A Computational Approach to the Study of Multilingualism

Ella Rabinovich

A thesis submitted for the degree "Doctor of Philosophy"

University of Haifa
Faculty of Social Sciences
Department of Computer Sciences

July 2018

A Computational Approach to the Study of Multilingualism

By: Ella Rabinovich

Supervised by: Prof. Shuly Wintner

A thesis submitted for the degree "Doctor of Philosophy"

University of Haifa
Faculty of Social Sciences
Department of Computer Sciences

Recommended by: _____ Date: _____
(Advisor)

Approved by: _____ Date: _____
(Chairman of Ph.D. committee)

December 2018

Acknowledgements

I was very fortunate to have Shuly Wintner as my Ph.D. advisor. His knowledge, patience, thoughtful guidance, unprecedented commitment and genuine passion for language helped shape me as researcher and guided me safely through the last few years. Shuly, the contribution of your wisdom, open-mindness, as well as broad and deep perspective, to this thesis cannot be overestimated. I will always be thankful for that.

I would like to express a special gratitude to Noam Ordan for long hours of fruitful discussions, for sharing his inexhaustible knowledge, educated advice and experience in my quest for answers. Noam, it has always been a great pleasure working with you.

I also have a pleasure in acknowledging my colleagues Sergio Nisioi, Gili Goldin, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Ofek Luis Lewinsohn for making this journey so colorful and interesting. I have learned and benefited a lot from our collaborations, and looking forward for more to come.

My heartfelt appreciation is given to Yulia Tsvetkov for fruitful collaborations, enlightening discussions and continuous encouragement to follow my dreams. Yulia, as both a researcher and a kind friend, you have been a pillar of support and a wonderful role model, from whom I draw constant inspiration. Our collaboration has always been pleasant and educational, and I truly hope that our paths will cross in the future.

I am grateful to Yuval Nov, Anat Prior, Oren Weimann, Michael Katz, Roy Bar-Haim, Haggai Roitman, Yosi Mass and Dafna Sheinwald for much advice and assistance throughout these years. I have no doubt that your extensive experience, educated suggestions and support helped me fulfilling my goals.

My sincere acknowledgment is given to the administrative staff members of the Department of Computer Science at the University of Haifa. Thank you for your warm welcome, continuous support, and dedicated and generous assistance in making the way to this achievement so enjoyable. Your help is always highly appreciated. I would also like to thank the Authority of Graduate Studies and the Dean of the Faculty of Social Sciences for their generous funding support throughout my studies.

Last, and most important, my deepest gratitude are to my family for their wholehearted and unconditional love and encouragement. My mother, who raised me with a love of science and supported me in all my pursuits. My children, Michal, Shira and Alon, for their patience, and for bringing so much light, happiness and joy to my life. And most of all – my beloved husband Sivan, for his love.

Contents

Abstract

A Computational Approach to the Study of Multilingualism

Ella Rabinovich

The goal of this thesis is to propose and evaluate an approach for bridging the gap between two related areas of research on bilingualism — translation studies and second language acquisition —- that may significantly benefit from cross-disciplinary study. We investigate the characteristics of language production that is influenced by the existence of another linguistic system – language that is produced by a variety of multilinguals, including learners, advanced non-native speakers and translators. We ask whether these language varieties are subject to unified principles, governed by phenomena that stem from the co-existence of multiple linguistic systems in a bilingual brain. By applying a range of computational methodologies, we highlight factors that account for the commonalities and the distinctions between various crosslingual languages varieties.

This thesis addresses fundamental questions related to the language of bilinguals (translators and non-native speakers), both in isolation and jointly. We begin with an in-depth analysis of the unique properties that characterize translated texts (both universal and source-language dependent), as well as the interplay between them. We further study lexical choices of advanced non-native speakers, highlighting the cognate facilitation phenomenon as one of the important factors shaping their language. Finally, we propose a unified computational umbrella for exploring these two related areas of research on bilingualism, identifying the similarities and the differences between translations and the language of advanced, highly-fluent non-native speakers.

Major features of bilingualism, including grammatical, cognitive, and social aspects, have been extensively studied by scholars for over half a century. Crucially, much of this research has been conducted with small, carefully-curated datasets or in a laboratory experimental setup. We show that the availability of large and diverse datasets of productions of non-native speakers stimulates new opportunities for pursuing the emerging direction of computational investigation of bilingualism, thereby tying empirical results with well-established theoretical foundations.

List of Tables

List of Figures

List of Publications

This thesis resulted in the following publications (relevant thesis chapters are listed):

1. Gili Goldin, Ella Rabinovich, and Shuly Wintner. Native language identification with user generated content, Under review.

2. Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018. (Chapter 5).

3. Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL-2017)*, pp. 530–540. Association for Computational Linguistics, 2017a. (Chapter 4).

4. Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017)*, pp. 1074–1084. Association for Computational Linguistics, 2017b.

5. Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1870–1881, 2016a. (Chapter 6).

6. Sergiu Nisioi, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

7. Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. A parallel corpus of translationese. In *Proccedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, 2016b.

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

Chapter 1

Introduction

Most people in the world today use more than one language in the course of their daily lives (Grosjean and Li, 2013), and it is estimated that most children today grow up with exposure to two or more languages. Multilingualism has been recognized as an educational and social goal, for example by the European Union: the European Council resolution on a European Strategy for Multilingualism (2008/C 320/01) urged member states to "provide young people, from the earliest age... with a diverse and high-quality supply of language and culture education options enabling them to master at least two foreign languages, which is a factor of integration in a knowledge-based society." In addition to individual multilingualism, culture, technology and information are increasingly characterized by a globalization of resources, mediated by translation of materials originating in a wide variety of languages.

Under a broad interpretation, a *bilingual* is a person who can communicate in more than one language.[1] This definition encompasses balanced bilinguals, whose command of two (or more) languages is roughly equivalent, but also unbalanced speakers (sometimes called dominant or receptive bilinguals), who master one of the languages better than the other(s). To some extent, this definition also includes language learners, at various stages of command of the learned language. In all of these cases, the language produced by a speaker is affected by the simultaneous presence of another linguistic system. Granger (2015) uses *crosslingual language varieties* as an umbrella term for such language production; we adopt this term here.

Another situation where one linguistic system is clearly influenced by another is *translation*, namely conveying meaning expressed in one language to another language. This variety is often termed *translationese* (Gellerstam, 1986). Much research in Translation Studies indicates that translated texts have unique characteristics, distinguishing them from texts written originally in a certain language.

Several factors may account for the differences between originals and translations;

[1] See Gass and Selinker (2008, Table 2.1) for various definitions of bilinguals. For simplicity we speak about a *second* language system but, of course, there could be more language systems involved.

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 language 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. In this thesis we view translationese as a special case of a crosslingual language variety.

Intriguingly, although crosslingual language varieties have been the subject of extensive research, the different varieties have primarily been studied in isolation, by different research communities, each with its own established methodologies, and little awareness of research in the other fields. Bilingualism in general is investigated mainly in *psycholinguistics*, using methods that include experiments with human participants; learner language is the focus of research in *second language acquisition* (SLA), where corpus-based research has become increasingly popular in recent years; and *translation studies* focuses on translationese, using both corpus-based methods and, recently, more sophisticated computational machinery. We suggest that this fragmentation and lack of cross-disciplinary engagement might be obscuring important commonalities among crosslingual varieties.

The goal of this thesis is to investigate the characteristics of language production that is influenced by the existence of another linguistic system. Such language can be produced by a variety of bilinguals, including learners, advanced non-native speakers and translators. Our view is that this diverse population forms a continuum, where different speakers exhibit various degrees of proficiency and experience in the languages they use. This thesis brings together methods from corpus linguistics, psycho-linguistics and computer science to achieve a better understanding of how individuals use the various languages at their disposal. The findings can be used to inform language education and translation studies.

Chapter 2

Related Work

2.1 Translationese

Much research in Translation Studies indicates that translated texts have unique characteristics. Translated texts (in any language) constitute a sub-language (sometimes referred to as a *genre*, or a *dialect*) of the target language, presumably reflecting both the artifacts of the translation process and traces of the original language from which the texts were translated (the *source* language). Gellerstam (1986) called this sub-language *translationese*, and suggested that the differences between original (O) and translated (T) do not indicate poor translation but rather a *statistical phenomenon*, caused by a systematic influence of the source language on the target language.

Corpus-based investigation of translationese has been a prolific field of recent research, laying out an empirical foundation for the theoretically motivated hypotheses on the characteristics of translationese. More specifically, identification of translated texts by means of automatic classification shed light on the manifestation of translation universals and cross-linguistic influences as markers of translated texts (Baroni and Bernardini, 2006; van Halteren, 2008; Gaspari and Bernardini, 2008; Kurokawa et al., 2009; Koppel and Ordan, 2011; Ilisei and Inkpen, 2011; Volansky et al., 2015; Nisioi, 2015b), while Gaspari and Bernardini (2008) introduced a dataset for investigation of potential common traits between translations and non-native texts. Such studies proved to be important for the development of parallel corpora (Resnik and Smith, 2003), the improvement in quality of plagiarism detection (Potthast et al., 2011), language modeling, and statistical machine translation (Lembersky et al., 2012, 2013).

Indeed, translated texts are so markedly different from original ones that automatic classification can identify them with very high accuracy (Baroni and Bernardini, 2006; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Popescu, 2011). Recently, Volansky et al. (2015) investigated several translation studies hypotheses by performing an extensive exploration of the ability of various feature sets to distinguish between O and T. Using SVM classifiers and ten-fold cross-validation evaluation, they listed several features that yield near perfect accuracy.

Most works mentioned above train and evaluate classifiers on texts drawn from the same corpus. When these classifiers were tested on texts from different domains, or in a

different genre, or translated from a different language, classification accuracy dramatically deteriorated. Koppel and Ordan (2011) trained classifiers on the Europarl corpus (Koehn, 2005), with English translated from five different languages. When the classifiers were evaluated on English translated from the same language they were trained on, accuracy is near 100%; but when evaluated on translations from a different language, accuracy dropped significantly, in some cases below 60%. This pattern recurred when the test corpus was different from the training corpus (newspaper articles vs. parliament proceedings). Similarly, Avner et al. (2016) reported excellent (near 100%) results identifying Hebrew translationese on a corpus of literary texts, using very simple word-level features. Evaluation on different domains (popular science) and on Hebrew translated from French, rather than English, however, showed much poorer results, with accuracies around 60% in many cases.

We hypothesize that the main reason for the deterioration in the accuracy of (supervised) translationese classifiers when evaluated out-of-domain stems from the fact that domain differences overshadow the differences between O and T. Diwersy et al. (2014) studied various sorts of linguistic variation by applying semi-supervised multivariate techniques. They investigated, among other factors, register variation in English and German originals and translations. By applying a series of supervised and unsupervised statistical analyses, they demonstrated that register-related properties are much better exhibited by the underlying texts than properties related to the documents' translation status. We address these challenges, thoroughly analyze, and further elaborate on them in Chapter 3.

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).

We further hypothesize that due to the interference phenomenon (Section 1), 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. While other works reconstructing phylogenetic trees rely on multilingual data (Nagata and Whittaker, 2013; Berzak et al., 2014), in Chapter 4 we approach reconstruction of phylogenetic trees from texts using solely monolingual translations from multiple source languages.

2.2 The language of bilingual speakers

The language of bilinguals is different. The mutual presence of two linguistic systems in the mind of the bilingual speaker involves a significant cognitive load (Shlesinger, 2003; Hvelplund, 2014; Prior, 2014; Kroll et al., 2014); this burden is likely to have a bearing on the linguistic productions of the bilingual speaker. Moreover, the presence of more than one linguistic system gives rise to *transfer*: traces of one linguistic system may be observed in the other language (Jarvis and Pavlenko, 2008). Psycholinguistic research over the last two decades has demonstrated that the two languages of bilinguals mostly rely on shared neural substrates and cognitive resources (Abutalebi and Green, 2007; Kroll et al., 2012; Prior, 2014). Current models of the bilingual and learner language systems agree that there is at least some degree of sharing between the languages (French and Jacquet, 2004; Kroll and Tokowicz, 2005), and there is wide agreement that conceptual representations are likely shared across the two languages (Duñabeitia et al., 2010; Francis, 2005).

These findings are corroborated by research in second language acquisition, which established the unique character of learner language. The entire linguistic system that emerges when second language learners—both children and adults—express meaning in the target language was termed *interlanguage* (Selinker, 1972; Selinker and Rutherford, 2013). The interlanguage hypothesis assumes the concept of *fossilization*: permanent cessation of language learning *before* the learner has attained the target language norms at all levels of linguistic structure (Han, 2013). The fossilization hypothesis predicts that even advanced, fluent non-natives preserve traces that expose their non-nativeness. Some such traces are a result of the influence of the native language of the learner.

There are clear similarities between translations and non-native language: both are affected by the simultaneous presence of (at least) two linguistic systems, which may result in a higher cognitive load (Shlesinger, 2003). The presence of the L1 may also cause similar CLI effects on the target language. On the other hand, there are reasons to believe that translationese and non-native language should differ from each other. Translations are produced by *native* speakers of the target language. Non-natives, in contrast, arguably never attain native-like abilities (Coppieters, 1987; Johnson and Newport, 1991).

In Chapter 6 we put to test the hypotheses on the similarities between translationese and the language of non-native speakers using several corpus-based computational methods. We use supervised and unsupervised classification (Section 6.2) to show that the three language varieties are easily distinguishable. In particular, we show that native and advanced non-native productions can be accurately separated. We also demonstrate that non-native utterances and translations comprise two distinct linguistic systems; yet, non-native productions and translationese exhibit higher mutual proximity than either of them with text produced by native speakers.

Several works addressed the translation choices of bilingual speakers, either within a rich linguistic context (e.g., given a source sentence), or decontextualized. For example, de Groot (1992) demonstrated that cognate translations are produced more rapidly and accurately than translations that do not exhibit phonetic or orthographic similarity with a source word. This observation was further articulated by Prior et al. (2007), who showed that translation choices of L2 speakers were positively correlated with cross-linguistic form overlap of a stimulus word with its target language translations. Prior et al. (2011) emphasized that "bilinguals are sensitive to the degree of form overlap between the translation equivalents in the two languages, and show a preference toward producing a cognate translation". As an example, they showed that the preferred translation of the Spanish *incidente* to English was *incident*, and not the alternative translation *event*, despite the much higher frequency of the latter.

More recent work is consistent with previous research and advances it by highlighting phonologically mediated cross-lingual influences on visual word processing of same- and different-script bilinguals (Degani and Tokowicz, 2010; Degani et al., 2017). Cognate facilitation was also studied using eye tracking (Libben and Titone, 2009; Cop et al., 2017), demonstrating that the reading of bilinguals is influenced by orthographic similarity of words with their translation equivalents in another language. Crucially, much of this research has been conducted in a laboratory experimental setup; this implies a small number of participants, a small number of target words, and focus on a very limited set of languages. While our research questions are similar, we present a computational analysis of the effects of cognates on L2 productions on a completely different scale: 31 languages, over 1000 words, and thousands of speakers whose spontaneous language production is recorded in a very large corpus.

Computational approaches also proved beneficial for theoretical research in second language acquisition (Jarvis and Pavlenko, 2008). Numerous studies address linguistic processes attributed to SLA, including automatic detection of highly competent non-native writers (Tomokiyo and Jones, 2001; Bergsma et al., 2012), identification of the mother tongue of English learners (Koppel et al., 2005; Tetreault et al., 2013; Tsvetkov et al., 2013; Nisioi, 2015a) and typology-driven error prediction in learners' speech (Berzak et al., 2015). These studies are instrumental for language teaching and student evaluation (Smith and Swan, 2001), and can improve NLP applications such as authorship profiling (Estival et al., 2007) or grammatical error correction (Chodorow et al., 2010). Most of these studies utilize techniques that are motivated by the same abstract principles associated with L1 influences on the target language.

From the *lexical* perspective, L2 writers have been shown to produce more overgeneralizations, use more frequent words and words with a lower degree of ambiguity (Hinkel, 2002; Crossley and McNamara, 2011). Several studies addressed cross-linguistic influences on semantic acquisition in L2, investigating the distribution of collocations

(Siyanova-Chanturia, 2015; Kochmar and Shutova, 2017) and formulaic language (Paquot and Granger, 2012) in learner corpora. We, in contrast, address highly-fluent, advanced non-natives in this work.

Nastase and Strapparava (2017) presented the first attempt to leverage etymological information for the task of native language identification of English learners. They sowed the seeds for exploitation of etymological clues in the study of non-native language, but their results were inconclusive. In Chapter 5 we lay out sound empirical foundations for the theoretical hypothesis on the cognate effect in L2 of non-native English speakers, highlighting the cognate facilitation phenomenon as one of the important factors shaping the language of non-native speakers.

Chapter 3

Unsupervised Identification of Translationese

Human-translated texts (in any language) have distinct features that distinguish them from original, non-translated texts. These differences stem either from the effect of the translation process on the translated outcomes, or from "fingerprints" of the source language on the target language product. The term *translationese* was coined to indicate the unique properties of translations.

