Natural Language Engineering (2021), **27**, pp. 339–372
doi:10.1017/S1351324920000182

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Investigating translated Chinese and its variants using machine learning

Hai Hu* and Sandra Kübler

Department of Linguistics, Indiana University, Bloomington, IN, USA
*Corresponding author. E-mail: huhai@indiana.edu

(Received 5 November 2018; revised 18 July 2019; accepted 27 August 2019; first published online 3 April 2020)

Abstract

Translations are generally assumed to share universal features that distinguish them from texts that are originally written in the same language. Thus, we can argue that these translations constitute their own variety of a language, often called translationese. However, translations are also influenced by their source languages and thus show different characteristics depending on the source language. Consequently, we argue that these variants constitute different "dialects" of translations into the same target language. Studies using machine learning techniques on Indo-European languages have investigated the universal characteristics of translationese and how translations from various source languages differ. However, for typologically very different languages such as Chinese, there are only few corpus studies that tap into the intricate relation between translations and the originals, as well as into the relations among translations themselves. In this contribution, we investigate the following questions: (1) What are the characteristics of Chinese translationese, both in general and with respect to different source languages? (2) Can we find differences not only at the lexical but also on the syntactic level? and (3) Based on the characteristics found in the previous questions, which of the proposed laws and universals can we corroborate based on our evidence from Chinese? We use machine learning to operationalize determining the importance of different characteristics and comparing their importance for our Chinese dataset with characteristics previously reported in studies on English. In addition, our methodology allows us to add syntactic features, which have rarely been used to study translations into Chinese. Our results show that Chinese translations as a whole can be reliably distinguished from non-translations, even based on only five features. More interestingly, typological traces from the source languages can often be found in their translations, therefore creating what we call dialects of translationese. For instance, translations from two Altaic languages exhibit more noun repetition and less frequent use of pronouns. Additionally, some characteristics that are not discriminative for English work well for Chinese, possibly because the distance between Chinese and the source languages is greater than that in English studies.

Keywords: Translationese; Text classification; Chinese; Translation universal

1. Introduction

Texts translated into a target language possess linguistic properties that are very different from comparable texts originally written in this language. These properties have been termed the "third code" (Frawley 1984) or "translationese" (Gellerstam 1986) in the literature. In fact, translationese has been shown to originate from two possible sources: one is assumed to be the translation process, which is independent of the source and target languages and results in "translation universals", as argued by Baker (1993, 1996). These universals include *simplification* (translations tend to be simpler) and *explicitation* (translations are more explicit), among several others. The second

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

source of differences concerns the assumption that translationese is bound to be influenced by the source language, from which the translations are produced. This is often referred to as the "law of interference" (Toury 1995) or source language "shining through" (Teich 2003).

Our work presented here investigates the hypothesis that translations constitute a separate language variety, different from other types of text in that language. In addition, we investigate the hypothesis that there is also interference from the source language, in which case we can consider translations from different source languages to be *dialects* of the translation variety (c.f. Koppel and Ordan 2011). We use the term "dialect" rather loosely, to easily refer to the differences we see based on the source language of a translation. We are not trying to draw parallels to regional or social dialects.

The different characteristics of translationese and its dialects have been studied for decades (Toury 1978; Baker 1993; Laviosa-Braithwaite 1996; Mauranen and Kujamäki, 2004; Baroni and Bernardini 2005; Xiao 2010, among many others). Identifying them is not only of great interests to theoretical translation studies, but can also help improve the training of translators, the evaluation of translation quality, and the performance of language models in machine translation systems (Lembersky *et al*. 2012). It is also the prerequisite for proposing and verifying any laws or universals with regard to translations. However, most of the existing translationese studies focus on Indo-European target languages, such as English (Laviosa-Braithwaite 1996; Olohan and Baker 2000), Spanish (Ilisei *et al*. 2010), German (Rubino *et al*. 2016), or Russian (Kunilovskaya and Kutuzov 2017). However, if we assume that these characteristics are universal, then they need to be tested on a wide range of languages and also on multiple linguistic levels. In our current work, we focus on Chinese as the target language in translations. While there exist studies on Chinese, they only looked at translations from one source language, namely English (Xiao 2010; Xiao and Hu 2015; Hu *et al*. 2018). As a result, they cannot be generalized to Chinese translations in general. More specifically, since only one source language was available, *interference* from different source languages could not be investigated and thus needs to be studied more thoroughly.

