respect to the common English *'s'* construction. Since copying word order, if possible across languages, is one of the major features of interference (Eetemadi and Toutanova, 2014), we anticipated that Balto-Slavic languages will exhibit the highest rate of noun-*'of'*-NP constructions. This would be followed by Romance languages, in which this construction is highly common, and then by Germanic languages, where noun compounds can often be copied as such. The results are consistent with our expectations.

6.3 Verb-particle constructions

Verb-particle constructions (e.g., *'turn down'*) consist of verbs that combine with a particle to create a new meaning (Dehé et al., 2002). Such constructions are much more common in Germanic languages (Iacobini and Masini, 2005), hence we expect to encounter their equivalents in English translations more frequently. We computed the frequencies of these constructions in the data; the results show a clear overuse of verb-particle constructions in translations from Germanic, and an underuse of such constructions in translations from Balto-Slavic.

6.4 Tense and aspect

Tense and aspect are expressed in different ways across languages. English, like other Germanic languages, uses a full system of aspectual distinctions, expressed via perfect and progressive forms (with the auxiliary verbs *'have'* or *'be'*). Balto-Slavic, in contrast, has no such system, and the distinction is marked lexically, by having two types of verbs. Romance languages are in between, with both lexical and grammatical distinctions. We computed the frequencies of perfect forms (defined as the auxiliary *'have'* followed by the past participle form), and the progressive forms (defined as the auxiliary *'be'* plus a present participle form). Indeed, Germanic overuses the perfect aspect significantly; the use of the progressive aspect also varies across language families, exhibiting the lowest frequency in translations from Balto-Slavic.

7 Conclusion

Translations may be considered distortions of the original text, but this distortion is far from random. It depicts a very clear picture, reflecting language typology to the extent that disregarding the sources altogether, a phylogenetic tree can be reconstructed from a monolingual corpus consisting of multiple translations. This holds for the product of highly professional translators, who conform to a common standard, and whose products are edited by native speakers, like themselves. It even holds after two phases of translations. We are presently trying to extend these results to translations in a different domain (literary texts) into a very different language (Hebrew).

Postulated universals in linguistics (Greenberg, 1963) were confronted with much contradicting evidence in recent years (Evans and Levinson, 2009), and the long quest for translation universals (Mauranen and Kujamäki, 2004) should now be viewed in light of our finding: more than anything else, translations are typified by interference. This does not undermine the force of translation universals: we demonstrated how explicitation, in the form of cohesive markers, can help identify translations. It may be possible to define classifiers implementing other universal facets of translation, e.g., simplification, which will yield good separation between O and T. However, explicitation fails in the reproduction of language typology, whereas interference-based features produce trees of considerable quality.

Remarkably, translations to contemporary English and French capture part of the millennium-old history of the source languages from which the translations were made. Our trees reflect some of the historical connections among the languages, but of course they are related in other ways, too (whether incidental, areal, etc.). This may explain the case of Romanian in our reconstructed trees: it has been isolated for many years from other Romance languages and was under heavy influence from Balto-Slavic languages.

Very little research has been done in historical linguistics on how translations impact the evolvement of languages. The major trends relate to loan translations (Jahr, 1999), or the impact of canonical texts, such as Luther's translation of the Bible to German (Russ, 1994) or the case of the King James translation to English (Crystal, 2010). It has been attested that for certain languages, up to 30% of published materials are mediated through translation (Pym and Chrupała, 2005). Given the fingerprints left on target language texts, translations very likely play a role in language change. We leave this as a direction for future research.

Acknowledgements

We wish to thank the three ACL anonymous reviewers for their constructive feedback. We are grateful to Sergiu Nisioi and Oren Weimann for their advice and helpful suggestions. We are also thankful to Yonatan Belinkov and Michael Katz for insightful and valuable comments.

References

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities* 31(1):30–54. http://dx.doi.org/10.1093/llc/fqu047.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and technology: in honour of John Sinclair*, John Benjamins, Amsterdam, pages 233–252.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3):259–274.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. pages 21–29. http://aclweb.org/anthology/W/W14/W14-1603.pdf.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication Discourse and cognition in translation and second language acquisition studies*, Gunter Narr Verlag, volume 35, pages 17–35.

Shoshana Blum-Kulka and Eddie A. Levenston. 1983. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, *Strategies in Interlanguage Communication*, Longman, pages 119–139.

Alix Boc, Anna Maria Di Sciullo, and Vladimir Makarenkov. 2010. Classification of the Indo-European languages using a phylogenetic network approach. In Hermann Locarek-Junge and Claus Weihs, editors, *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V., Dresden, March 13-18, 2009*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 647–655.

Sebastian Böcker, Stefan Canzar, and Gunnar W Klau. 2013. The generalized Robinson-Foulds metric. In *International Workshop on Algorithms in Bioinformatics*. Springer, pages 156–169.

David Crystal. 2010. *Begat: The King James Bible and the English Language*. Oxford University Press.

Nicole Dehé, Ray Jackendoff, Andrew McIntyre, and Silke Urban, editors. 2002. *Verb-particle Explorations*. Interface explorations. Mouton de Gruyter.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean classification. a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5):iii–132.

Sauleh Eetemadi and Kristina Toutanova. 2014. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 159–164. http://www.aclweb.org/anthology/D14-1018.