Awareness to translationese can improve statistical machine translation (SMT). First, for training translation models, parallel texts that were translated in the direction of the SMT task are preferable to texts translated in the opposite direction; second, for training language models, monolingual corpora of translated texts are better than original texts.

It is possible to automatically distinguish between original (O) and translated (T) texts, with very high accuracy, by employing text classification methods. Existing approaches, however, only employ *supervised* machine-learning; they therefore suffer from two main drawbacks: (i) they inherently depend on data annotated with the translation direction, and (ii) they may not be generalized to unseen (related or unrelated) domains.[1] These shortcomings undermine the usability of supervised methods for translationese identification in a typical real-life scenario, where no labelled in-domain data are available.

We explore *unsupervised* techniques for reliable discrimination of original and translated texts. More precisely, we apply *dimension reduction* and *centroid-based clustering* methods (enhanced by internal clustering evaluation), for telling O from T in an unsupervised scenario. Furthermore, we introduce a robust methodology for labelling the obtained clusters, i.e., annotating them as "original" or "translated", by inspecting similarities between the clustering outcomes and O and T *prototypical* examples. Rigorous experiments with four diverse corpora demonstrate that clustering of in-domain texts using lexical, content-independent features systematically yields very high accuracy, only 10 percent points lower than the performance of supervised classification on

[1] We use "domain" rather freely henceforth to indicate not only the topic of a corpus but also its modality (written vs. spoken), register, genre, date, etc.

the same data (in most cases). Accuracy can be improved even further by *clustering consensus* techniques.

We further scrutinize the tension between domain-related and translationese-based text properties. Using a series of experiments in a *mixed-domain* setup, we show that clustering (in particular, relying on content-independent features) perfectly groups the data into domains, rather than into the (desirable) cross-domain O and T; that is, domain-related properties clearly dominate and overshadow the translationese-based characteristics of the underlying texts. We address the challenge of discriminating O from T in a mixed-domain setup by proposing two simple methodologies (*flat* and *two-phase*) and empirically demonstrate their soundness.

The clustering experiments throughout this chapter were conducted in a setup similar to that of supervised classification, determining the status (O vs. T) of logical units (chunks) of 2,000 tokens. We also show that clustering accuracy remains stable even when the number of available chunks decreases dramatically and remains satisfactory when the chunk size is reduced.

The main contribution of this chapter is therefore two-fold: (i) we establish a robust approach for reliable unsupervised identification of translated texts, thereby eliminating the need for in-domain labeled data; (ii) we provide an extensive empirical foundation for the dominance of domain-based properties over translationese-related characteristics of a text, and propose a methodology for identification of translationese in a mixed-domain scenario.

3.1 Experimental Setup

3.1.1 Datasets

Our main dataset[2] consists of texts originally written in English and texts translated to English from French. We use various corpora:(i) Europarl, the proceedings of the European Parliament (Koehn, 2005), between the years 2001-2006; (ii) the Canadian Hansard, transcripts of the Canadian Parliament, spanning years 2001-2009; (iii) literary classics written (or translated) mainly in the 19th century; and (iv) transcripts of TED and TEDx talks. This collection suggests diversity in genre, register, modality (written vs. spoken) and era. Table 6.1 details some statistical data on the corpora (after tokenization).[3] We now briefly describe each dataset.

Europarl is probably the most popular parallel corpus in natural language processing, and it was indeed used for many of the translationese tasks surveyed in Section 2. This corpus has been used extensively in SMT (Koehn et al., 2009), and was even adapted specifically for research in translation studies: Islam and Mehler (2012) compiled a customized version of Europarl, where the direction of translation is indicated. We use a

[2]The dataset is available at `http://cl.haifa.ac.il/projects/translationese`.

[3]We use "EUR", "HAN", "LIT" and "TED" to denote the four corpora in the discussion below.

Corpus	Number of sentences			Number of tokens		Number of types	
	Original E	F→E	Total	Original E	F→E	Original E	F→E
EUR	134,725	71,816	206,541	3,406,513	2,112,085	37,203	28,119
HAN	3,441,984	757,573	4,199,557	65,491,960	13,457,613	158,645	63,192
LIT	36,123	85,210	121,333	858,297	1,750,525	25,113	38,842
TED	7,551	4,827	12,378	129,334	87,214	9,667	7,441

TABLE 3.1: Corpus statistics

version of Europarl (Rabinovich et al., 2016b) that aims to further increase the confidence in the direction of translation, through a comprehensive cross-lingual validation of the original language of the speakers.

The Hansard is a parallel corpus consisting of transcriptions of the Canadian parliament in English and French between 2001 and 2009. This is the largest available source of English–French sentence pairs. We use a version that is annotated with the original language of each parallel sentence. Relying on metadata available in the corpus, we filtered out all segments not referring to speech, i.e., retaining only sentences annotated as *Content ParaText*.

The Literature corpus consists of literary classics written (and translated) in the 18th–20th centuries by English and French authors; the raw material is available from the Gutenberg project. We use subsets that were manually or automatically paragraph-aligned. Note that classifying literary texts is considered a more challenging task than classifying more "technical" translations, such as parliament proceedings, since translators of literature typically enjoy more literary freedom, thereby rendering the translation product more similar to original writing (Lynch and Vogel, 2012; Avner et al., 2016).

Our TED talks corpus consists of talks originally given in English and talks translated to English from French. The quality of translations in this corpus is very high: not only are translators assumed to be competent, but the common practice is that each translation passes through a review before being published. This corpus consists of talks delivered orally, but we assume that they were meticulously prepared, so the language is not spontaneous but rather planned. Compared to the other sub-corpora, the TED dataset has some unique characteristics that stem from the following reasons: (i) its size is relatively small; (ii) it exhibits stylistic disparity between the original and translated texts (the former contains more "oral" markers of a spoken language, while the latter is a written translation); and finally (iii) TED talks are not transcribed but are rather subtitled, so they undergo some editing and rephrasing.[4]

The vast majority of TED talks are publicly available online, which makes this corpus easily extendable for future research.

[4] http://translations.ted.org/wiki/How_to_Compress_Subtitles

3.1.2 Processing and Tools

All datasets are first tokenized using the Stanford tools (Manning et al., 2014) and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary). We assume that translationese-related features are present in the texts across author or speaker, thus we allow some chunks to contain linguistic information from two or more different texts simultaneously. For the main (single-corpus) classification experiments we use 2000 text chunks each from Europarl and Hansard, 800 from Literature and 88 chunks from TED; each sub-corpus consists of an equal number of original and translated chunks. For every classification experiment we use the maximal equal number of chunks from each class, thus we always (randomly) down-sample the datasets in order to have a comparable number of training/testing examples for supervised classification, and comparable cluster size for clustering.

We use Weka (Hall et al., 2009) as the main tool for classification, clustering, and dimension reduction. In all the classification experiments, we use SVM (SMO) as the classification algorithm with the default linear kernel. For clustering we use Weka's KMeans implementation (SimpleKMeans) with the KMeans++ initialization strategy. We use Eucledian distance as the similarity measure for KMeans, and apply a custom clustering-evaluation-based wrapper (see Section 5.2.4) to further enhance Weka's basic clustering implementation.

We use Principal Component Analysis (PCA, Jolliffe (2002)) for dimension reduction. PCA is a statistical procedure that discovers variables with the largest possible variance, i.e., features that account for most variability in the data (*principal components*). It performs a linear mapping of the data to a lower-dimensional space in a way that maximizes the variance of the data in the low-dimensional representation, by removing highly correlated or superfluous variables. The outcome of PCA is a new set of features, each of which is a linear combination of the discovered components. The number of the newly generated variables varies from one to the number of variables originally used to describe the data, and is typically controlled by a parameter.

Apart from the enhanced efficiency (due to the reduced computational costs), dimensionality reduction often carries a positive effect on the accuracy of the underlying classification task, especially when the data are meager or feature vectors are sparse. The (accuracy-wise) optimization gains of PCA, when followed by the KMeans clustering algorithm, were reported by Ng et al. (2001). We perform dimension reduction using the Weka implementation of PCA, with the "variance_covered" parameter set to 0.1 across all feature types and datasets, prior to applying a clustering procedure.

3.1.3 Features

We focus on a set of features that reflect lexical and structural properties of the text, and have been shown to be effective for supervised classification of translationese (Volansky

et al., 2015). Specifically, we use *function words* (FW), more precisely, the same list that was used in previous works on classification of translationese (Koppel and Ordan, 2011; Volansky et al., 2015). Feature values are raw counts (further denoted by *term frequency, tf*), normalized by the number of tokens in the chunk; the chunk size may slightly vary, since the chunks respect sentence boundaries. For the clustering experiments we further scale the normalized *tf* by the *inverse document frequency (idf)*, which offsets the importance of a term by a factor proportional to its frequency in the corpus. The *tf-idf* statistic has been shown to be effective with *lexical* features, and is often used as a weighting factor in information retrieval and text mining. While function words are assumed to be very frequent, their counts within a text vary greatly (e.g., "the" vs. "whereas"). We therefore opt for *tf-idf* weighting of FW across all sub-corpora.

In addition to function words, we experiment with several other feature sets, including character trigrams, part-of-speech (POS) trigrams, *contextual function words* and *cohesive markers*. Contextual function words are a variation of POS trigrams where a trigram can be anchored by specific function words: these are consecutive triplets $\langle w_1,w_2,w_3\rangle$ where at least two of the elements are function words, and at most one is a POS tag. Cohesive markers are words or phrases that signal the underlying flow of thought: they organize a composition of phrases by specifying the type, purpose or direction of upcoming ideas, and can therefore serve as evidence of the translation process. We use the list of 40 cohesive markers defined in Volansky et al. (2015).

Character, POS, and contextual FW trigrams are calculated as detailed in Volansky et al. (2015), but we only consider the 1000 most frequent feature values extracted from each dataset (or a combination of datasets) being classified. This subset yields the same classification quality as the full set, reducing computation complexity.

3.2 Supervised Classification

We begin with supervised classification, re-establishing the high accuracy of in-domain (supervised) classification of translationese, but highlighting the deterioration in accuracy when cross-domain classification is considered. We first reproduce the Europarl classification results with the best performing feature sets, as reported by Volansky et al. (2015), and present results for three additional sub-corpora: Hansard, Literature and TED. Table 3.2 lists the ten-fold cross-validation classification accuracy with various features. All features (except perhaps cohesive markers) yield excellent accuracy.

A few previous works suggested that cross-domain classification of translationese results in low accuracy (Koppel and Ordan, 2011; Avner et al., 2016). Our experiments corroborate this observation; Table 3.3 depicts the cross-domain classification accuracy

feature / corpus	EUR	HAN	LIT	TED
FW	96.3	98.1	97.3	97.7
char-trigrams	98.8	97.1	99.5	100.0
POS-trigrams	98.5	97.2	98.7	92.0
contextual FW	95.2	96.8	94.1	86.3
cohesive markers	83.6	86.9	78.6	81.8

TABLE 3.2: In-domain (cross-validation) classification accuracy using various feature sets

on the Europarl, Hansard and Literature corpora, when training on one corpus and testing on another (using function words).[5] A balanced setup for this experiment was generated by randomly selecting 800 chunks from each corpus, divided equally to O and T. The results only slightly outperform chance level, even for the Europarl–Hansard seemingly domain-related pair: we obtain 59.7% to 60.8% accuracy in the two directions.

train / test	EUR	HAN	LIT	10-fold x-validation
EUR		60.8	56.2	94.7
HAN	59.7		58.7	98.1
LIT	64.3	61.5		97.3

TABLE 3.3: Pairwise cross-domain classification using function words

Attempting to enrich the classifier's training "experience" we conducted additional experiments, where we train on two sub-corpora out of Europarl, Hansard and Literature, and test on the remaining one. The results are depicted in Table 3.4. Here, too, accuracy is very low, implying that training on diverse data does not necessarily provide a solution for cross-domain classification of translationese. The right-hand column of the table reports ten-fold cross-validation results of the two sub-corpora that are subject for training. Excellent in-domain classification results on the one hand and poor cross-domain predictive performance on the other, imply that the model describing the relation in a certain domain is inapplicable to a different (even seemingly similar) domain due to significant differences in the distribution of the underlying data.

train / test	EUR	HAN	LIT	10-fold x-validation
EUR + HAN			63.8	94.0
EUR + LIT		64.1		92.9
HAN + LIT	59.8			96.0

TABLE 3.4: Leave-one-out cross-domain classification using function words

[5] We focus mainly on function words, because they are known to reflect stylistic differences rather than contents or specific corpus features, and are therefore less susceptible to domain overfitting. Other feature sets yielded similar results.

Reflecting the poor generalization capability of translationese features, these results call for developing other methodologies for reliably discriminating O from T, specifically, methodologies that are independent of in-domain labeled data.

3.3 Clustering

3.3.1 Initial results

To overcome the domain-dependence of supervised classification, we experiment in this section with unsupervised methods. We begin with the KMeans clustering algorithm, using KMeans++ initialization policy and dimension reduction. To evaluate the accuracy of the algorithms, each cluster is labeled by the majority of (O or T) instances it includes (using ground truth annotations), and the overall precision is the percentage of instances correctly assigned to their respective clusters (we discuss *unsupervised* cluster labeling in Section 3.3.2).

The KMeans clustering algorithm (with any initialization policy) is sensitive to the initial settings of its parameters, in particular the initial choice of centroids. A cluster *centroid* is the geometrical center of all observations within the cluster. The result of the KMeans algorithm may significantly vary according to its first step: the initial assignment of (random) points to cluster centroids. We address this potential pitfall by performing N clustering iterations, randomly varying the initial parameter settings, outputting the outcome that exhibits the highest similarity of points within a cluster. Formally, let C_i^j denote cluster i in iteration j, and let m_i^j denote this cluster's centroid, so that $i \in [1,2]$, and $j \in [1..N]$. *Sum-of-Square-Error (SSE)* is an intrinsic clustering evaluation metric that measures the similarity of elements in a cluster. The SSE of C_i^j is defined by

$$SSE_i^j = \sum_{x \in C_i^j} (x - m_i^j)^2$$

We aim to optimize the clustering result by choosing an outcome that minimizes the accumulative SSE:

$$\arg\min_j SSE^j = \arg\min_{j \in [1..N]} \sum_{i \in [1,2]} SSE_i^j$$

The selected clustering outcome represents the result of a *single* clustering experiment. The described method for selecting a clustering outcome can be viewed as a binary version of the *Bisecting* KMeans algorithm; it is applied in all experiments throughout the chapter, with number of iterations (N) fixed to 5, following the recommendation by Steinbach et al. (2000, p. 13).

We conducted a series of experiments with various feature sets; the main results are depicted in Table 3.5. The reported numbers reflect the average accuracy over 30 experiments (the only difference being a random choice of the initial conditions).[6]

[6] Standard deviation in most experiments was close to 0.

feature / corpus	EUR	HAN	LIT	TED
FW	88.6	88.9	**78.8**	**87.5**
char-trigrams	72.1	63.8	70.3	78.6
POS-trigrams	**96.9**	76.0	70.7	76.1
contextual FW	92.9	**93.2**	68.2	67.0
cohesive markers	63.1	81.2	67.1	63.0

TABLE 3.5: Clustering results using various feature sets

First and foremost, the results are very good, ranging from a few percent points lower than supervised classification (Table 3.2, Europarl and Hansard) to approximately 25 percent points lower in a few cases (e.g., Literature). Function words systematically yield very high accuracy; the quality of clustering with other features varies across the sub-corpora. Cohesive markers perform poorly (with a single exception, Hansard), which mirrors the moderate supervised classification precision achieved with the same feature set.

The exceptionally high result of Europarl with POS-trigrams can be attributed to the excessive frequency of specific phrases in the translated Europarl texts (in contrast to their original counterparts).[7] We explain the lower precision achieved on the Literature corpus by its diverse character: it comprises works attributed to a variety of authors, periods and genres, which is challenging for the unsupervised algorithm (see Section 3.4). A notably high accuracy is obtained on the small TED corpus, which implies the applicability of our clustering methodology to data-meager scenarios.

We conducted an additional set of experiments with unequal proportions of original and translated texts, considering twice the number of O chunks compared to T and vice versa. The average clustering accuracy using FW is similar to that obtained in the balanced setup (Table 3.5): 87.5% on Europarl, 88.9% on Hansard, 73.2% on Literature, and 88.6% on the TED sub-corpus.

3.3.2 Cluster labeling

As is always the case with unsupervised methods, clustering can divide observations into classes but cannot label those classes. A *cluster labeling* algorithm examines the contents of each cluster in order to find labels that best summarize its members, and distinguish the clusters from each other.

In the context of translationese identification, the task of cluster labeling is to determine which of the produced clusters represents O, and which T. We address this challenge by exploring similarities between the *language models* of the obtained clusters, and

[7] As an example (and in line with van Halteren (2008)), in the 2000 Europarl chunks, the phrase *ladies and gentlemen* appears 1258 times in T, but only 12 times in O.

language models of (presumably) *prototypical* O and T samples. A simple unigram language model assigns each word a probability proportional to its frequency in the underlying text; we use smoothed term frequencies scaled by the inverse total term frequencies. We then compare language models to reveal similarities between the prototypical O and T samples and the chunk sets produced by clustering.