Therefore, in the research reported here, we first introduce a new corpus of translated Chinese, balanced with seven source languages from different language families, and then investigate three research questions: (1) What are the characteristics of Chinese translationese, both in general and with respect to different source languages? (2) Can we find differences not only at the lexical but also on the syntactic level? and (3) Based on the characteristics found in the previous questions, which of the proposed laws and universals can we corroborate based on our evidence from Chinese?

We use machine learning, classifying text into either translation or original text, in order to determine which characteristics reach a high level of accuracy. From a high accuracy, we can conclude that the respective characteristics are important differences between originals and translations or between different translation dialects. Note that this methodology allows us to investigate scenarios where combinations of features are important. This is much more difficult in count-based corpus studies. We have shown in previous work (Hu *et al*. 2018) that a machine learning approach can reliably distinguish between translations and Chinese originals. In essence, the goal of this work is twofold: one goal is to determine empirically which features[a] can distinguish translated and non-translated Chinese as well as translations from different source languages. The other goal is more theoretical in nature, that is, we investigate whether Chinese translations can be described by the universal hypotheses proposed by Baker (1993) and whether the law of interference is applicable.

The remainder of this article is structured as follows: Section 2 reviews the relevant literature in translation studies, focusing on the characteristics that have been investigated and on machine learning approaches, on which our work is based. Then, the new corpus is introduced (Section 3).

[a] We use the terms "characteristics" and "features" as loose synonyms; "characteristics" is used when we talk about linguistic or theoretical work, "features" is used in machine learning work.

Next, we perform machine learning experiments to distinguish translations from originals, with the goal of identifying and interpreting the discriminative features (Section 4). We will start with a range of features based on a previous study in English (Volansky *et al.* 2013) and the syntactic features from our previous work (Hu *et al.* 2018). Then, new features will be introduced and tested. Finally, we compare our results with the study on English by Volansky *et al.* (2013) and discuss implications for the universal hypothesis and the law of interference. Our findings are summarized in Section 5.

2. Related work

In this section, we will review relevant literature from two perspectives. The first angle is more linguistic in nature and revolves around the characteristics of translated text proposed over the decades, such as simplification and explicitation. Note that each category of features can be realized on different linguistic levels ranging from the most obvious and more easily accessible lexical level to the less obvious syntactic level. The second perspective is more computational in nature and concerns the methods and techniques for analyzing or utilizing these features. We will review studies that used frequency comparisons, focusing on the more recent machine learning research, which is the basis for our experimental work.

2.1 Features of translations

Over the years, research in translation studies has used the term “translationese” to describe the distinctive linguistic characteristics of translations that set them apart from both the source text and comparable text in the same target language (Gellerstam 1986). Detailing what exactly those features are has been a major goal in corpus-based translation studies. Several high-level linguistic properties of translated text have been proposed, some of which are termed “translation universals” that are supposed to hold across all translations, regardless of source or target language (Baker 1993, 1995, 1996). For example, “explicitation” was first discussed by Blum-Kulka (1986), referring to the tendency that target text is usually rendered more explicit than the *source* text, as a result of the translation process. Later, explicitation was extended to describe the relation between translations and *comparable* original text (Baker 1993, 1995). For example, in translations, pronouns are more likely to be spelled out, and connectives are often added to enhance cohesion and aid understanding. “Simplification” refers to the observation that translations are in general lexically and syntactically simpler and less ambiguous than original text (Baker 1993, 1996). The “normalization” universal describes the tendency to conform to the “typical patterns” of the target language (Baker 1996). Finally, there is “levelling out,” a “tendency of translated text to gravitate toward the center of a [linguistic] continuum” (Baker 1996). An example is that in translated English, the type-token ratio (TTR) has a smaller standard deviation than original texts (Laviosa-Braithwaite 1996).

There are several other characteristics of translationese in the literature. One of them is “implicitation,” that is, the opposite of explicitation. Implicitation refers to the tendency of translators to make a text more implicit, by either leaving out connectives or using pronouns instead of the nouns in the source text. Implicitation may depend on the source–target language pair in question. For example, Cartoni *et al.* (2011) show that French-to-English translations use more clausal connectives than original English, but the opposite is true for English-to-French translations. There is evidence for implicitation in other languages as well; see Becher (2011) for German, Meyer and Webber (2013) for French and German, and Ke (2005) for Chinese. However, this characteristic is difficult to operationalize. For this reason, we refrain from using it in our work.