T. Mark Ellison and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th*

Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 273–280. https://doi.org/10.3115/1220175.1220210.

Nicholas Evans and Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5):429–494.

William Frawley. 1984. Prolegomenon to a theory of translation. In William Frawley, editor, *Translation. Literary, Linguistic and Philosophical Perspectives*, University of Delaware Press, Newark, pages 159–175.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, CWK Gleerup, Lund, pages 88–95.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435–439.

Joseph H. Greenberg, editor. 1963. *Universals of Human Language*. MIT Press, Cambridge, Mass.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11(1):10–18. https://doi.org/10.1145/1656274.1656278.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering* 12(02):115–129.

Katherine A Heller and Zoubin Ghahramani. 2005. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*. ACM, pages 297–304.

Eli Hinkel. 2001. Matters of cohesion in L2 academic texts. *Applied Language Learning* 12(2):111–132.

Hagen Hirschmann, Anke Lüdeling, Ines Rehbein, Marc Reznicek, and Amir Zeldes. 2013. Underuse of syntactic categories in Falko. a case study on modification. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *20 Years of Learner Corpus Research. Looking Back, Moving Ahead.*, Presses Universitaires de Louvain, Louvain la Neuve, pages 223–234.

Claudio Iacobini and Francesca Masini. 2005. Verb-particle constructions and prefixed verbs in Italian: typology, diachrony and semantics. In *Mediterranean Morphology Meetings*. volume 5, pages 157–184.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. http://dx.doi.org/10.1007/978-3-642-12116-6.

Ernst Håkon Jahr. 1999. *Language change: advances in historical sociolinguistics*, volume 114. Walter de Gruyter.

S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation* 13(3):637–649.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 1318–1326. http://www.aclweb.org/anthology/P11-1132.

Mary K Kuhner and Joseph Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11(3):459–468.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Anna Mauranen and Pekka Kujamäki, editors. 2004. *Translation universals: Do they exist?*. John Benjamins.

Ryo Nagata and Edward W. D. Whittaker. 2013. Reconstructing an Indo-European family tree from non-native English texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 1137–1147. http://aclweb.org/anthology/P/P13/P13-1112.pdf.

Luay Nakhleh, Don Ringe, and Tandy Warnow. 2005a. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2):382–420.

Luay Nakhleh, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005b. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103(2):171–192. https://doi.org/10.1111/j.1467-968X.2005.00149.x.

Javad Nouri and Roman Yangarber. 2016. Modeling language evolution with codes that utilize context and phonetic features. *CoNLL 2016* page 136.

Lin Øverås. 1998. In search of the third code: An investigation of norms in literary translation. *Meta* 43(4):557–570.

Asya Pereltsvaig and Martin W. Lewis. 2015. *The Indo-European Controversy*. Cambridge University Press, Cambridge.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one* 6(6):e20109.

Anthony Pym. 2008. On Toury's laws of how translators translate. *BENJAMINS TRANSLATION LIBRARY* 75:311.

Anthony Pym and Grzegorz Chrupała. 2005. The quantitative analysis of translation flows in the age of an international language. In Albert Branchadell and Lovell M. West, editors, *Less Translated Languages*, John Benjamins, Amsterdam, pages 27–38.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics* 3:419–432.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2015. The Haifa corpus of translationese. Unpublished manuscript. http://arxiv.org/abs/1509.03611.

Kateřina Rexová, Daniel Frynta, and Jan Zrzavỳ. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19(2):120–127.

Kateřina Rexová, Daniel Frynta, and Jan Zrzavý. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics-the International Journal of the Willi Hennig Society* 19(2):120–127.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1):59–129. https://doi.org/10.1111/1467-968X.00091.

David F Robinson and Leslie R Foulds. 1981. Comparison of phylogenetic trees. *Mathematical biosciences* 53(1):131–147.

Charles VJ Russ. 1994. *The German language today: A linguistic introduction*. Psychology Press.

Maurizio Serva and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *Europhysics Letters* 81(6):68005. http://stacks.iop.org/0295-5075/81/i=6/a=68005.

George Steiner. 1975. *After Babel*. University Press.

Michael Swan and Bernard Smith. 2001. *Learner English*. Cambridge University Press, Cambridge, second edition.

Yee Whye Teh, Hal Daumé III, and Daniel Roy. 2009. Bayesian agglomerative clustering with coalescents. *arXiv preprint arXiv:0907.0781* .

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics.

Gideon Toury. 1980. *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv.

Gideon Toury. 1995. *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam / Philadelphia.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 2013. Identifying the L1 of non-native writers: the CMU-Haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, pages 279–287. http://www.aclweb.org/anthology/W13-1736.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*. pages 937–944. http://www.aclweb.org/anthology/C08-1118.

Ria Vanderauwerea. 1985. *Dutch novels translated into English: the transformation of a 'minority' literature*. Rodopi, Amsterdam.

Lawrence Venuti. 2008. *The translator's invisibility: A history of translation*. Routledge.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30(1):98–118.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58(301):236–244.

Søren Wichmann and Anthony P Grant. 2012. *Quantitative approaches to linguistic diversity: commemorating the centenary of the birth of Morris Swadesh*, volume 46. John Benjamins Publishing.