The construction method of prototypical LMs is motivated by (i) abstracting from content, by utilizing only function words for this purpose; and (ii) attempting to avoid the interference of domain-related properties, by considering only (presumably) *universal* markers: words that share similar frequency patterns in several datasets w.r.t. to O vs. T.

Let O_m (O-markers) denote a set of function words that tend to be associated with O. We select this set by picking words whose frequency in O is excessive, compared to T; more precisely, the ratio of their frequency in O and T is above (1+δ), where δ=0.05. Similarly, T_m (T-markers) is a set of words with O-to-T frequency ratio below (1-δ). We create a prototypical O example by the concatenation of O_m, and a prototypical T example by the concatenation of T_m. The language model of these examples is then constructed by the ϵ-smoothed likelihood of each term in the markers vocabulary $V = O_m \bigcup T_m$, where ϵ=0.001.

Formally, for $w \in V$,

$$p(w \mid O_m) = \frac{tf(w) + \epsilon}{|O_m| + \epsilon \times |V|},\ p(w \mid T_m) = \frac{tf(w) + \epsilon}{|T_m| + \epsilon \times |V|}$$

We denote the resulting language models by P_O and P_T, respectively. Given two clusters, C_1 and C_2, we similarly compute their language models, denoted by P_{C_1} and P_{C_2}, respectively, over the vocabulary V. We measure the similarity between a class X (either O or T) and a cluster C_i using the Jensen-Shannon divergence (JSD) (Lin, 1991) on the respective probability distributions. Specifically, we define the *distance* between the language models as the square root of the divergence value, which is a metric, often referred to as *Jensen-Shannon distance*:

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X||P_{C_i})}$$

The assignment of the label X to the cluster C_1 is then supported by both C_1's proximity to the class X and C_2's proximity to the other class:

$$label(C_1) = \begin{cases} \text{“O”} & \text{if } D_{JS}(O, C_1) \times D_{JS}(T, C_2) \ < \ \alpha \times D_{JS}(O, C_2) \times D_{JS}(T, C_1) \\ \text{“T”} & \text{otherwise} \end{cases}$$

C_2 is assigned the complementary label. The value of α is fixed to 1 in this equation, but we note that it can be varied for further investigation of the relatedness of the underlying

language models.

We apply the cluster labeling technique described above to determine the labels of generated clusters. We construct prototypical O- and T-texts by selecting O- and T-markers from a random sample of Europarl and Hansard texts, using 600 chunks from each corpus.[8] We then compare the language models induced by these samples to those of the generated clusters (tested on different chunks, of course) to determine the cluster labels; the predicted labels are then verified against the majority-driven labeling, based on ground truth annotations. We apply this procedure to the outcome of all clustering experiments (per domain, using various features), achieving overall precision of 100%. In other words, the labeling procedure yields prefect accuracy not only on Europarl and Hansard texts that were not used for generation of O and T prototypical examples, but also on unseen Literature and TED datasets. We conclude that it is possible, in general, to determine the labels of clusters produced by our clustering algorithm with perfect accuracy.

3.3.3 Clustering consensus among feature sets

Since different feature sets have different predictions on our data, we hypothesize that consensus voting can improve the accuracy of clustering. We treat each individual clustering result (based on a certain feature set) as a judge, voting whether a single text chunk belongs to O or to T. We use the cluster labeling method of Section 3.3.2 to determine labels. The final assignment of a label to a cluster is determined by the majority vote of the various judges.

Table 3.6 presents the results of these experiments. We compare consensus results to the accuracy achieved by function words, the best-performing single feature set (on average), see Table 3.5. Both three judges and five judges yield a consistent increase in accuracy. Five judges systematically (and, on Europarl and Hansard, significantly) outperform the result of clustering with functions words only. This indicates that various features tend to capture different aspects of translationese, that are eventually leveraged by the "fusion" of different clustering results into a single, higher-quality outcome.

3.3.4 Sensitivity analysis

In supervised classification, the amount of labeled data has a critical effect on the classification accuracy. This does not seem to be the case with clustering: accuracy remains stable when the number of chunks used for classification decreases (Figure 3.1a). Evidently, as few as 300 chunks are sufficient for excellent classification.[9] We attribute the (slight) fluctuations in the graph to the random choice of the subset of chunks that are subject for clustering. Naturally, clustering accuracy stabilizes when the number of

[8] This subset of the Europarl and Hansard corpora was used for one-time generation of prototypical O and T language models, and excluded from further use.

[9] The results on the Literature corpus are limited by the amount of available data in this dataset.

method / corpus	EUR	HAN	LIT	TED
FW	88.6	88.9	78.8	87.5
FW char-trigrams POS-trigrams	91.1*	86.2	78.2	**90.9***
FW POS-trigrams contextual FW	**95.8***	89.8	72.3	86.3
FW char-trigrams POS-trigrams contextual FW cohesive markers	94.1*	**91.0***	**79.2**	88.6

TABLE 3.6: Clustering consensus by voting; statistically significant improvements, compared to using FW only, are marked with '*'

chunks increases, since the effect of random noise diminishes with more data. This result is of clear practical importance, as in real-life situations only a limited amount of data may be available.

The accuracy of supervised classification deteriorates when the size of the underlying logical units (here, chunks) decreases (Kurokawa et al., 2009). We corroborate this observation in the context of clustering, but note that reasonable accuracy (over 70%) can be obtained even with 1000-token chunks (Figure 3.1b). This further supports the applicability of unsupervised classification of translationese to real-world scenarios.

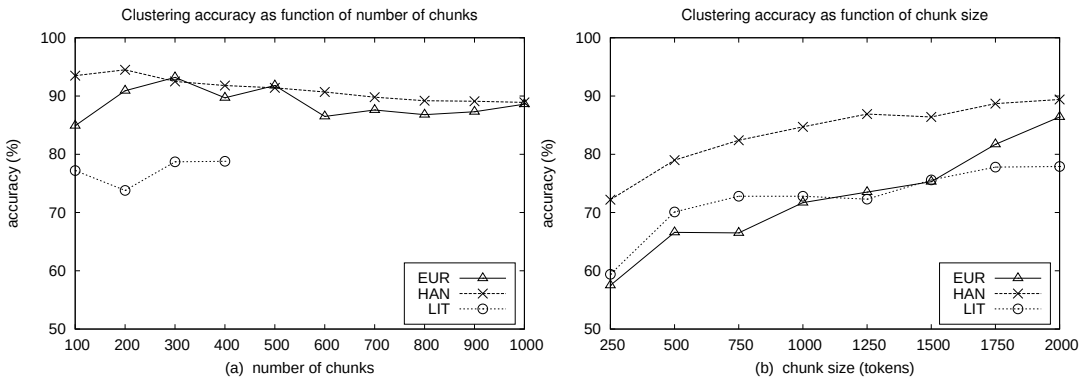

FIGURE 3.1: The effect of varying the number of chunks and chunk size (in tokens) on clustering accuracy

3.4 Mixed-domain classification

Poor cross-domain classification results, as described in Section 3.2, demonstrate that the in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains. In this section we explore the tension between the discriminative power of domain- and translationese-related properties, in the *unsupervised* scenario. Our underlying hypothesis is that domain-specific features overshadow the

features of translationese. The next series of experiments involves (a balanced) combination of various datasets; we excluded the small TED corpus from these experiments to prevent downsampling of other sub-corpora.

3.4.1 Domain-related vs. translationese-based characteristics

We begin with an investigation of the mutual effect of the domain- and translationese-specific characteristics on the accuracy of clustering. We first merged equal numbers of O and T chunks from two corpora: 800 chunks each from Europarl and Hansard, yielding 1,600 chunks, half of them O and half T. We applied the clustering algorithm of Section 5.2.4 to this dataset; the result was a perfect domain-driven separation of all Europarl and Hansard chunks, yielding poor (chance-level) translationese accuracy. In other words, we obtained two clusters, one consisting of Europarl chunks and the other of Hansard chunks, independently of their O-vs.-T status. We repeated the experiment with additional corpus pairs, and further extended it by adding equal numbers of Literature chunks (400 O and 400 T), this time fixing the number of clusters to three. Again, the result was separation by domain: Europarl, Hansard and Literature chunks were grouped into distinct clusters (Table 3.7, top).

As an additional experiment, we attempted to leave the decision on the "best" number of clusters to the algorithm. To that end, we employed the XMeans clustering procedure (Pelleg and Moore, 2000), which uses KMeans but applies additional statistical cues to decide on the number of clusters that best explain the data. We also applied PCA for dimension reduction prior to XMeans invocation. We repeated both experiments (two- and three-domain mixes) with XMeans, expecting to obtain two and three clusters, respectively. The result is a replication of the more constrained KMeans in three out of four cases (Table 3.7, bottom).

method / corpus	**EUR + HAN**	**EUR + LIT**	**HAN + LIT**	**EUR + HAN + LIT**
KMeans				
accuracy by domain	93.7	99.5	99.8	92.2
accuracy by translation status	50.3	50.0	50.0	–
XMeans				
generated # of clusters	2	2	3	3
accuracy by domain	93.6	99.5	99.9	92.2
accuracy by translation status	50.3	50.0	–	–

TABLE 3.7: Clustering a chunk-level mix of Europarl, Hansard and Literature using function words; accuracy by translation status (O vs. T) is reported where applicable (i.e., the outcome constitutes two clusters)

These observations have a crucial effect on understanding the tension between the domain- and translationese-based characteristics of the underlying texts. Not only are domains accurately separated given a fixed number of clusters, but even when the decision on the number of clusters is left to the clustering procedure, classification into domains explains the data best (as shown by XMeans). Recall that these experiments

all rely on the set of function words: topic-independent features, that have been proven effective for telling O from T in both supervised (Section 2) and unsupervised scenarios (Section 5.2.4). The fact that this translationese-oriented feature set yields the results presented in Table 3.7 clearly demonstrates the dominance of domain-specific properties over the characteristics of translationese.[10]

3.4.2 Clustering in a mixed-domain setup

Driven by the results of Section 3.4.1, we turn to explore a methodology for identification of translationese in a mixed-domain setup. We assume that we are given a set of text chunks that come from multiple domains, such that some chunks are O and some are T; the task is to classify the texts to O vs. T, *independently of their domain*. For that purpose, we investigate two approaches: *two-phase* and *flat*. Both methods assume that the number of domains, k, is known (it can be discovered by XMeans, as in Section 3.4.1, or fixed to a somewhat higher value than estimated in order to capture unsuspected differences within domains). The two-phase method first clusters a mixture of texts into domains (e.g., using KMeans), and then separates each of the resulting (presumably, domain-coherent) clusters into two sub-clusters, presumably O and T. The flat approach applies KMeans, attempting to divide the dataset into $2 \times k$ clusters; that is, we expect classification by domains and by translationese status, simultaneously.

method / corpus	**EUR + HAN**	**EUR + LIT**	**HAN + LIT**	**EUR + HAN + LIT**
Flat	**92.5**	60.7	77.5	66.8
Two-phase	91.3	**79.4**	**85.3**	**67.5**

TABLE 3.8: Flat and two-phase clustering of domain-mix using function words

We experimented with two setups:(i) mixture of two datasets out of Europarl, Hansard and Literature (1600 chunks in total); and (ii) mixture of all three of them (2400 chunks in total). We applied both methods to each of the two setups. We invoked PCA prior to clustering in the flat approach; in the two-phase approach, we applied PCA on *raw* data instances that are subject to clustering at each hierarchy level.[11] As our goal is identification of translationese, we define the accuracy of the classification as the ratio of O and T instances classified correctly (i.e., we ignore the accuracy of identifying the correct domain).

Table 3.8 reports the results. Both methods yield similarly high accuracy in the Europarl+Hansard setup, and much lower accuracy in the setup of all three datasets (with a single exception of EUR+LIT). This implies that the difficulty of telling O from T increases as the number of domains in the mixed-domain setup grows. The two-phase approach outperforms the flat one in most cases: the latter attempts to cluster data

[10] Other feature sets yielded similar outcomes.

[11] Note that our two-phase approach differs from the traditional hierarchical clustering in this sense.

instances by domain and translation status *simultaneously*, and is therefore potentially more error-prone. As a concrete example, in the Europarl+Literature setup, attempting to produce four clusters, we obtained a single cluster of Europarl chunks and three clusters of Literature chunks. The two-phase approach avoids such pitfalls by explicitly separating the steps of domain- and translationese-based clustering.

Table 3.8 clearly demonstrates that in a real-world scenario, where a dataset can be assumed to include texts from multiple domains, it is possible to overcome the dominance of domain-related features over translationese-related ones by splitting the task into two. The result is highly accurate identification of translated texts, even in an extremely challenging setup. Compare the results of Table 3.8 to the *supervised* case (Tables 3.3, 3.4): while clustering cannot compete with ten-fold cross-validation results of heterogenous datasets (93–96%), it is far superior to training a classifier on one or more datasets and then using it on a data from a new source (60–64%).

3.5 Discussion

Distinguishing between original and translated texts has been proven useful for SMT, as awareness to translationese can improve the quality of SMT systems. So far, classifying texts into original vs. translated has been done almost exclusively by supervised methods. In this work we advocated the use of *unsupervised* classification as an effective way to address this task. We demonstrated that simple feature sets, coupled with standard clustering algorithms, a novel cluster labeling technique, and voting among several features, can yield very high accuracy, over 90% in several cases.

Using diverse datasets we robustly demonstrated that the approach we advocate is effective for identification of translationese, even when only little data are available, and text chunks are small. We further highlighted the dominance of domain-based characteristics of the texts over their translationese-related properties and proposed a simple methodology for identification of translationese in a mixed-domain setup. We concluded that the proposed (two-phase) clustering approach is a robust method for distinguishing O from T in heterogenous datasets.

By conducting a series of experiments with unbalanced proportions of O and T texts, we demonstrated that the proposed methodology is also applicable to scenarios where the original and translated data are unevenly distributed.

We applied PCA for dimension reduction and the *tf-idf* weighting scheme with FW throughout all experiments in this work. The latter had a slight positive effect on clustering accuracy in most scenarios, and no impact in some cases. Dimension reduction improved computational efficiency, especially with large feature sets (e.g., character and POS trigrams). However, its effect on clustering accuracy was not uniform: the most prominent improvement (over 15 percent points) was obtained on the TED dataset, while a slight accuracy deterioration was observed in a few cases (e.g., 5 percent points

on Europarl with FW). We conclude that while carrying an overall positive value, the application of dimension reduction in similar scenarios calls for further investigation.

3.6 Conclusions and Future Work

To the best of our knowledge, this is the first work to extensively explore unsupervised classification of translationese. We only scratched the surface of this research direction. In the future, we intend to explore the robustness of our approach even further, with more datasets in various language pairs. We will first attempt to identify translationese in *French*, using the current dataset (in the reverse direction). We will also experiment with English-German, in both directions, and hopefully also with English-Hebrew, a more challenging setup.

The potential value of unsupervised identification of translationese leaves much room for further exploratory activities. Our future plans include using various datasets and reduced amount of data for LMs compiled for cluster labeling; in particular, we plan to explore the correlation between these two parameters and the scaling factor α used for association of a label with a clustering outcome.

Furthermore, to highlight the contribution of these results to SMT, we plan to replicate the results of Lembersky et al. (2012, 2013), using *predicted* rather than ground-truth indication of the translationese status of the texts that are used to train SMT systems. We believe that we will be able to show an improvement in the quality of SMT with extremely little supervision. We leave this direction for future research.

Relevant publications:

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

Chapter 4

Reconstructing Phylogenetic Language Trees from Translations

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 language 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 chapter 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.

4.1 Methodology

4.1.1 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., 2016b).

All parliament speeches were translated[1] from the original language into all other

[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.

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.

4.1.2 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 overrepresented 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.

[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.

4.1.3 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 sub-groupings 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 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 4.1 depicts the resulting gold-standard subtree.

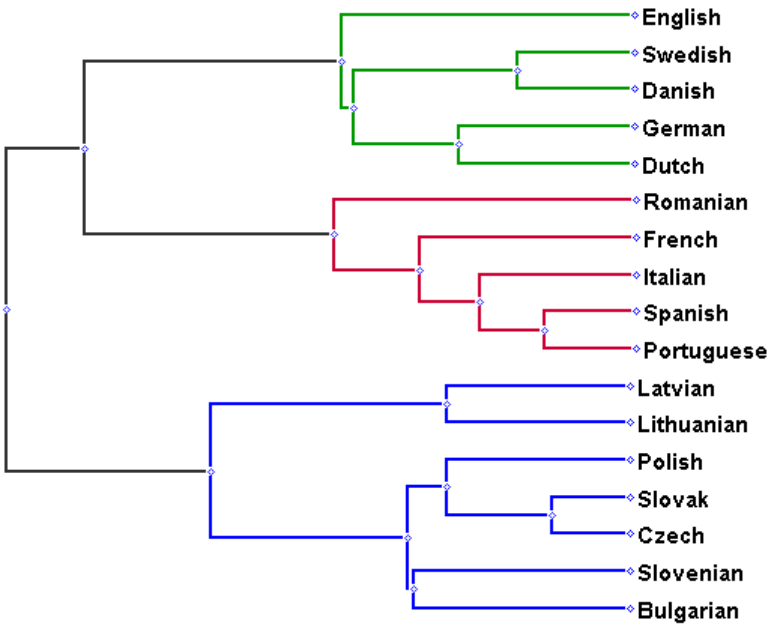

FIGURE 4.1: Gold standard tree, pruned

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.