All of the above-mentioned characteristics tend to focus on the discrepancies between translations and original text in the target language, therefore neglecting the influence from

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

the source language. However, translations are by definition a product of the source text. Thus, it goes without saying that the source language plays an indispensable role in the creation of translationese. This is often referred to as the "law of interference" (Toury 1995), which is also described as the phenomenon of the source language "shining through" (Teich 2003). This effect has been demonstrated in empirical studies, for example, by Volansky *et al.* (2013) and by Evert and Neumann (2017).

To verify whether these proposed characteristics of translationese hold, be they universals or interference, researchers have traditionally used methods from corpus linguistics. For example, to study simplification, we can compare the measures of textual complexity such as TTR, mean sentence length, lexical density to see if there are statistically significant differences between a comparable corpus of translations and originals (Laviosa-Braithwaite 1996; Xiao 2010; Hu *et al.* 2016). The same methodology also applies for explicitation. Scholars usually start with connectives and other function words, such as conjunctions and the complementizer *that*, which are considered signs of explicitness (Olohan and Baker 2000; Puurtinen 2004; Chen 2006). This line of work has produced fruitful yet mixed results. For example, studies on Hungarian (Pápai 2004) and Chinese (Chen 2006; Xiao 2010) found evidence supporting explicitation in translations via a comparison of frequencies, whereas a study on Finnish by Puurtinen (2004) reported "no clear overall tendency of either translated or original Finnish using connectives more frequently" in a corpus of translated and original children's literature.

It should be noted that in a comparable corpus setting, where one has only access to (a) translations from a single source language, and (b) text originally written in the target language, there is no easy way of investigating the interference from the source language. To properly study source language interference, we usually either need both translated and source text to examine the translation process (as in the study by Blum-Kulka (1986)) or have translations from multiple source languages for comparison (as in the study by Volansky *et al.* (2013)).

Starting from work by Baroni and Bernardini (2005), however, a new paradigm in translation studies has gradually gained popularity. Instead of using occurrence counts of lexico-grammatical properties, the importance of linguistic characteristics is determined using them as features in machine learning tasks that classify text into translations and originals. Our study follows this line of work, and we will discuss relevant studies in the next section.

2.2 Classification between translated and original text

Baroni and Bernardini (2005) use machine learning with features such as word or part-of-speech (POS) *n*-grams to classify journal articles on geopolitics that are either originally written in Italian or translated from English. They demonstrate that an ensemble of support vector machines can reach very high accuracy (86.7%), demonstrating convincingly that translationese is an experimentally verifiable phenomenon. They show that a machine learner can recognize linguistic differences between translations and originals with an accuracy that is higher than that of professional translators.

Similar classification tasks have been carried out for other Indo-European languages, for example, by Ilisei *et al.* (2010) for Spanish medical and technical texts, by Ilisei and Inkpen (2011) for Romanian news text, by Rubino *et al.* (2016) for a mixed genre of German, and by many for English (Koppel and Ordan 2011; Volansky *et al.* 2013; Rabinovich and Wintner 2015; Rabinovich *et al.* 2016).

While some of these studies aim to explore classification under a more complex setting, for example, using a multi-genre, multi-source language corpus (Koppel and Ordan 2011), employing unsupervised methods (Rabinovich and Wintner 2015), others focus on examining the hypotheses reviewed above. In these studies, machine learning experiments serve as empirical evaluation of linguistic hypotheses. For example, Ilisei *et al.* (2010) and Ilisei and Inkpen (2011) operationalize several simplification features, such as mean sentence length, parse tree depth, average senses per

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

word, and show that adding these features improves the classification accuracy, providing support for the simplification hypothesis. They also show that there are syntactic differences in addition to lexical ones.

Of particular interest to our work is the study by Volansky *et al.* (2013). They build classifiers to distinguish English original texts from translations into English aggregated from 10 source languages in the Europarl corpus (Koehn 2005). They categorize their features under four hypotheses: simplification, explicitation, normalization, and interference, with the goal of investigating the proposed translation universals and other characteristics. They not only cite the work of Baker (1993) as the basis for their work but also include interference in their feature design.

More specifically, Volansky *et al.* (2013) experiment with a number of simplification features. Different operationalizations of TTR achieve >70% accuracy. Other features, such as lexical density, only reach an accuracy of 53%. Sentence length yields an accuracy of 65%, but this does not support the simplification hypothesis since 7 out of the 10 translations (from Italian, Portuguese, Spanish, etc.) have longer sentences than the English originals, contrary to the predictions of the simplification hypothesis. Overall, translations are 2.5 words longer than the originals.

