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

Ella Rabinovich[△⋆] **Sergiu Nisioi**[◇] **Noam Ordan**[†] **Shuly Wintner**[⋆]
[△]IBM Haifa Research Labs
[⋆]Department of Computer Science, University of Haifa
[◇]Solomon Marcus Center for Computational Linguistics, University of Bucharest
[†]The Arab College for Education, Haifa
{ellarabi,sergiu.nisioi,noam.ordan}@gmail.com, shuly@cs.haifa.ac.il

Abstract

We present a computational analysis of three language varieties: native, advanced non-native, and translation. Our goal is to investigate the similarities and differences between non-native language productions and translations, contrasting both with native language. Using a collection of computational methods we establish three main results: (1) the three types of texts are easily distinguishable; (2) non-native language and translations are closer to each other than each of them is to native language; and (3) some of these characteristics depend on the source or native language, while others do not, reflecting, perhaps, unified principles that similarly affect translations and non-native language.

1 Introduction

This paper 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?

Much research in translation studies indicates that translated texts have unique characteristics. Translated texts (in any language) constitute a sublanguage of the target language, sometimes referred to as *translationese* (Gellerstam, 1986). The unique characteristics of translationese have been traditionally classified into two categories: properties that stem from *interference* of the source language (Toury, 1979), and universal traits resulting from the translation process itself, independently of the specific source and target languages (Baker, 1993; Toury, 1995). The latter so-called *translation universals* have triggered a continuous debate among translation studies researchers (Mauranen and Kujamäki, 2004; House, 2008; Becher, 2010).

Similarly, over half a century of research on second language acquisition (SLA) established the presence of *cross-linguistic influences* (CLI) in non-native utterances (Jarvis and Pavlenko, 2008). CLI is a cover term proposed by Kellerman and Sharwood-Smith (1986) to denote various phenomena that stem from language contact situations such as transfer, interference, avoidance, borrowing, etc.[1] In addition, universal traits resulting from the learning process itself have been noticed regardless of the native language, L1.[2] For example, similar developmental sequences have been observed for negation, question formation, and other sentence structures in English (Dulay and Burt, 1974; Odlin, 1989) for both Chinese and Spanish natives. Phenomena such as overgeneralization, strategies of learning (Selinker, 1972), psychological factors (Ellis, 1985), and cultural distance (Giles and Byrne, 1982) are also influential in the acquisition process.

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

[1]To avoid terminological conflicts, we shall henceforth use CLI to denote any influence of one linguistic system over another, w.r.t. both translations and non-native productions.

[2]For simplicity, we will use *L1* to refer both to the native language of a speaker and to the source language of a translated text. We use *target language* to refer to second and translation languages (English in this paper).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1870–1881,
Berlin, Germany, August 7-12, 2016.

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 4) 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.

In Section 5, we use statistical analysis to explore the unique properties of each language variety. We show that the two varieties of constrained language are much closer to each other than they are to native language: they exhibit poorer lexical richness, a tendency to use more frequent words, a different distribution of idiomatic expressions and pronouns, and excessive use of cohesive devices. This is an unexpected finding, given that both natives and translators (in contrast to non-natives) produce texts in their mother tongue.

Finally, in Section 6 we use language modeling to show that translations and non-native language exhibit similar statistical properties that clearly reflect cross-linguistic influences: experiments with distinct language families reveal salient ties between the two varieties of constrained language.

The main contribution of this work 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 import, as we briefly mention in the following section.

2 Related work

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; Rabinovich and Wintner, 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 prove 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).

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 associ-

ated with L1 influences on the target language.

To the best of our knowledge, our work is the first to address both translations and non-native language under a unifying computational framework, and in particular to compare both with native language.

3 Methodology and experimental setup

3.1 Dataset

Our dataset[3] 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 1 depicts statistics of the dataset.[4] 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.

3.2 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 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.

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

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.

3.3 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

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

[4] Appendix A provides details on the distribution of NN and T texts by various L1s.

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).

4 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 2 reports the results; the figures reflect average tenfold 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 (Vanderauwera, 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 2: Pairwise and three-way classification results of N, NN and T texts.

A careful inspection of the results in Table 2 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.