4.1.4 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:

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

4.2 Detection of Translations and their Source Language

4.2.1 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 4.1 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 4.1: Classification accuracy (%) of English and French O vs. T

4.2.2 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.

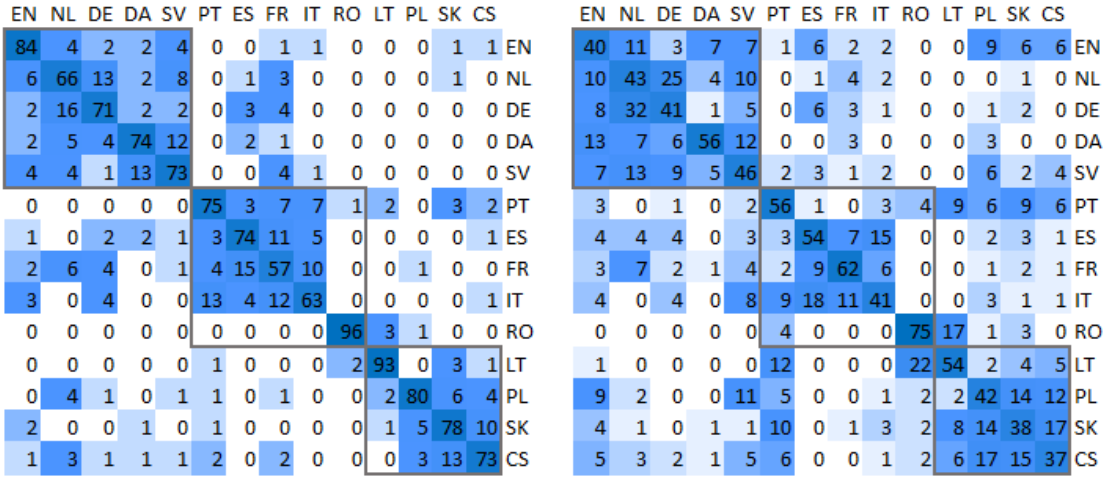

Figure 4.2: Confusion matrix of 14-way classification of English (left) and French (right) translations. The actual class is represented by rows and the predicted one by columns.

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 4.2 (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 4.2 (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.

4.3 Reconstruction of Phylogenetic Language Trees

4.3.1 Reconstructing language typology

Inspired by the results reported in Section 4.2.2, 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 4.1).

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 6.1.3). Figure 4.3 depicts the trees produced from English and French translations using POS-trigrams. Both trees reasonably group individual languages into three

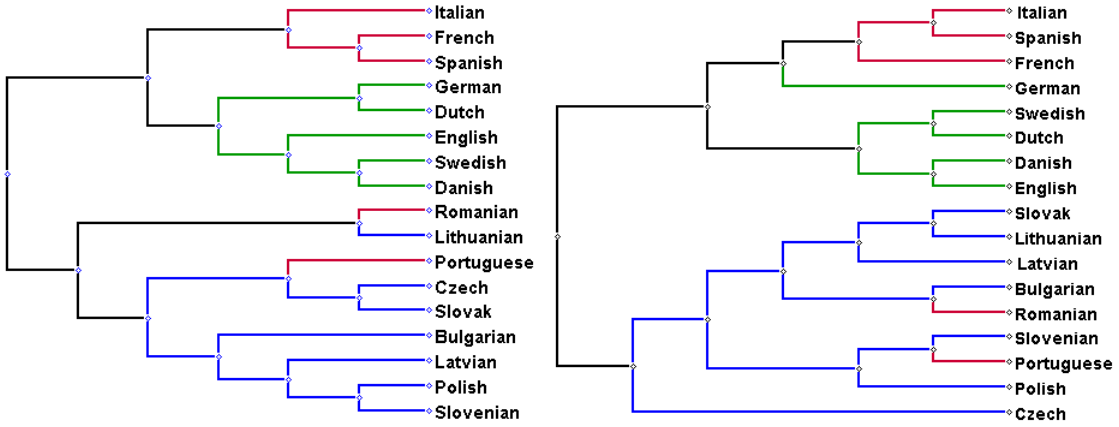

FIGURE 4.3: Phylogenetic language trees generated with English (left) and French (right) translations

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 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]

4.3.2 Evaluation results

The *unweighted* evaluation results are listed in Table 4.2. 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 4.3. 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 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 4). Furthermore, function words achieve more moderate

[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 4.2: 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 4.3: 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.

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 6.1.3). 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.

4.4 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

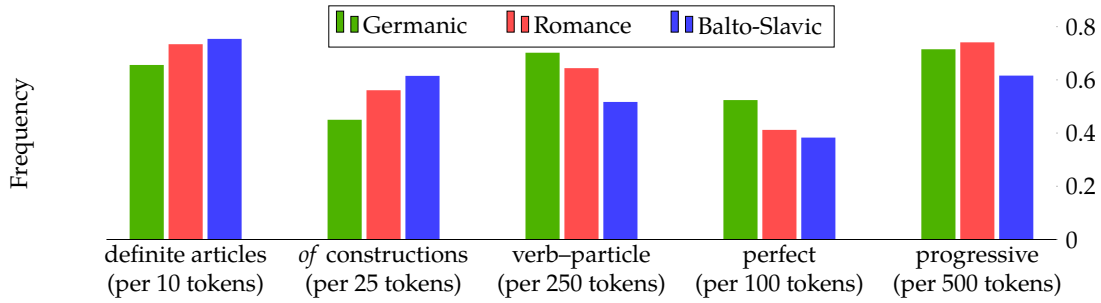

FIGURE 4.4: Frequencies reflecting various linguistic phenomena (Sections 4.4.1– 4.4.4) in English translations

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 4.4 depicts the results.

4.4.1 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.

4.4.2 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 (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.

4.4.3 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.

4.4.4 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.

4.5 Conclusion and Future Work

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.

Relevant publications:

Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL-2017)*, pp. 530–540. Association for Computational Linguistics, 2017a.

Chapter 5

Native Language Cognate Effects on Second Language Lexical Choice

Acquisition of vocabulary and semantic knowledge of a second language, including appropriate word choice and awareness of subtle word meaning contours, are recognized as a notoriously hard task, even for advanced non-native speakers. When non-native authors produce utterances in a foreign language (*L2*), these utterances are marked by traces of their native language (*L1*). Such traces are known as *transfer* effects, and they can be phonological (a foreign accent), morphological, lexical, or syntactic. Specifically, psycholinguistic research has shown that the choice of lexical items is influenced by the author's L1, and that non-native speakers tend to choose words that happen to have *cognates* in their native language.

Cognates are words in two languages that share both a similar meaning and a similar phonetic (and, sometimes, also orthographic) form, due to a common ancestor in some protolanguage. The definition is sometimes also extended to words that have similar forms and meanings due to *borrowing*. Most studies on cognate facilitation have been conducted with few human subjects, focusing on few words, and the experimental setup was such that participants were asked to produce lexical choices in an artificial setting. We demonstrate that cognates affect lexical choice in L2 spontaneous production on a much larger scale.

Using a new and unique large corpus of non-native English that we introduce as part of this work, we identify a *focus set* of over 1000 words, and show that they are distributed very differently across the "Englishes" of authors with various L1s. Importantly, we go to great lengths to guarantee that these words do not reflect specific properties of the various native languages, the cultures associated with them, or the topics that may be relevant for particular geographic regions. Rather, these are "ordinary" words, with very little culture-specific weight, that happen to have synonyms in English that may reflect cognates in some L1s, but not all of them. Consequently, they are used differently by authors with different linguistic backgrounds, to the extent that the authors' L1s can be identified through their use of the words in the focus set. The signal of L1 is so powerful, that we are able to reconstruct a linguistic typology tree from the distribution of these words in the Englishes witnessed in the corpus.

We propose a methodology for creating a focus set of highly frequent, unbiased words that we expect to be distributed differently across different Englishes simply because they happen to have synonyms with different etymologies, even though they carry very limited cultural weight. Then, we show that simple lexical semantic features (based on the focus set of words) suffice for clustering together English texts authored by speakers of "closer" languages; we generate a phylogenetic tree of 31 languages solely by looking at lexical semantic properties of the English spoken by non-native speakers from 31 countries.

The contribution of this work is twofold. First, we introduce the *L2-Reddit corpus*: a large corpus of highly-advanced, fluent, diverse, non-native English, with sentence-level annotations of the native language of each author. Second, we lay out sound empirical foundations for the theoretical hypothesis on the cognate effect in L2 of non-native English speakers, highlighting the cognate facilitation phenomenon as one of the important factors shaping the language of non-native speakers.

5.1 The *L2-Reddit corpus*

One contribution of this work is the collection, organization and annotation of a large corpus of highly-fluent non-native English. We describe this new and unique corpus in this section.

5.1.1 Corpus mining

Reddit[1] is an online community-driven platform consisting of numerous forums for news aggregation, content rating, and discussions. As of 2017, it has over 200 million unique users, ranking the fourth most visited website in the US. Content entries are organized by areas of interest called *subreddits*,[2] ranging from main forums that receive much attention to smaller ones that foster discussion on niche areas. Subreddit topics include news, science, movies, books, music, fitness and many others.

Collection of author metadata We collected a large dataset of posts (both initial submissions and subsequent comments) using an API especially designed for providing search capabilities on Reddit content.[3] We focused on several subreddits (r/Europe, r/AskEurope, r/EuropeanCulture, r/EuropeanFederalists, r/Eurosceptics) whose content is generated by users who specified their country as a *flair* (metadata attribute). Although categorized as 'European', these subreddits are used by people from all over the world, expressing views on politics, legislation, economics, culture, etc.

In the absence of a restrictive policy, multiple flair alternatives often exist for the same country, e.g., 'CROA' and 'Croatia' for Croatia. Additionally, distinct flairs are

[1] https://www.reddit.com/

[2] Subreddits are typically denoted with a leading r/, for example r/linguistics is the 'linguistics' subreddit.

[3] https://github.com/pushshift/api

sometimes used for regions, cities, or states of big European countries, e.g., 'Bavaria' for Germany. We (manually) grouped flairs representing the same country into a single cluster, reducing 489 distinct flairs into 50 countries, from Albania to Vietnam. The posts in the Europe-related subreddits constitute our *seed corpus*, comprising 9M sentences (160M tokens) by over 45K distinct users.

Dataset expansion A typical user activity in Reddit is not limited to a single thread, but rather spreads across multiple, not necessarily related, areas of interest. Once the authors' country is determined based on their European submissions, their entire Reddit footprint can be associated with their profile, and, therefore, with their country of origin. We extended our seed corpus by mining *all* submissions of users whose country flair is known, querying all Reddit data spanning years 2005-2017. The final dataset thus contains over 250M sentences (3.8B tokens) of native and non-native English speakers, where each sentence is annotated with its author's country of origin. The data covers posts by over 45K authors and spans over 80K subreddits.[4]

Focus on "large" languages For the sake of robustness, we limited the scope of this work to (countries whose L1s are) the Indo-European (IE) languages; and only to those countries whose users had at least 500K sentences in the corpus. Additionally, we excluded multilingual countries, such as Belgium and Switzerland. Consequently, the final set of Reddit authors considered in this work originate from 31 countries, which represent the three main IE language families: *Germanic* (Austria, Denmark, Germany, Iceland, Netherlands, Norway, Sweden); *Romance* (France, Italy, Mexico, Portugal, Romania, Spain); and *Balto-Slavic* (Bosnia, Bulgaria, Croatia, Czech, Latvia, Lithuania, Poland, Russia, Serbia, Slovakia, Slovenia, Ukraine). In addition, we have data authored by native English speakers from Australia, Canada, Ireland, New Zealand, the UK and the US.

Correlation of country annotation with L1 We view the country information as an accurate, albeit not perfect, proxy for the native language of the author.[5] We acknowledge that the L1 information is noisy and may occasionally be inaccurate. We therefore evaluated the correlation of the country flair with L1 by means of supervised classification: our assumption is that if we can accurately distinguish among users from various countries using features that reflect language, rather than culture or content, then such a correlation indeed exists.

We assume that the native language of speakers "shines through" mainly in their syntactic choices. Consequently, we opted for (shallow) syntactic structures, realized by function words (FW) and n-grams of part-of-speech (POS) tags, rather than geographical and topical markers, that are reflected best by content words. Aiming to disentangle the

[4]The annotated dataset will freely available at http://cl.haifa.ac.il/projects/L2. To protect the anonymity of Reddit users, the released dataset does not expose any author identifying information.

[5]We therefore use the terms 'user country', 'native language' and 'L1' interchangeably henceforth.

effect of native language we randomly shuffled texts produced by all authors from each country, thereby "blurring out" any topical (i.e., subreddit-specific) or authorial trace. Consequently, we assume that the separability of texts by country can be attributed to the only distinguishing linguistic variable left: the dimension of the native language of a speaker.

We classified 200 chunks of 100 randomly sampled sentences from each country into (i) native vs. non-native English speakers, (ii) the three IE language families, and (iii) 45 individual L1s, where the six English-speaking countries are unified under the native-English umbrella. Using over 400 function words and top-300 most frequent POS-trigrams, we obtained 10-fold cross-validation accuracy of 90.8%, 82.5% and 60.8%, for the three scenarios, respectively. We conclude, therefore, that the country flair can be viewed as a plausible proxy for the native language of Reddit authors.

Initial preprocessing Several preprocessing steps were applied on the dataset. We (i) removed text by users who changed their country flair within their period of activity; (ii) excluded non-English sentences;[6] and (iii) eliminated sentences containing single non-alphabetic tokens. The final corpus comprises over 230M sentences and 3.5B tokens.

5.1.2 Evaluation of author proficiency

Unlike most corpora of non-native speakers, which focus on *learners* (e.g., ICLE (Granger, 2003), EFCAMDAT (Geertzen et al., 2013), or the TOEFL dataset (Blanchard et al., 2013)), our corpus is unique in that it is composed by fluent, advanced non-native speakers of English. We verified that, on average, Reddit users possess excellent, near-native command of English by comparing three distinct populations: (i) Reddit native English authors, defined as those tagged for one of the English-speaking countries: Australia, Canada, Ireland, New Zealand, and the UK. We excluded texts produced by US authors due to the high ratio of the US immigrant population; (ii) Reddit non-native English authors; and (iii) A population of English learners, using the TOEFL dataset (Blanchard et al., 2013); here, the proficiency of authors is classified as low, intermediate, or high.

We compared these populations across various indices, assessing their proficiency with several commonly accepted lexical and syntactic complexity measures (Lu and Ai, 2015; Kyle and Crossley, 2015). Lexical richness was evaluated through type-to-token ratio (TTR), average age-of-acquisition (in years) of lexical items (Kuperman et al., 2012), and mean word rank, where the rank was retrieved from a list of the entire Reddit dataset vocabulary, sorted by word frequency in the corpus. Syntactic complexity was assessed using mean length of T-units (TU; the minimal terminable unit of language that can be considered a grammatical sentence), and the ratio of complex T-units (those containing a dependent clause) to all T-units in a sentence.

[6] We used the *polyglot* language detection tool (`http://polyglot.readthedocs.io`).

Population	Mean TU length	Complex TU ratio	TTR	Mean word rank	AoA
Learners (low)	15.583	0.513	0.089	1172.19	5.186
Learners (medium)	16.357	0.534	0.106	1504.01	5.317
Learners (high)	17.468	0.528	0.124	1852.64	5.562
Reddit non-natives	*19.528*	*0.633*	*0.174*	*1960.62*	*5.524*
Reddit natives	20.154	0.658	0.179	2063.89	5.575

TABLE 5.1: Evaluation of the English proficiency of non-native Reddit users.

Table 5.1 reports the results. Across almost all indices, the level of Reddit non-natives is much higher than even the advanced TOEFL learners, and almost on par with Reddit natives.

5.2 L1 cognate effects on L2 lexical choice

5.2.1 Hypotheses

Cognates are words in two languages that share both a similar meaning and a similar form. Our main hypothesis is that non-native speakers, when required to pick an English word that has a set of synonyms, are more likely to select a lexical item that has a cognate in their L1. We therefore expect the effect of L1 cognates to be reflected in the frequency of their English counterparts in the spontaneous productions of L2 speakers. Moreover, we expect similar effects, perhaps to a lesser extent, in the contextual usage of certain words, reflecting collocations and subtle contours of word meanings that are transferred from L1. The different contexts that certain words are embedded in (in the Englishes of speakers with different L1 backgrounds) can be captured by the means of distributional semantics.

Furthermore, we hypothesize that the effect of L1 is powerful to an extent that facilitates clustering of Englishes produced by non-natives with "similar" L1s; specifically, L1s that belong to the same language family. "Similar" L1s may reflect both typological and areal closeness: for example, we expect the English spoken by Romanians to be similar both to the English of Italians (as both are Romance languages) and to the English of Bulgarians (as both are Balkan languages). Ultimately, we aim to reconstruct the IE language phylogeny, reflecting historical and areal evolution of the subsets of Germanic, Romance and Balto-Slavic languages over thousands of years, from non-native English only.

While lexical transfer from L1 is a known phenomenon in *learner* language, we hypothesize that its signal is present also in the language of highly competent non-native speakers. Mastering the nuances of lexical choice, including subtle contours of word meaning and the correct context in which words tend to occur, are key factors in advanced language competence. The L2-Reddit corpus provides a perfect environment for testing this hypothesis.

5.2.2 Selection of a focus set of words

Our goal is to investigate non-native speakers' choice of lexical items in English. We address this task by defining a set of English words that have at least one synonym; ideally, we would like the various synonyms to have different etymologies, and in particular, to have different cognates in different language families. English happens to be a particularly good choice for this task, since in spite of its Germanic origins, much of its vocabulary evolved from Romance, as a great number of words were borrowed from Old French during the Norman occupation of Britain in the 11th century.

To trace the etymological history of English words we used Etymological WordNet (EW), a database that contains information about the ancestors of over 100K English words, about 25K of them in contemporary English (de Melo, 2014). For each word recorded in EW, the full path to its root can be reconstructed. Intuitively, an English word with Latin roots may exhibit higher (phonetic and orthographic) proximity to its Romance languages' counterparts. Conversely, an English word with a Proto-Germanic ancestor may better resemble its equivalents in Germanic languages.

We selected from EW all the nouns, verbs, and adjectives. For each such word w, we identified the synset of w in WordNet, choosing only the first (i.e., most prominent) sense of w (and, in particular, corresponding to the most frequent part-of-speech (POS) category of w in the L2-Reddit dataset). Then, we retained only those words that had synonyms, and only those whose synonyms had at least two different etymological paths, i.e., synonyms rooted in different ancestors. For example, we retained the synset {*heaven*, *paradise*}, since the former is derived from Proto-Germanic **himin-* , while the latter is derived from Greek (via Latin and Old French).

Furthermore, to capture the bias of non-native speakers toward their L1 cognate, it makes sense to focus on a set of easily interchangeable synonyms, e.g., {*divide*, *split*}. In contrast, consider an unbalanced synset {*kiss*, *buss*, *osculation*}: presumably, the prevalent alternative *kiss* is likely to be used by all speakers, regardless of their native language. To eliminate such cases, we excluded synsets that were dominated by a single alternative (with a frequency of over 90% in our corpus), compared to other synonymous choices. Table 5.2 illustrates a few examples of synonym sets with their etymological origins.

Synonym set	**Etymological path to root**
cargo (N)	Spanish: *cargo* ← Spanish: *cargar* ← Late Latin: *carricare*
freight (N)	Mid. English: *freyght* ← Mid. Low German: *vrecht* ← Proto-Germanic **fra-* + **aihtiz*
weary (Adj)	Mid. English: *wery* ← Old English: *wērig* ← Proto-Germanic: **wōrīgaz*
fatigue (Adj)	French: *fatigue* ← French: *fatiguer* ← Latin: *fatigare*
exaggerate (V)	Latin: *exaggerare* ← Latin: *ex-* + Latin: *aggerare*
overdo (V)	English: *over* + *do*

Table 5.2: Etymological roots of example synonym sets with corresponding part-of-speech.

Eliminating cultural bias Although our Reddit corpus spans over 80K topical threads and 45K users, posts produced by authors from neighboring countries may carry over markers with similar geographical or cultural flavor. For example, we may expect to encounter *soviet* more frequently in posts by Russians and Ukrainians, *wine* in texts of French or Italian authors, and *refugees* in posts by German users. While they may be typical to a certain population group, such terms are totally unrelated to the phenomenon we address here, and we therefore wish to eliminate them from the focus set of words.

A common way to identify elements that are statistically over-represented in a particular population, compared to another, is *log-odds ratio informative Dirichlet prior* (Monroe et al., 2008). We employed this approach to discover words that were overused by authors of a certain country, where posts from each country (a category under test) were compared to all the others (the background). We used the strict log-odds score of -5 as a threshold for filtering out terms associated with a certain country.[7] Among the terms eliminated by this procedure were *genocide* for Armenia, *hockey* for Canada and *independence* for the UK. The final focus set of words thus consists of neutral, ubiquitous sets of synonyms, varying in their etymological roots. It comprises 540 synonym sets and 1143 distinct words.

5.2.3 Model

We hypothesize (Section 5.2.1) that L1 effects on lexical choice are so powerful, even with advanced non-native speakers, that it is possible to reconstruct the IE language phylogeny, reflecting historical and areal evolution over thousands of years, from non-native English only. We now describe a simple yet effective framework for clustering the Englishes of authors with different L1s, integrating both word frequencies and semantic word representations of the words in our focus set (Section 5.2.2).

Data cleanup and abstraction

Aiming to learn word representations for the lexical items in our focus set, we want the contextual information to be as free as possible from strong geographical and cultural cues. We therefore process the corpus further. First, we identified named entities (NEs) and systematically replaced them by their type. We used the implementation available in the *spacy* Python package,[8] which supports a wide range of entities (e.g., names of people, nationalities, countries, products, events, book titles, etc.), at state-of-the-art accuracy. Like other web-based user generated content, the Reddit corpus does not adhere to strict casing rules, which has detrimental effects on the accuracy of NE identification. To improve the tagging accuracy, we applied a preprocessing step of 'truecasing', where each token w was assigned the case (lower, upper, or upper-initial) that maximized the likelihood of the consecutive tri-gram $\langle w_{pre}, w, w_{post} \rangle$ in the Corpus of Contemporary

[7] The threshold was set by preliminary experiments, without any further tuning.
[8] https://spacy.io

American English (COCA).[9] For example, the tri-gram 'the us people' was converted to 'the US people', but 'let us know' remained unchanged. When a tri-gram was not found in the COCA n-gram corpus, we employed fallback to unigram probability estimation. Additionally, we replaced all non-English words with the token 'UNK'; and all web links, subreddit (e.g., r/compling) and user (u/userid) pointers with the 'URL' token.[10]

Distance estimation and clustering

Bamman et al. (2014) introduced a model for incorporating contextual information (such as geography) in learning vector representations. They proposed a joint model for learning word representations in a situated language, a model that "includes information about a subject (i.e., the speaker), allowing to learn the contours of a word's meaning that are shaped by the context in which it is uttered". Using a large corpus of tweets, their joint model learned word representations that were sensitive to geographical factors, demonstrating that the usage of *wicked* in the United States (meaning *bad* or *evil*) differs from that in New England, where it is used as an adverbial intensifier (*my boy's wicked smart*).

We leveraged this model to uncover linguistic variation grounded in the different L1 backgrounds of non-native Reddit speakers. We used equal-sized random samples of 500K sentences from each country to train a model of vector representations. The model comprises representation of every vocabulary item in each of the 31 Englishes; e.g., 31 vectors are generated for the word *fatigue*, presumably reflecting the subtle divergences of word semantics, rooted in the various L1 backgrounds of the authors.

In order to cluster together Englishes of speakers with "similar" L1s, we need a measure of distance between two English texts. This measure is based on two constituents: word frequencies and word embeddings. Given two English texts originating from different countries, we computed for each word w in our focus set (i) the difference in the frequency of w in the two texts; and (ii) the distance between the vector representations of w in these texts, estimated by cosine similarity of the two corresponding word vectors. We employed the popular *weighted product model* to integrate the two arguments. The word vector component was assigned a higher weight as the frequency of w in the collection increases; this is motivated by the intuition that learning the semantic relationships of a word benefits from vast usage examples. We therefore weigh the embedding constituent proportionally to the word's frequency in the dataset, and assign the complementary weight to the difference of frequencies.

Formally, given two English texts E_{L_i} and E_{L_j}, with L_i and L_j native languages, and given a word w in the focus set, let f_i and f_j denote the frequencies of w in E_{L_i} and E_{L_j},

[9] https://www.ngrams.info

[10] The cleaned, abstracted subset of the corpus is also available at http://cl.haifa.ac.il/projects/L2. The cleanup code is available at https://github.com/ellarabi/reddit-l2.

respectively. Let p_w be the probability of w in the entire collection. We further denote the vector space representation of w in E_{L_i} by v_i, and the representation of w in E_{L_j} by v_j. Then, the distance between E_{L_i} and E_{L_j} with respect to the word w is:

$$D_{ij}(w) = (|f_i - f_j|)^{1-p_w} \times (1 - cos(v_i, v_j))^{p_w}. \quad (5.1)$$

The final distance between E_{L_i} and E_{L_j} is given by averaging D_{ij} over all words in the focus set FS:

$$D_{ij} = \frac{(\sum_{w \in FS} D_{ij}(w))}{|FS|}.$$

Finally, we constructed a symmetric distance matrix (31×31) M by setting $M[i, j] = D_{ij}$. We used Ward's hierarchical clustering[11] with the Euclidean distance metric to derive a tree from the distance matrix M.

We considered several other weighting alternatives, including assignment of constant weights to the two factors in Equation 5.1; they all resulted in inferior outcomes. We also corroborated the relative contribution of the two components by using each of them alone. While considering only frequencies resulted in a slightly inferior outcome (see Section 5.2.5), using word representations alone produced a completely arbitrary result.

5.2.4 Results

The resulting tree is depicted in Figure 5.1. The reconstructed language typology reveals several interesting observations. First, and much expectedly, all native English speakers are grouped together into a single, distant sub-tree, implying that similarities exhibited by the lexical choices of native speakers go beyond geographical and cultural differences. The Englishes of non-native speakers are clustered into three main language families: Germanic, Romance, and Balto-Slavic. Notably, Spanish-speaking Mexico is clustered with its Romance counterparts. The firm Balto-Slavic cluster reveals historical relations between languages by generating coherent sub-branches: the Czech Republic and Slovakia, Latvia and Lithuania, as well as the relative proximity of Serbia and Croatia. In fact, former Yugoslavia is clustered together, except for Bosnia, which is somewhat detached. Similar close ties can be seen between Austria and Germany, and between Portugal and Spain.

Another interesting phenomenon is captured by English texts of authors from Romania: their language is assigned to the Balto-Slavic family, implying that the deep-rooted areal and cultural Balkan influences left their traces in the Romanian language, which in turn, is reflected in the English productions of native Romanian authors. Unfortunately, we cannot explain the location of Iceland.

[11] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

A geographical view mirroring the language phylogeny is presented in Figure 5.3. Flat clusters were obtained from the hierarchy using the *scipy fcluster* method[12] with defaults.

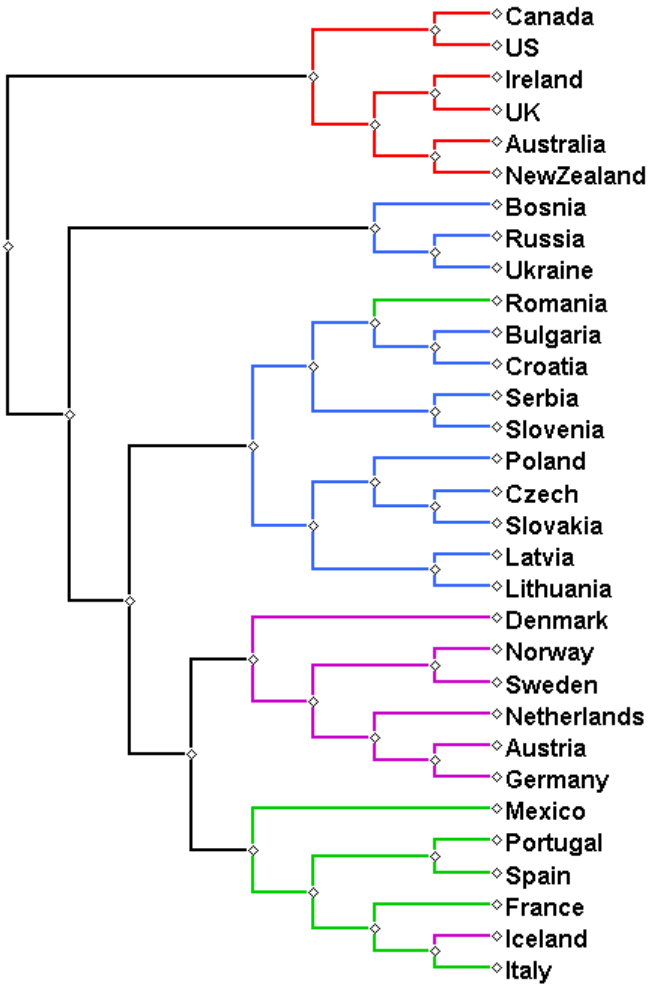

FIGURE 5.1: Language typology reconstructed from non-native Englishes using features reflecting lexical choice. Countries that belong to the same phylogenetic family (according to the gold tree) share identical color. E.g., Iceland is colored purple, like other Germanic languages, even though it is assigned to the Romance cluster.

This outcome, obtained using only lexical semantic properties (word frequencies and word embeddings) of English authored by various non-native speakers, is a strong indication of the power of L1 influence on L2 speakers, even highly fluent ones. These results are strongly dependent on the choice of focus words: we carefully selected words that on one hand lack any cultural or geographical bias toward one group of non-natives, but on the other hand have synonyms with different etymologies. As an additional validation step, we generated a language tree using exactly the same methodology but a different set of focus words. We randomly sampled 1143 words from the corpus, controlling for

[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html

country-specific bias but *not* for the existence of synonyms with different etymologies. Although some of the intra-family ties were captured (in particular, all native speakers were clustered together), the resulting tree (Figure 5.2) is far inferior.

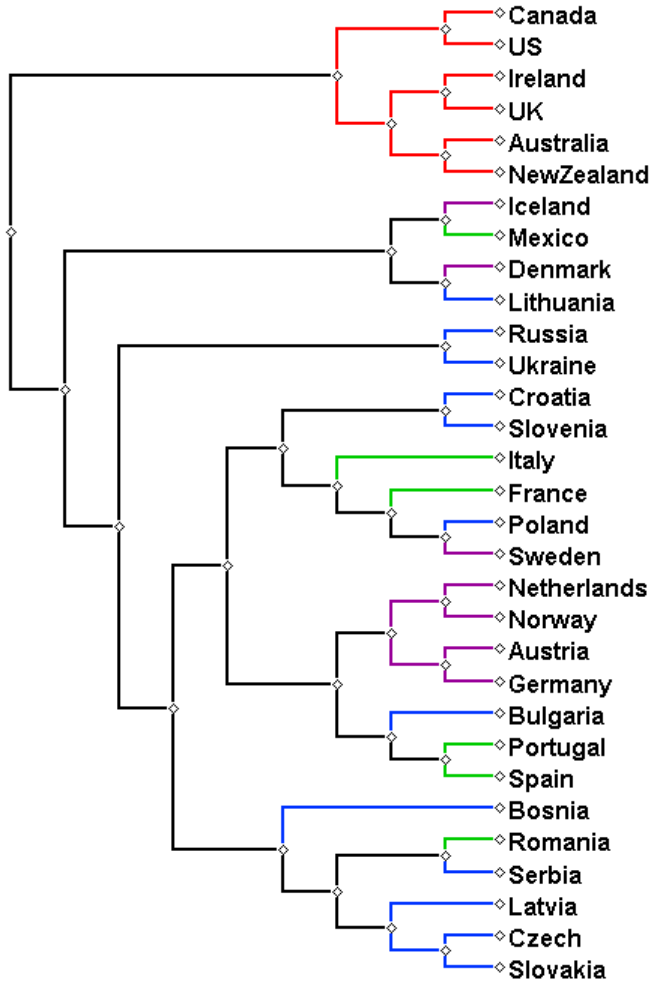

FIGURE 5.2: Language typology reconstructed from a randomly selected focus set of 1143 words.

We also conducted an additional experiment, including multilingual Belgium and Switzerland in the set of countries. While the L1 of speakers cannot be determined for these two countries, presumably Belgium is dominated by Dutch and French, and Switzerland by German and French. Indeed, both countries were assigned into the Germanic language family in our clustering experiments.