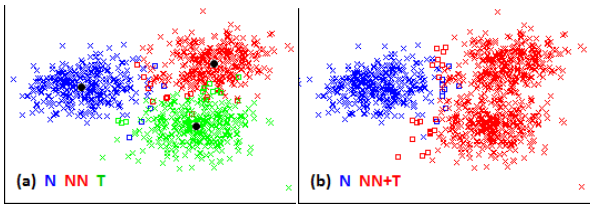

Figure 1: Clustering of N, NN and T into three (a) and two (b) clusters using function words. Clusters' centroids in (a) are marked by black circles; square sign stands for instances clustered wrongly.

The results are depicted in Figure 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 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 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.

5 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 2 graphically depicts the normalized metric values. We now describe and motivate each metric. We analyze the results in Section 5.1 and establish their statistical significance in Section 5.2.

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[5], excluding the list of function words (see Section 3.3). 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.

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[6] 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.

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 fre-

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

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

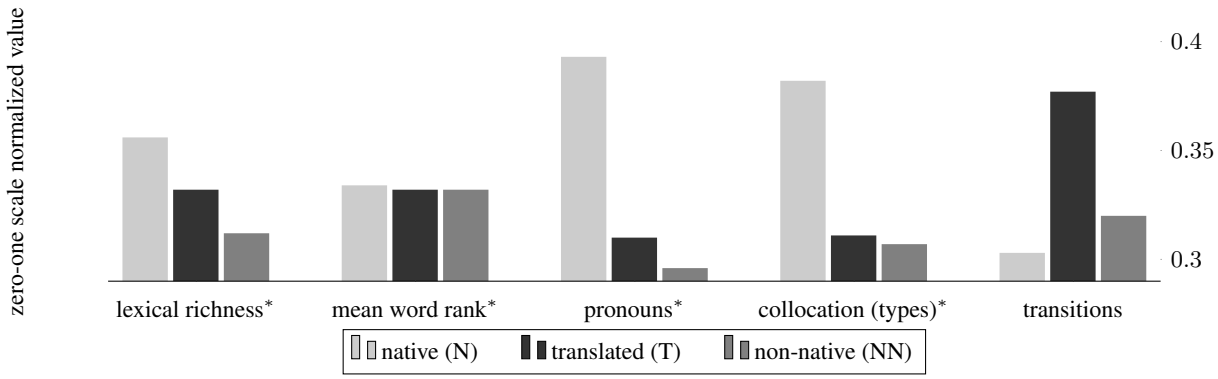

Figure 2: Metric values in N, NN and T. Tree-way differences are significant in all metric categories and "*" indicates metrics with higher pairwise similarity of NN and T, compared individually to N.

quently; 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.[7]

5.1 Analysis

Evidently (see Figure 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 5.2). 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 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.

5.2 Statistical significance

Inspired by the results depicted in Figure 2, we now put to test two statistical hypotheses: (1) N, NN and T productions do not represent identical underlying distributions, 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 pair-

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

wise distances between the three individual dataset metrics is denoted by D_{total}:

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

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:[8] we sample C_{ALL} with replacement (at sentence granularity), generating in the j-th iteration equal-sized samples $\widehat{C_{\text{N}}}^j$, $\widehat{C_{\text{NN}}}^j$, $\widehat{C_{\text{T}}}^j$. The corresponding distance estimate, therefore, is:

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

We repeat random re-sampling and computation of $\widehat{D_{\text{total}}}^j$ 1,000 times, and estimate the p-value of $\widehat{D_{\text{total}}}$ by calculation of its percentile within the series of (sorted) $\widehat{D_{\text{total}}}^j$ values, where $j \in (1, \ldots, 1000)$. In all our experiments the original distance D_{total} exceeds the maximum estimate in the series of $\widehat{D_{\text{total}}}^j$, 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 $\widetilde{C_{\text{N}}}$, $\widetilde{C_{\text{NN}}}$ and $\widetilde{C_{\text{T}}}$, respectively, and define the following distance function:

$$\widetilde{D_{\text{dif}}}^j = |f^m(\widetilde{C_{\text{N}}}^j) - f^m(\widetilde{C}_K^j)| - |f^m(\widetilde{C_{\text{NN}}}^j) - f^m(\widetilde{C_{\text{T}}}^j)|$$

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 $\widetilde{D_{\text{dif}}}^j$ 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 2.