5.2.5 Evaluation

To better assess the quality of the reconstructed trees we now provide a quantitative evaluation of the language typologies obtained by the various experiments. We adopt the evaluation approach of Rabinovich et al. (2017a), who introduced a distance metric between two trees, defined as the sum of the square differences between all leaf-pair distances in the two trees. More specifically, given a tree of N leaves, l_i, $i \in [1..N]$, the

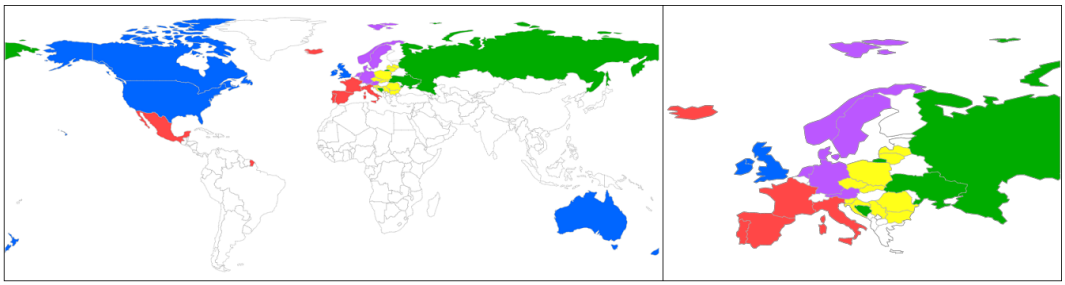

FIGURE 5.3: Countries by clusters: World (on the left) and Europe (on the right) views. Countries assigned to the same flat cluster by the *clustering procedure* (Section 5.2.4) share identical color, e.g., the wrongly assigned Iceland shares the red color with the Romance-language speaking countries. Countries not included in this work are uncolored.

distance between two leaves l_i, l_j in a tree τ, denoted $D_\tau(l_i, l_j)$, is defined as the length of the shortest path between l_i and l_j. 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 in the two trees:

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

We used the Indo-European tree in *Glottolog*[13] as our gold standard, pruning it to contain the set of 31 languages considered in this work. For the sake of comparison, we also present the distance obtained for a completely random tree, generated by sampling a random distance matrix from the uniform (0, 1) distribution. The reported random tree evaluation score is averaged over 100 experiments.

Table 5.3 presents the results. All distances are normalized to a zero-one scale, where the bounds, zero and one, represent the identical and the most distant tree with respect to the gold standard, respectively. Much expectedly, the random tree is the worst one, followed closely by the tree reconstructed from a random sample of over 1000 words sampled from the corpus (Figure 5.2). The best result is obtained by considering both word frequencies and representations, being only slightly superior to the tree reconstructed using word frequencies alone. The latter result corroborates the aforementioned observation (Section 5.2.3) and further posits word frequencies as the major factor affecting the shape of the obtained phylogeny.

5.3 Analysis

The results described in Section 5.2.4 empirically support the intuition that cognates are one of the factors that shape lexical choice in productions of non-native authors. In this section we perform a closer analysis of the data, aiming to capture the subtle

[13] http://glottolog.org/

Features used	Distance
Random tree	1.000
Randomly sampled words (Figure 5.2)	0.857
Focus set with frequencies only	0.497
+ embeddings (Figure 5.1)	**0.469**

TABLE 5.3: Normalized distance between a reconstructed and the gold tree; lower distances indicate better result.

yet systematic distortions that help distinguish between English texts of speakers with different L1s.

Quantitative analysis Given a synonym set $s \in FS$, consisting of words $\langle w_1, w_2, ..., w_n \rangle$, and two English texts with two different L1s, E_{L_i} and E_{L_j}, we computed the counts of the synset words in these texts, and further normalized the counts by the total sum, yielding probabilities. We denote the probability distribution of a synset $s = \langle w_1, w_2, ..., w_n \rangle$ in E_{L_i} by:

$$P_i^s = \langle p_i(w_1), p_i(w_2), ..., p_i(w_n) \rangle.$$

The different usage patterns of a synonym set s across two Englishes can then be estimated using the Jensen-Shannon divergence (JSD) between the two probability distributions:

$$div_{ij}(s) = JSD(P_i^s, P_j^s). \tag{5.2}$$

We expect that "close" L1s will have lower divergence, whereas L1s from different language families will exhibit higher divergences.

Table 5.4 presents the top twenty synonym sets for the arbitrarily chosen Germany–Spain country pair, ranked by divergence (Equation 5.2). The overuse of *hinder* by German authors may be attributed to its German *behindern* cognate, whereas Spanish users' preference of *impede* is probably attributable to its Spanish *impedir* equivalent. A Spanish cognate for *plantation*, *plantación*, possibly explains the clear preference of Spanish native speakers for this alternative, compared to the more popular choice of German authors, *grove*, which has Germanic etymological origins.

The {*weariness, tiredness, fatigue*} synset reveals the preference of Spanish native speakers for *fatigue*, whose Spanish equivalent *fatiga* resembles it to a great extent; *weariness*, however, is slightly more frequent in the texts of German speakers, potentially reflecting its Proto-Germanic **wōrīgaz* ancestor. An interesting phenomenon is revealed by the synset {*conceivable, imaginable*}: while both words have Latin origins, *imaginable* is more ubiquitous in the English language, rendering it more frequent in texts of German native speakers, compared to the more balanced choice of Spanish authors. Usage patterns in {*overdo, exaggerate*} and {*inspect, audit, scrutinize*} can be attributed to the same

Synonym set s	$P^s_{Germany}$	P^s_{Spain}
⟨hinder impede⟩	(0.909, 0.091)	(0.69, 0.31)
⟨grove orchard plantation⟩	(0.643, 0.214, 0.143)	(0.227, 0.068, 0.705)
⟨weariness tiredness fatigue⟩	(0.167, 0.208, 0.625)	(0.017, 0.119, 0.864)
⟨yarn recital narration⟩	(0.55, 0.1, 0.35)	(0.22, 0.15, 0.63)
⟨bloom blossom flower⟩	(0.25, 0.143, 0.607)	(0.085, 0.098, 0.817)
⟨conceivable imaginable⟩	(0.22, 0.78)	(0.415, 0.585)
⟨overdo exaggerate⟩	(0.556, 0.444)	(0.319, 0.681)
⟨inspect audit scrutinize ⟩	(0.667, 0.25, 0.083)	(0.446, 0.429, 0.125)
⟨sharp acute⟩	(0.886, 0.114)	(0.717, 0.283)
⟨steady stiff unwavering firm⟩	(0.364, 0.172, 0.017, 0.447)	(0.278, 0.083, 0.007, 0.632)
⟨ecstasy rapture⟩	(0.593, 0.407)	(0.412, 0.588)
⟨sizeable ample⟩	(0.597, 0.403)	(0.429, 0.571)
⟨scummy abject miserable⟩	(0.167, 0.028, 0.806)	(0.067, 0.053, 0.88)
⟨drift displace⟩	(0.835, 0.165)	(0.734, 0.266)
⟨waive abandon forego⟩	(0.095, 0.845, 0.061)	(0.043, 0.899, 0.058)
⟨weigh consider count⟩	(0.028, 0.605, 0.367)	(0.024, 0.582, 0.394)
⟨quick fast rapid⟩	(0.328, 0.649, 0.024)	(0.326, 0.643, 0.031)
⟨stumble stagger lurch⟩	(0.889, 0.097, 0.014)	(0.7, 0.114, 0.186)
⟨omen presage⟩	(1.0, 0.0)	(0.9, 0.1)
⟨freight cargo⟩	(0.215, 0.785)	(0.19, 0.81)

TABLE 5.4: Top-20 examples of the most divergent usage patterns of synsets in texts of German vs. Spanish authors. Words with (recorded) Germanic origins are in blue and words with (recorded) Latin origins are in red.

phenomenon, where the German equivalent for *inspect* (*inspizieren*) resembles its English counterpart despite a different etymological root.

L1	Sentence
French	*I have to go to the Dr. to do a **rapid** check on my heart stability.*
French	*Maybe put every name through a manual **approbation** pipeline so it ensures quality.*
French	*Polls have shown public **approbation** for this law is somewhere between 58% and 65%, and it has been a strong promise during the presidential campaign.*
Italian	*The event was even more shocking because the **precedent** evening he wasn't sick at all.*

TABLE 5.5: Cognate facilitation phenomena in usage examples by Reddit authors.

Usage examples Table 5.5 presents example sentences written by Reddit authors with French and Italian L1s, further illustrating discrepancies in lexical choice (presumably) stemming from cognate facilitation effects. The French *rapide* is a translation equivalent of the English synset {*rapid, quick, fast*}, but its English *rapid* cognate is more constrained to contexts of movement or growth, rendering the collocation *rapid check* somewhat marked. The French noun *approbation* is more frequent in contemporary French than its English (practically unused) equivalent *approbation*; this makes its use in English sound unnatural. In our Reddit corpus, *approbation* appears 48 times in L1-French texts, compared to 5, 4, and 4 in equal-sized texts by authors from the UK, Ireland and

Canada, respectively. One of the frequent English synonym alternatives {*approval, acceptance*} would better fit this context. Finally, while the Italian expression *sera precedente* is common, its English equivalent *precedent evening* is very infrequent, yet it is used in English productions of Italian speakers.

5.4 Conclusions and Future Work

We presented an investigation of L1 cognate effects on the productions of advanced non-native Reddit authors. The results are accompanied by a large dataset of native and non-native English speakers, annotated for author country (and, presumably, also L1) at the sentence level.

Several open questions remain for future research. From a theoretical perspective, we would like to extend this work by studying whether the tendency to choose an English cognate is more powerful in L1s with both phonetic and orthographic similarity to English (Roman script) than in L1s with phonetic similarity only (e.g., Cyrillic script). We also plan to more carefully investigate productions of speakers from multilingual countries, like Belgium and Switzerland. Another extension of this work may broaden the analysis to include additional language families.

There are also various potential practical applications to this work. First, we plan to exploit the potential benefits of our findings to the task of native language identification of (highly advanced) non-native authors, in various domains. Second, our results will be instrumental for personalization of language learning applications, based on the L1 background of the learner. For example, error correction systems can be enhanced with the native language of the author to offer root cause analysis of subtle discrepancies in the usage of lexical items, considering both their frequencies and context. Given the L1 of the target audience, lexical simplification systems can also benefit from cognate cues, e.g., by providing an informed choice of potentially challenging candidates for substitution with a simplified alternative. We leave such applications for future research.

Relevant publications:

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018.

Chapter 6

On the Similarities Between Native, Non-native and Translated Texts

This chapter addresses two linguistic phenomena: translation and non-native language. Our main goal is to investigate the similarities and differences between these two phenomena, and contrast them with native language. In particular, we are interested in the reasons for the differences between translations and originals, on one hand, and native and non-native language, on the other. Do they reflect "universal" principles, or are they dependent on the source/native language?

There are clear similarities between translations and non-native language: both are affected by the simultaneous presence of (at least) two linguistic systems, which may result in a higher cognitive load (Shlesinger, 2003). The presence of the L1 may also cause similar CLI effects on the target language.

On the other hand, there are reasons to believe that translationese and non-native language should differ from each other. Translations are produced by *native* speakers of the target language. Non-natives, in contrast, arguably never attain native-like abilities (Coppieters, 1987; Johnson and Newport, 1991), however this hypothesis is strongly debated in the SLA community (Birdsong, 1992; Lardiere, 2006).

Our goal in this work is to investigate three language *varieties*: the language of native speakers (N), the language of advanced, highly fluent non-native speakers (NN), and translationese (T). We use the term *constrained language* to refer to the latter two varieties. We propose a unified computational umbrella for exploring two related areas of research on bilingualism: translation studies and second language acquisition. Specifically, we put forward three main hypotheses: (1) The three language varieties have unique characteristics that make them easily distinguishable. (2) Non-native language and translations are closer to each other than either of them is to native language. (3) Some of these characteristics are dependent on the specific L1, but many are not, and may reflect unified principles that similarly affect translations and non-native language.

We test these hypotheses using several corpus-based computational methods. We

use supervised and unsupervised classification (Section 6.2) to show that the three language varieties are easily distinguishable. In particular, we show that native and advanced non-native productions can be accurately separated. More pertinently, we demonstrate that non-native utterances and translations comprise two distinct linguistic systems.

The main contribution of this chapter is thus theoretical: it sheds light on some fundamental questions regarding bilingualism, and we expect it to motivate and drive future research in both SLA and translation studies. Moreover, a better understanding of constrained language may also have some practical importance, as we briefly mention in the following section.

6.1 Methodology and experimental setup

6.1.1 Dataset

Our dataset[1] is based on the highly homogeneous corpus of the European Parliament Proceedings (Koehn, 2005). Note that the proceedings are produced as follows: (1) the utterances of the speakers are transcribed; (2) the transcriptions are sent to the speaker who may suggest minimal editing without changing the content; (3) the edited version is then translated by native speakers. Note in particular that the texts are *not* a product of simultaneous interpretation.

In this work we utilize a subset of Europarl in which each sentence is manually annotated with speaker information, including the EU state represented and the original language in which the sentence was uttered (Nisioi et al., 2016). The texts in the corpus are uniform in terms of style, respecting the European Parliament's formal standards. Translations are produced by native English speakers and all non-native utterances are selected from members not representing UK or Ireland. Europarl N consists of texts delivered by native speakers from England.

Table 6.1 depicts statistics of the dataset, and Table 6.2 provides details on the distribution of NN and T texts by various L1s. In contrast to other learner corpora such as ICLE (Granger, 2003), EFCAMDAT (Geertzen et al., 2013) or TOEFL-11 (Blanchard et al., 2013), this corpus contains translations, native, and non-native English of high proficiency speakers. Members of the European Parliament have the right to use any of the EU's 24 official languages when speaking in Parliament, and the fact that some of them prefer to use English suggests a high degree of confidence in their language skills.

6.1.2 Preprocessing

All datasets were split by sentence, cleaned (text lowercased, punctuation and empty lines removed) and tokenized using the Stanford tools (Manning et al., 2014). For the classification experiments we randomly shuffled the sentences within each language

[1]The dataset is available at http://nlp.unibuc.ro/resources.html

sub-corpus	**sentences**	**tokens**	**types**
native (N)	60,182	1,589,215	28,004
non-native (NN)	29,734	783,742	18,419
translated (T)	738,597	22,309,296	71,144
total	828,513	24,682,253	117,567

TABLE 6.1: Europarl corpus statistics: native, non-native and translated texts.

country of origin	tokens(T)	tokens(NN)
Austria	-	2K
Belgium	-	67K
Bulgaria	25K	6K
Cyprus	-	35K
Czech Republic	21K	3K
Denmark	444K	14K
Estonia	32K	50K
Finland	500K	81K
France	3,486K	28K
Germany	3,768K	17K
Greece	944K	13K
Hungary	167K	38K
Italy	1,690K	15K
Latvia	38K	13K
Lithuania	177K	18K
Luxembourg	-	46K
Malta	28K	40K
Netherlands	1,746K	64K
Poland	522K	36K
Portugal	1,633K	54K
Romania	244K	29K
Slovakia	88K	6K
Slovenia	43K	1K
Spain	1,836K	54K
Sweden	951K	52K

TABLE 6.2: Distribution of L1s by country.

variety to prevent interference of other artifacts (e.g., authorship, topic) into the classification procedure. We divided the data into chunks of approximately 2,000 tokens, respecting sentence boundaries, 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.

In all the experiments we used (the maximal) equal amount of data from each category, thus we always randomly down-sampled the datasets in order to have a comparable number of examples in each class; specifically, 354 chunks were used for each language variety: N, NN and T.

6.1.3 Features

The first feature set we utilized for the classification tasks comprises *function words* (FW), probably the most popular choice ever since Mosteller and Wallace (1963) used it successfully for the Federalist Papers. Function words proved to be suitable features for multiple reasons: (1) they abstract away from contents and are therefore less biased by topic; (2) their frequency is so high that by and large they are assumed to be selected unconsciously by authors; (3) although not easily interpretable, they are assumed to reflect grammar, and therefore facilitate the study of how structures are carried over from one language to another. We used the list of approximately 400 function words provided in Koppel and Ordan (2011).

A more informative way to capture (admittedly shallow) syntax is to use *part-of-speech (POS) trigrams*. Triplets such as PP(personal pronoun) + VHZ (*have*, 3sg present) + VBN (*be*, past participle) reflect a complex tense form, represented distinctively across languages. In Europarl, for example, this triplet is highly frequent in translations from Finnish and Danish and much rarer in translations from Portuguese and Greek. In this work we used the top-3,000 most frequent POS trigrams in each corpus.

We also used *positional token frequency* (Grieve, 2007). The feature is defined as counts of words occupying the first, second, third, penultimate and last positions in a sentence. The motivation behind this feature is that sentences open and close differently across languages, and it should be expected that these opening and closing devices will be transferred from L1 if they do not violate the grammaticality of the target language. Positional tokens were previously used for translationese identification (Volansky et al., 2015) and for native language detection (Nisioi, 2015a).

Translations are assumed to exhibit *explicitation*: the tendency to render implicit utterances in the source text more explicit in the translation product. For example, causality, even though not always explicitly expressed in the source, is expressed in the target by the introduction of cohesive markers such as *because*, *due to*, etc. (Blum-Kulka, 1986). Similarly, Hinkel (2001) conducted a comparative analysis of *explicit cohesive devices* in academic texts by non-native English students, and found that cohesive markers are distributed differently in non-native English productions, compared to their native counterparts. To study this phenomenon, we used the set of over 100 cohesive markers introduced in Hinkel (2001).

6.2 The status of constrained language

To establish the unique nature of each language variety in our dataset, we perform multiple pairwise binary classifications between N, NN, and T, as well as three-way classifications. Table 6.3 reports the results; the figures reflect average ten-fold cross-validation accuracy (the best result in each column is boldfaced).

In line with previous works (see Section 2), classification of N–T, as well as N–NN, yields excellent results with most features and feature combinations. NN–T appears to be easily distinguishable as well; specifically, FW+POS-trigrams combination with/without positional tokens yields 99.57% accuracy. The word *maybe* is among the most discriminative feature for NN vs. T, being overused in NN, as opposed to *perhaps*, which exhibits a much higher frequency in T; this may indicate a certain degree of formality, typical of translated texts (Olohan, 2003). The words *or*, *which* and *too* are considerably more frequent in T, implying higher sentence complexity. This trait is also reflected by shorter NN sentences, compared to T: the average sentence length in Europarl is 26 tokens for NN vs. 30 for T. Certain decisiveness devices (*sure*, *very*) are underused in T, in accordance with Toury (1995)'s law of standardization (Vanderauwerea, 1985). The three-way classification yields excellent results as well; the highest accuracy is obtained using FW+positional tokens with/without POS-trigrams.

feature / dataset	N-NN	N-T	NN-T	3-way
FW	98.72	98.72	96.89	96.60
POS (trigrams)	97.45	98.02	97.45	95.10
pos. tok	99.01	99.01	98.30	98.11
cohesive markers	85.59	87.14	82.06	74.19
FW+POS	99.43	99.57	**99.57**	99.34
FW+pos. tok	99.71	99.85	98.30	**99.52**
POS+pos. tok	99.57	99.57	99.01	99.15
FW+POS+pos. tok	**99.85**	**99.85**	**99.57**	**99.52**

TABLE 6.3: Pairwise and three-way classification results of N, NN and T texts.

A careful inspection of the results in Table 6.3 reveals that NN–T classification is a slightly yet systematically harder task than N–T or N–NN; this implies that NN and T texts are more similar to each other than either of them is to N.

To emphasize this last point, we analyze the separability of the three language varieties by applying unsupervised classification. We perform *bisecting KMeans* clustering procedure previously used for unsupervised identification of translationese by Rabinovich and Wintner (2015). Clustering of N, NN and T using function words into three clusters yields high accuracy, above 90%. For the sake of clusters' visualization in a bidimensional plane, we applied principal component analysis for dimensionality reduction.

The results are depicted in Figure 6.1 (a). Evidently, NN and T exhibit higher mutual proximity than either of them with N. Fixing the number of expected clusters to 2 further highlights this observation, as demonstrated in Figure 6.1 (b): both NN and T instances were assigned to a single cluster, distinctively separable from the N cluster.

We conclude that the three language varieties (N, NN, and T) constitute three different, distinguishable ontological categories, characterized by various lexical, syntactic

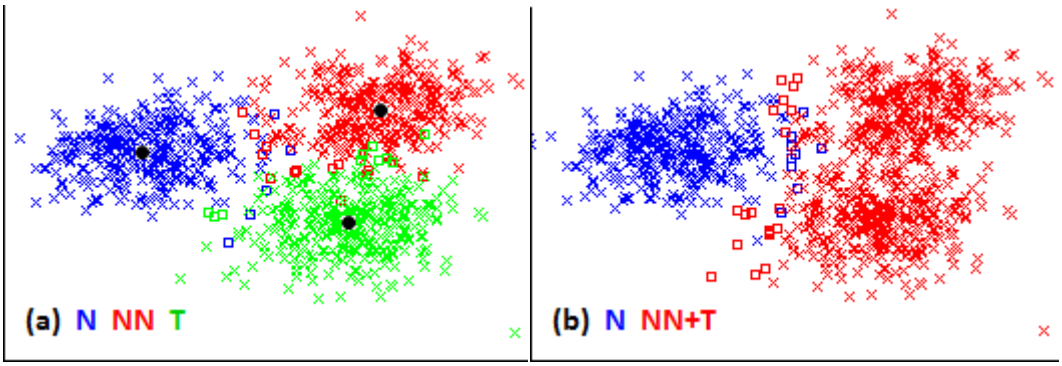

FIGURE 6.1: Clustering of N, NN and T into three (a) and two (b) clusters using function words. Clusters' centroids in (a) are marked by black circles; square sign stands for instances clustered wrongly.

and grammatical properties; in particular, the two varieties of constrained language (NN and T) represent two distinct linguistic systems. Nevertheless, we anticipate NN and T to share more common tendencies and regularities, when compared to N. In the following sections, we put this hypothesis to the test.

6.3 L1-independent similarities

In this section we address L1-independent similarities between NN and T, distinguishing them from N. We focus on characteristics which are theoretically motivated by translation studies and which are considered to be L1-independent, i.e., unrelated to cross-linguistic influences. We hypothesize that linguistic devices over- or under-represented in translation would behave similarly in highly competent non-native productions, compared to native texts.

To test this hypothesis, we realized various linguistic phenomena as properties that can be easily computed from N, NN and T texts. We refer to the computed characteristics as *metrics*. Our hypothesis is that NN metric values will be similar to T, and that both will differ from N. We used equally-sized texts of 780K tokens for N, NN and T; the exact computation is specified for each metric.

For the sake of visualization, the three values of each metric (for N, NN and T) were zero-one scaled by total-sum normalization. Figure 6.2 graphically depicts the normalized metric values. We now describe and motivate each metric. We analyze the results in Section 6.3.1 and establish their statistical significance in Section 6.3.2.

Lexical richness Translated texts tend to exhibit less lexical diversity (Al-Shabab, 1996). Blum-Kulka (1986) suggested that translated texts *make do with less words*, which is reflected by their lower type-to-token ratio (TTR) compared to that of native productions. We computed the TTR metric by dividing the number of unique (lemmatized) tokens by the total number of tokens.

Mean word rank Halverson (2003) claims that translators use more prototypical language, i.e., *they regress to the mean* (Shlesinger, 1989). We, therefore, hypothesize that rarer words are used more often in native texts than in non-native productions and translationese. To compute this metric we used a BNC-based ranked list of 50K English words[2], excluding the list of function words (see Section 6.1.3). The metric value was calculated by averaging the rank of all tokens in a text; tokens that do not appear in the list of 50K were excluded.

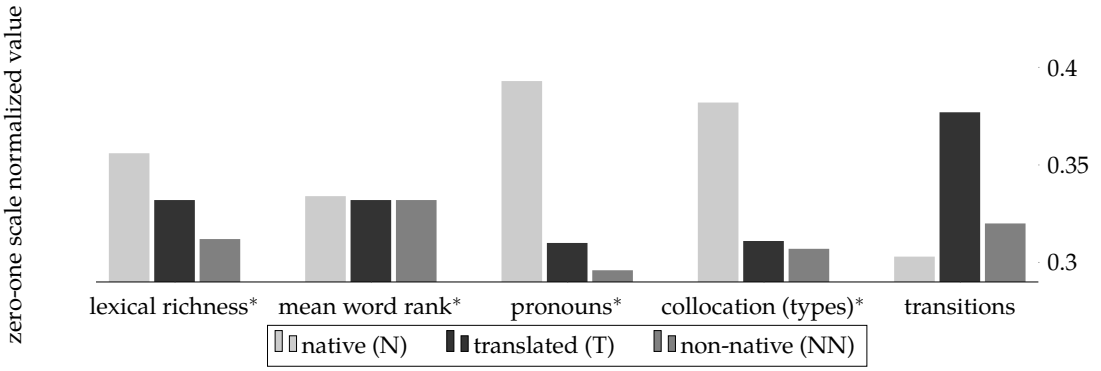

FIGURE 6.2: Metric values in N, NN and T. Tree-way differences are significant in all metric categories and "*" indicates metrics with higher pairwise similarity of NN and T, compared individually to N.

Collocations Collocations are distributed differently in translations and in originals (Toury, 1980; Kenny, 2001). Common and frequent collocations are used almost subconsciously by native speakers, but will be subjected to a more careful choice by translators and, presumably, by fluent non-native speakers (Erman et al., 2014). For example, the phrase *make sure* appears twice more often in native Europarl texts than in NN, and five times more than in T; *bear in mind* has almost double frequency in N, compared to NN and T. Expressions such as: *bring forward*, *figure out*, *in light of*, *food chain* and *red tape* appear dozens of times in N, as opposed to zero occurrences in NN and T Europarl texts. This metric is defined by computing the frequency of idiomatic expressions[3] in terms of types.

Cohesive markers Translations were proven to employ cohesion intensively (Blum-Kulka, 1986; Øverås, 1998; Koppel and Ordan, 2011). Non-native texts tend to use cohesive markers differently as well: *sentence transitions*, the major cohesion category, was shown to be overused by non-native speakers regardless of their native language (Hinkel, 2001). The metric is defined as the frequency of sentence transitions in the three language varieties.

[2] https://www.kilgarriff.co.uk we used the list extracted from both spoken and written text.

[3] Idioms were taken from https://en.wiktionary.org/wiki/Category:English_idioms. The list was minimally cleaned up.

Qualitative comparison of various markers between NN and T productions, compared to N in the Europarl texts, highlights this phenomenon: *in addition* is twice as frequent in NN and T than in N; *according*, *at the same time* and *thus* occur three times more frequently in NN and T, compared to N; *moreover* is used four times more frequently; and *to conclude* is almost six times more frequent.

Personal pronouns We expect both non-native speakers and translators to spell out entities (both nouns and proper nouns) more frequently, as a means of *explicitation* (Olohan, 2002), thus leading to under-use of personal pronouns, in contrast to native texts. As an example, *his* and *she* are twice more frequent in N than in NN and T.

We define this metric as the frequency of (all) personal and possessive pronouns used in the three language varieties. The over-use of personal pronouns in N utterances, is indeed balanced out by lower frequency of proper and regular nouns in these texts, compared to T and NN.[4]

6.3.1 Analysis

Evidently (see Figure 6.2), translationese and non-native productions exhibit a consistent pattern in both datasets, compared to native texts: NN and T systematically demonstrate lower metric values than N for all characteristics (except sentence transitions, where both NN and T expectedly share a higher value). All metrics except mean word rank exhibit substantial (sometimes dramatic) differences between N, on the one hand, and NN and T, on the other, thus corroborating our hypothesis. Mean word rank exhibits a more moderate variability in the three language varieties, yielding near identical value in NN and T; yet, it shows excessive usage in N.

The differences between metric values are statistically significant for all metrics (Section 6.3.2). Moreover, in all cases (except transitions), the difference between NN and T metrics is significantly lower than the difference between either of them and N, implying a higher proximity of NN and T distributions, compared individually to N. This finding further emphasizes the common tendencies between NN and T.

As shown in Figure 6.2, NN and T are systematically and significantly different from N. Additionally, we can see that T is consistently positioned between N and NN (except for sentence transitions), implying that translations produced by native speakers tend to resemble native utterances to a higher degree than non-native productions.

6.3.2 Statistical significance

Inspired by the results depicted in Figure 6.2, we now put to test two statistical hypotheses: (1) N, NN and T productions do not represent identical underlying distributions,

[4]Normalized frequencies of nouns and proper nouns are 0.323, 0.331 and 0.345 for N, T, and NN, respectively.

i.e., at least one pair is distributed differently; and consequently, (2) NN and T productions exhibit higher similarity (in terms of *distance*) than either of them with N. We test these hypotheses by applying the *bootstrapping* statistical analysis.

Bootstrapping is a statistical technique involving random re-sampling (with replacement) from the original sample; it is often used to assign a measure of accuracy (e.g., a confidence interval) to an estimate. Specifically, let C_{N}, C_{NN} and C_{T} denote native, non-native and translated sub-corpora of equal size (780K tokens). Let C_{ALL} denote the concatenation of all three sub-corpora, resulting in a total of 2,340M tokens. We further denote a function computing a metric m by f^m; when applied to C, its value is $f^m(C)$. The sum of pairwise distances between the three individual dataset metrics is denoted by D_{total}:

$$D_{\mathrm{total}} = |f^m(C_{\mathrm{N}}) - f^m(C_{\mathrm{NN}})| + f^m(C_{\mathrm{N}}) - f^m(C_{\mathrm{T}})| + |f^m(C_{\mathrm{NN}}) - f^m(C_{\mathrm{T}})|$$

High values of D_{total} indicate a difference between the three language varieties. To examine whether the observed D_{total} is high beyond chance level, we use the bootstrap approach, and repeat the following process 1,000 times:[5] we sample C_{ALL} with replacement (at sentence granularity), generating in the j-th iteration equal-sized samples $\widehat{C}^j_{\mathrm{N}}$, $\widehat{C}^j_{\mathrm{NN}}$, $\widehat{C}^j_{\mathrm{T}}$. The corresponding distance estimate, therefore, is:

$$\widehat{D}^j_{\mathrm{total}} = |f^m(\widehat{C}^j_{\mathrm{N}}) - f^m(\widehat{C}^j_{\mathrm{NN}})| + f^m(\widehat{C}^j_{\mathrm{N}}) - f^m(\widehat{C}^j_{\mathrm{T}})| + |f^m(\widehat{C}^j_{\mathrm{NN}}) - f^m(\widehat{C}^j_{\mathrm{T}})|$$

We repeat random re-sampling and computation of $\widehat{D}^j_{\mathrm{total}}$ 1,000 times, and estimate the p-value of $\widehat{D}_{\mathrm{total}}$ by calculation of its percentile within the series of (sorted) $\widehat{D}^j_{\mathrm{total}}$ values, where $j \in (1, \ldots, 1000)$. In all our experiments the original distance D_{total} exceeds the maximum estimate in the series of $\widehat{D}^j_{\mathrm{total}}$, implying highly significant difference, with p-value<0.001 for all metrics.

In order to stress this outcome even further, we now test whether (the constrained) NN and T exhibit higher pairwise similarity, as opposed to N. We achieve this by assessment of the distance between NN and T productions, compared to the distance between N and its closest production (again, in terms of distance): either NN or T. We sample C_{N}, C_{NN} and C_{T} (with replacement) separately, constructing $\widetilde{C}_{\mathrm{N}}$, $\widetilde{C}_{\mathrm{NN}}$ and $\widetilde{C}_{\mathrm{T}}$, respectively, and define the following distance function:

$$\widetilde{D}^j_{\mathrm{dif}} = |f^m(\widetilde{C}^j_{\mathrm{N}}) - |f^m(\widetilde{C}^j_{K})| - |f^m(\widetilde{C}^j_{\mathrm{NN}}) - |f^m(\widetilde{C}^j_{\mathrm{T}})|$$

[5] This sample size is proven sufficient by the highly significant results (very low p-value).

where

$$K = \begin{cases} \text{NN} & \text{if } |f^m(C_{\text{N}}) - f^m(C_{\text{NN}})| < |f^m(C_{\text{N}}) - f^m(C_{\text{T}})| \\ \text{T} & \text{otherwise} \end{cases}$$

We repeat re-sampling and computation of $\widetilde{D}^j_{\text{dif}}$ 1,000 times for each metric value in both datasets and sort the results. The end points of the 95% confidence interval are defined by estimate values with 2.5% deviation from the minimum (*min-end-point*) and the maximum (*max-end-point*) estimates. We assess the *p*-value of the test by inspecting the estimate underlying the min-end-point; specifically, in case the min-end-point is greater than 0, we consider p<0.05. Metric categories exhibiting higher NN-T similarity than either N-NN or N-T are marked with "*" in Figure 6.2.

6.4 L1-related similarities

We hypothesize that both varieties of constrained language exhibit similar (lexical, grammatical, and structural) patterns due to the influence of L1 over the target language. Consequently, we anticipate that non-native productions of speakers of a certain native language (L1) will be closer to translations from L1 than to translations from other languages.

Limited by the amount of text available for each individual language, we set out to test this hypothesis by inspection of two language *families*, Germanic and Romance. Specifically, the Germanic family consists of NN texts delivered by speakers from Austria, Germany, Netherlands and Sweden; and the Romance family includes NN speakers from Portugal, Italy, Spain, France and Romania. The respective T families comprise translations from Germanic and Romance originals, corresponding to the same countries. Table 6.4 provides details on the datasets.

	sentences	**tokens**	**types**
Germanic NN	5,384	132,880	7,841
Germanic T	269,222	7,145,930	43,931
Romance NN	6,384	180,416	9,838
Romance T	307,296	9,846,215	49,925

Table 6.4: Europarl Germanic and Romance families: NN and T.

We estimate L1-related traces in the two varieties of constrained language by the fitness of a translationese-based *language model* (LM) to utterances of non-native speakers from the same language family. Attempting to trace structural and grammatical, rather than content similarities, we compile five-gram *POS* language models from Germanic and Romance translationese (GerT and RomT, respectively).[6] We examine the prediction power of these models on non-native productions of speakers with Germanic and Romance native languages (GerNN and RomNN), hypothesizing that an LM compiled

[6]For building LMs we used the closed vocabulary of Penn Treebank POS tag set.

from Germanic translationese will better predict non-native productions of a Germanic speaker and vice versa. The fitness of a language model to a set of sentences is estimated in terms of *perplexity* (Jelinek et al., 1977).