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

6 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 3 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 3: 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).[9] We examine the predic-

[9] For building LMs we used the closed vocabulary of Penn

tion power of these models on non-native productions of speakers with Germanic and Romance native languages (GerNN and RomNN), hypothesizing that an LM compiled 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 4 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 4: 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 3 presents a scatter plot with the results.

A clear pattern, evident from the plot, reveals

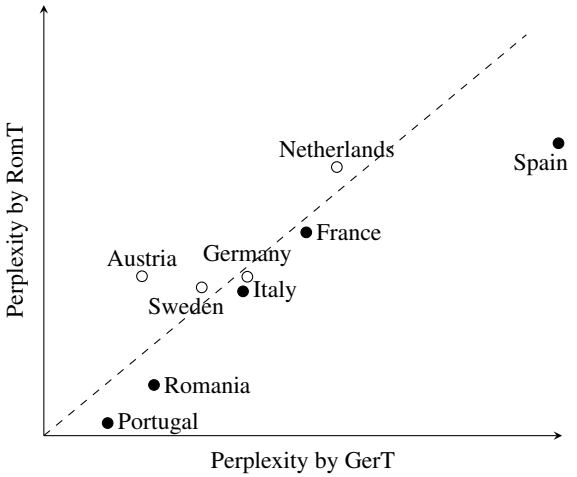

Figure 3: Perplexity of the GerT and RomT language models with respect to non-native utterances of speakers from various countries.

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.

7 Conclusion

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

Treebank POS tag set.

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.

Acknowledgments

This research was supported by the Israeli Ministry of Science and Technology. We are immensely grateful to Yuval Nov for much advice and helpful suggestions. We are also indebted to Roy Bar-Haim, Anca Bucur, Liviu P. Dinu, and Yosi Mass for their support and guidance during the early stages of this work, and we thank our anonymous reviewers for their valuable insights.

References

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

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

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

Viktor Becher. 2010. Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1):1–28.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. 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*. Association for Computational Linguistics, pages 327–337.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2015. 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*. pages 94–102.

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

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

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

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

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

Heidi C. Dulay and Marina K. Burt. 1974. Natural sequences in child second language acquisition. *Language learning* 24(1):37–53.

Rod Ellis. 1985. *Understanding Second Language Acquisition*. Oxford Applied Linguistics. Oxford University Press.

Britt Erman, Annika Denke, Lars Fant, and Fanny Forsberg Lundell. 2014. 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* .

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

Federico Gaspari and Silvia Bernardini. 2008. 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.

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

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

Howard Giles and Jane L. Byrne. 1982. An intergroup approach to second language acquisition. *Journal of Multilingual and Multicultural Development* 3(1):17–40.

Sylviane Granger. 2003. The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly* pages 538–546.

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

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

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

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

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

Juliane House. 2008. Beyond intervention: Universals in translation? *trans-kom* 1(1):6–19.

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

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

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

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

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

Eric Kellerman and Michael Sharwood-Smith. 1986. *Crosslinguistic Influence in Second Language Acquisition*. Language Teaching Methodology Series. Pearson College Division.

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

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

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

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. 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*. ACM, pages 624–628.

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

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

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

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

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

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

Frederick Mosteller and David L Wallace. 1963. 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.

Sergiu Nisioi. 2015a. 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)*. Springer, Lecture Notes in Computer Science.

Sergiu Nisioi. 2015b. 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*. Springer, volume 9103 of *Lecture Notes in Computer Science*, pages 323–334.

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

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge Applied Linguistics. Cambridge University Press.

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

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

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

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

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

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

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

Miriam Shlesinger. 1989. *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.

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

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

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

Laura Mayfield Tomokiyo and Rosie Jones. 2001. 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 on Language technologies*. Association for Computational Linguistics, pages 1–8.

Gideon Toury. 1979. Interlanguage and its manifestations in translation. *Meta* 24(2):223–231.

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

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

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

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

Ria Vanderauwera. 1985. *Dutch Novels Translated into English: The transformation of a "minority" literature*. Rodopi.

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

Appendix A - Distribution of L1s in Translations and Non-native Texts

We assume that native languages of non-native speakers are highly correlated with (although not strictly identical to) their country of origin.

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 5: Distribution of L1s by country.