For building and estimating language models we used the KenLM toolkit (Heafield, 2011), employing modified Kneser-Ney smoothing without pruning. Compilation of language-family-specific models was done using 7M tokens of Germanic and Romance translationese each; the test data consisted of 5350 sentences of Germanic and Romance non-native productions. Consequently, for perplexity experiments with individual languages we utilized 500 sentences from each language. We excluded OOVs from all perplexity computations.

Table 6.5 reports the results. Prediction of GerNN by the GerT language model yields a slightly lower perplexity (i.e., a better prediction) than prediction by RomT. Similarly, RomNN is much better predicted by RomT than by GerT. These differences are statistically significant: we divided the NN texts into 50 chunks of 100 sentences each, and computed perplexity values by the two LMs for each chunk. Significance was then computed by a two-tailed paired t-test, yielding p-values of 0.015 for GerNN and 6e-22 for RomNN.

LM / NN	GerNN	LM / NN	RomNN
GerT	8.77	GerT	8.64
RomT	8.79	RomT	8.43

Table 6.5: Perplexity: fitness of Germanic and Romance translationese LMs to Germanic and Romance NN test sets.

As a further corroboration of the above result, we computed the perplexity of the GerT and RomT language models with respect to the language of NN speakers, this time distinguishing speakers by their country of origin. We used the same language models and non-native test chunks of 500 sentences each. Inspired by the outcome of the previous experiment, we expect that NN productions by Germanic speakers will be better predicted by GerT LM, and vice versa. Figure 6.3 presents a scatter plot with the results.

A clear pattern, evident from the plot, reveals that all English texts with underlying Romance native languages (under the diagonal) are better predicted (i.e., obtain lower perplexity) by the RomT LM. All Germanic native languages (except German), on the other hand, are better predicted by the GerT LM. This finding further supports the hypothesis that non-native productions and translationese tend to exhibit similar L1-related traits.

6.5 Conclusions and Future Work

We presented a unified computational approach for studying constrained language, where many of the features were theoretically motivated. We demonstrated that while

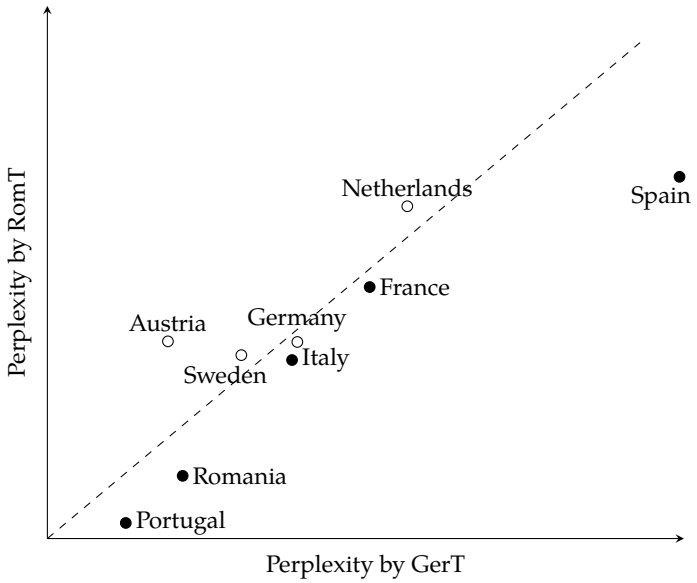

FIGURE 6.3: Perplexity of the GerT and RomT language models with respect to non-native utterances of speakers from various countries.

translations and non-native productions are two distinct language varieties, they share similarities that stem from lower lexical richness, more careful choice of idiomatic expressions and pronouns, and (presumably) subconscious excessive usage of explicitation cohesive devices. More dramatically, the language modeling experiments reveal salient ties between the native language of non-native speakers and the source language of translationese, highlighting the unified L1-related traces of L1 in both scenarios. Our findings are intriguing: native speakers and translators, in contrast to non-native speakers, use their native language, yet translation seems to gravitate towards non-native language use.

The main contribution of this work is empirical, establishing the connection between these types of language production. While we believe that these common tendencies are not incidental, more research is needed in order to establish a theoretical explanation for the empirical findings, presumably (at least partially) on the basis of the cognitive load resulting from the simultaneous presence of two linguistic systems. We are interested in expanding the preliminary results of this work: we intend to replicate the experiments with more languages and more domains, investigate additional varieties of constrained language and employ more complex lexical, syntactic and discourse features. We also plan to investigate how the results vary when limited to specific L1s.

Relevant publications:

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1870–1881, 2016a.

Chapter 7

Summary

Much research in translation studies indicates that translated texts have unique characteristics, that have been traditionally classified into two categories: properties that stem from *interference* of the source language, and universal traits resulting from the translation process itself, independently of the specific source and target languages. Similarly, over half a century of research on second language acquisition established the presence of *cross-linguistic influences* in non-native utterances, stemming from the interference of the mother tongue on productions in a foreign language. In addition, universal traits resulting from the learning process itself have been noticed regardless of the native language of a speaker.

This thesis investigated the two manifestations of crosslingual language varieties, namely translations and non-native language, and in particular explored similarities and differences between them, and the extent to which such similarities are "universal" or language-pair-specific.

Embarking on unsupervised classification of translated texts in a multi-variate setup (Chapter 3) we analyzed the manifestation of the signal of translation status compared to register-related properties (e.g., domain, modality, era, etc.). We concluded that the subtle, albeit systematic, signal of translationese is completely overshadowed by the other properties, raising the need for more sophisticated approaches to this task, e.g., techniques driven by various types of *transfer learning*. We deepened the analysis of the unique characteristics of translated text, focusing on source-language interference as the main factor shaping the unique dialect of translated texts (Chapter 4). The long quest for translation universals may now be viewed in light of our findings: more than anything else, translations are typified by interference.

We further studied the effect of cognate facilitation in productions of advanced non-native speakers using a very large and diverse corpus of non-native utterances collected from social-media platforms (Chapter 5). Our analysis lays out sound empirical foundations for the theoretical hypothesis on the cognate effect in L2 of English non-native speakers, highlighting the *cognate facilitation* phenomenon as one of the important factors shaping the language of non-native speakers.

Finally, Chapter 6 proposed a unified computational umbrella for exploring two related areas of research on bilingualism: translation studies and second language acquisition. We investigated the similarities and the differences between two related language varieties: the language of (highly advanced, fluent) non-native speakers, and translationese. This work highlights the power of transfer, or interference, as a major force in shaping these two crosslingual language varieties.

Bibliography

Jubin Abutalebi and David Green. Bilingual language production: The neurocognition of language representation and control. *Journal of neurolinguistics*, 20(3):242–275, 2007.

Omar S. Al-Shabab. *Interpretation and the language of translation: creativity and conventions in translation*. Janus, Edinburgh, 1996.

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

Mona Baker. 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*, pp. 233–252. John Benjamins, Amsterdam, 1993.

David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 828–834, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Marco Baroni and Silvia Bernardini. 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, 2006.

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 327–337. Association for Computational Linguistics, 2012.

Yevgeni Berzak, Roi Reichart, and Boris Katz. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 21–29, 2014.

Yevgeni Berzak, Roi Reichart, and Boris Katz. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pp. 94–102, 2015.

David Birdsong. Ultimate attainment in second language acquisition. *Language*, 68(4):706–755, 1992.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15, 2013.

Shoshana Blum-Kulka. 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*, volume 35, pp. 17–35. Gunter Narr Verlag, 1986.

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

Alix Boc, Anna Maria Di Sciullo, and Vladimir Makarenkov. 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*, pp. 647–655, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

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

Martin Chodorow, Michael Gamon, and Joel Tetreault. The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3):419–436, 2010.

Uschi Cop, Nicolas Dirix, Eva Van Assche, Denis Drieghe, and Wouter Duyck. Reading a book in one or two languages? an eye movement study of cognate facilitation in L1 and L2 reading. *Bilingualism: Language and Cognition*, 20(4):747–769, 2017.

Rene Coppieters. Competence differences between native and near-native speakers. *Language*, 63(3):544–573, 1987.

Scott A. Crossley and Danielle S. McNamara. Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, 20(4):271–285, 2011.

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

Annette M. de Groot. Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5):1001, 1992.

Gerard de Melo. Etymological Wordnet: Tracing the history of words. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Paris, France, 2014. ELRA.

Tamar Degani, Anat Prior, and Walaa Hajajra. Cross-language semantic influences in different script bilinguals. *Bilingualism: Language and Cognition*, pp. 1–23, 2017.

Tamar Degani and Natasha Tokowicz. Semantic ambiguity within and across languages: An integrative review. *The Quarterly Journal of Experimental Psychology*, 63(7):1266–1303, 2010.

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

Sascha Diwersy, Stefan Evert, and Stella Neumann. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*, pp. 174–204. De Gruyter, Berlin, Boston, 2014.

Jon Andoni Duñabeitia, Manuel Perea, and Manuel Carreiras. Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology*, 57(2):98–107, 2010.

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

Sauleh Eetemadi and Kristina Toutanova. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 159–164. Association for Computational Linguistics, 2014.

T. Mark Ellison and Simon Kirby. 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*, pp. 273–280, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Britt Erman, Annika Denke, Lars Fant, and Fanny Forsberg Lundell. Nativelike expression in the speech of long-residency L2 users: A study of multiword structures in L2 English, French and Spanish. *International Journal of Applied Linguistics*, 2014.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pp. 263–272, 2007.

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

Wendy S. Francis. Bilingual semantic and conceptual representation. In *Handbook of bilingualism: Psycholinguistic approaches*, chapter 12, pp. 251–267. Oxford University Press, New York, 2005.

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

Robert M. French and Maud Jacquet. Understanding bilingual memory: models and data. *Trends in Cognitive Sciences*, 8(2):87–93, 2004.

Federico Gaspari and Silvia Bernardini. Comparing non-native and translated language: Monolingual comparable corpora with a twist. In *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*, 2008.

Susan M. Gass and Larry Selinker. *Second Language Acquisition: An Introductory Course*. Digital Online. Taylor & Francis, 2008.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*, Somerville, MA, 2013. Cascadilla Proceedings Project.

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

Gili Goldin, Ella Rabinovich, and Shuly Wintner. Native language identification with user generated content, Under review.

Sylviane Granger. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, pp. 538–546, 2003.

Sylviane Granger. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1):7–24, 2015.

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

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

Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

François Grosjean and Ping Li. *The Psycholinguistics of Bilingualism*. Wiley-Blackwell, 2013.

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

Sandra Halverson. The cognitive basis of translation universals. *Target*, 15(2):197–241, 2003.

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

ZhaoHong Han. Forty years later: Updating the fossilization hypothesis. *Language teaching*, 46 (2):133–171, 2013.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197. Association for Computational Linguistics, 2011.

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

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

Eli Hinkel. Second language writers' text: Linguistic and rhetorical features. 2002.

Hagen Hirschmann, Anke Lüdeling, Ines Rehbein, Marc Reznicek, and Amir Zeldes. 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.*, pp. 223–234. Presses Universitaires de Louvain, Louvain la Neuve, 2013.

Kristian Tangsgaard Hvelplund. Eye tracking and the translation process: reflections on the analysis and interpretation of eye-tracking data. *MonTI. Monografías de Traducción e Interpretación*, pp. 201–223, 2014.

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

Iustina Ilisei and Diana Inkpen. Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1-2), 2011.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 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*, volume 6008 of *Lecture Notes in Computer Science*, pp. 503–511. Springer, 2010.

Zahurul Islam and Alexander Mehler. Customization of the Europarl corpus for translation studies. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), 2012.

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

Scott Jarvis and Aneta Pavlenko. *Crosslinguistic influence in language and cognition*. Routledge, 2008.

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62:S63, 1977. Supplement 1.

Jacqueline S. Johnson and Elissa L. Newport. Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language. *Cognition*, 39(3): 215–258, 1991.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2nd edition, 2002.

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

Dorothy Kenny. *Lexis and creativity in translation: a corpus-based study*. St. Jerome, 2001.

Ekaterina Kochmar and Ekaterina Shutova. Modelling semantic acquisition in second language learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 293–302, 2017.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pp. 79–86. AAMT, 2005.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for Europe. In *Proceedings of the Twelfth Machine Translation Summit*, pp. 65–72, 2009.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1318–1326, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 624–628. ACM, 2005.

Judith F. Kroll, Susan C. Bobb, and Noriko Hoshino. Two languages in mind: Bilingualism as a tool to investigate language, cognition, and the brain. *Current Directions in Psychological Science*, 23(3):159–163, 2014.

Judith F. Kroll, Paola E. Dussias, Cari A. Bogulski, and Jorge R. Valdes Kroff. Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. *Psychology of Learning and Motivation*, 56:229–262, 2012.

Judith F. Kroll and Natasha Tokowicz. Models of bilingual representation and processing. In JUDITH F. KROLL and ANNETTE M. B. DE GROOT, editors, *Handbook of bilingualism: Psycholinguistic approaches*, chapter 26, pp. 531–553. Oxford University Press, New York, 2005.

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

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990, 2012.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pp. 81–88, 2009.

Kristopher Kyle and Scott A. Crossley. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786, 2015.

Donna Lardiere. *Ultimate Attainment in Second Language Acquisition: A Case Study*. L. Erlbaum, 2006.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825, 2012.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023, 2013.

Maya R. Libben and Debra A. Titone. Bilingual lexical access in context: evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35 (2):381, 2009.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Xiaofei Lu and Haiyang Ai. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29, 2015. New developments in the study of L2 writing complexity.

Gerard Lynch and Carl Vogel. Towards the automatic detection of the source language of a literary translation. In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics: Posters*, pp. 775–784, 2012.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, Baltimore, Maryland, 2014. Association for Computational Linguistics.

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

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403, 2008.

Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

Ryo Nagata and Edward W. D. Whittaker. Reconstructing an Indo-European family tree from non-native English texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1137–1147, 2013.

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

Luay Nakhleh, Tandy Warnow, Don Ringe, and Steven N. Evans. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society*, 103(2):171–192, 2005b.

Vivi Nastase and Carlo Strapparava. Word etymology as native language interference. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2702–2707. Association for Computational Linguistics, 2017.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pp. 849–856. MIT Press, 2001.

Sergiu Nisioi. Feature analysis for native language identification. In Alexander F. Gelbukh, editor, *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2015)*, Lecture Notes in Computer Science. Springer, 2015a.

Sergiu Nisioi. Unsupervised classification of translated texts. In Chris Biemann, Siegfried Handschuh, André Freitas, Farid Meziane, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems: Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems, NLDB*, volume 9103 of *Lecture Notes in Computer Science*, pp. 323–334. Springer, 2015b.

Sergiu Nisioi, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

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

Maeve Olohan. Leave it out! using a comparable corpus to investigate aspects of explicitation in translation. *Cadernos de Tradução*, 1(9):153–169, 2002.

Maeve Olohan. How frequent are the contractions? A study of contracted forms in the translational English corpus. *Target*, 15(1):59–89, 2003.

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

Magali Paquot and Sylviane Granger. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32:130–149, 2012.

Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

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

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

Marius Popescu. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *Proceedings of RANLP-2011*, pp. 634–639, 2011.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62, 2011.

Anat Prior. Bilingualism: Interactions between languages. In Patricia J. Brook and Vera Kempe, editors, *Encyclopedia of Language Development*. Sage Publications, 2014.

Anat Prior, Brian MacWhinney, and Judith F Kroll. Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39(4):1029–1038, 2007.

Anat Prior, Shuly Wintner, Brian Macwhinney, and Alon Lavie. Translation ambiguity in and out of context. *Applied Psycholinguistics*, 32(1):93–111, 2011.

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

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

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1870–1881, 2016a.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL-2017)*, pp. 530–540. Association for Computational Linguistics, 2017a.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017)*, pp. 1074–1084. Association for Computational Linguistics, 2017b.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018.

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

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. A parallel corpus of translationese. In *Proccedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, 2016b.

Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3): 349–380, 2003.

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

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

Don Ringe, Tandy Warnow, and Ann Taylor. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129, 2002.

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

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

Larry Selinker. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10 (1–4):209–232, 1972.

Larry Selinker and William E. Rutherford. *Rediscovering interlanguage*. Routledge, 2013.

Maurizio Serva and Filippo Petroni. Indo-European languages tree by Levenshtein distance. *Europhysics Letters*, 81(6):68005, 2008.

Miriam Shlesinger. Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum. Master's thesis, Tel Aviv University, Faculty of the Humanities, Department of Poetics and Comparative Literature, 1989.

Miriam Shlesinger. Effects of presentation rate on working memory in simultaneous interpreting. *The Interpreters' Newsletter*, 12:37–49, 2003.

Anna Siyanova-Chanturia. Collocation in beginner learner writing: A longitudinal study. *System*, 53:148–160, 2015.

Bernard Smith and Michael Swan. *Learner English: A teacher's guide to interference and other problems*. Ernst Klett Sprachen, 2001.

Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *KDD-2000 Workshop on Text Mining*, 2000.

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

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

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

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 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, 2013.

Laura Mayfield Tomokiyo and Rosie Jones. You're not from 'round here, are you?: naive Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 1–8. Association for Computational Linguistics, 2001.

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

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

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 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*, pp. 279–287. Association for Computational Linguistics, 2013.

Hans van Halteren. 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*, pp. 937–944, 2008.

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

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

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

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

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