The interference features used by Volansky *et al.* (2013) include POS *n*-grams, prefixes, and suffixes, among others. POS *n*-grams achieve a 90–98% accuracy, prefixes, and suffixes 80%. Crucially, Volansky *et al.* (2013) analyze the POS trigrams and find that translations from all source languages have a much higher number of the trigram MD + VB + VVN (modal + verb base form + verb past participle, e.g., *must be given*) compared with English originals, with translations from Finnish at the top, followed by Swedish and Danish. The analysis of affix features again reveals that source language-specific prefixes are carried over to their English translations. For instance, *mono-* is more frequent in translations from Greek because of its Greek origin; the Latinate suffix *-ible* is more prominent in translations from Romance languages. These results not only prove the existence of interference but they also identify linguistic structures that distinguish translations from different source languages.

For Chinese, to our knowledge, our investigation (Hu *et al.* 2018) is the only classification study, and no study has utilized translations from multiple source languages. Therefore, the interference effect has only been investigated for English-to-Chinese translations. The current study aims to fill the two gaps and set the baseline for text classification of original and translated Chinese. More interestingly, it will provide a direct comparison between English (Volansky *et al.* 2013) and Chinese translationese to see whether those hypotheses hold for a language from a language family different from the Indo-European languages that have been intensely investigated so far.

On the methodological side, we will experiment with syntactic features that have rarely been used in Chinese translation studies and focus on providing a detailed and meaningful linguistic interpretation. Syntactic features have been used in several studies on Indo-European translations. As described above, Ilisei and Inkpen (2011) use parse tree depth as an indicator for sentence complexity in Spanish. In another study, Rubino *et al.* (2016) compute surprisal, complexity, and distortion features based on flattened constituent parse trees, which are then used as features for their classifier. While Rubino *et al.* (2016) operationalize information density as surprisal, we use another information-theoretic measure—entropy—to measure information density and complexity, since Hale (2016) suggests that entropy is a better predictor in modeling human sentence processing. There are also other studies that look at syntactic features without using parse trees. For instance, Kunilovskaya and Kutuzov (2017) use words, POS tags, and length features to detect unnaturalness (i.e., translationese) in English-to-Russian translations. Some studies in the edited volume by De Sutter *et al.* (2017) also investigate structural properties of translations, for example, phrasal verbs in English translations (Cappelle and Loock 2017), noun + modifier, and modifier + noun combinations in Italian translations and interpretations (Ferraresi and Miličević 2017). However, they all extract relevant structures by wordform matching, rather than by relying on parse trees. One innovation in our current study is to measure syntactic complexity by the entropy

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

of various types of grammar rules and to use them as features for classification, as we will explain in the next section.

3. Methodology

In this section, we describe the features, dataset, and classifier in the current study. Our starting point constitutes the work by Volansky *et al.* (2013) (and our previous work (Hu *et al.* 2018)) in that we also use a machine learning approach to text classification as the method to measure the validity of the translation hypotheses. Volansky *et al.* (2013) have based their work on the theoretical work by Baker (1993, 1995, 1996) and Toury (1995), that is, they investigate (mostly) Baker's translation universals and the law of interference. Following Volansky *et al.* (2013), we first explore features representing the three universals used in their work. However, if we see differences between source languages with regard to those features, we can interpret those differences as support for the law of interference.

3.1 Features

The features used in this study are explained below. A complete list can be found in Appendix B. Some of the features are taken from Ilisei *et al.* (2010) and Volansky *et al.* (2013), with modifications suitable for Chinese. We add syntactic features and features specific to Chinese.

3.1.1 Explicitation features

Variants of cohesive markers. The most widely used type of features for explicitation consists of cohesive markers. We test a list of 160 cohesive markers from Chen (2006). We next experiment with features that measure the richness of the markers, that is, their total counts, TTR, and entropy.

Explicit naming. This is operationalized as the ratio of personal pronouns to proper nouns, based on the assumption that the more explicit the text, the more proper nouns, rather than pronouns, are used.

Single naming and mean multiple naming. *Single naming* counts the number of single-token pronouns, whereas *mean multiple naming* counts the average number of tokens in proper noun phrases (NPs), both with the hypothesis that more explicit texts are likely to elaborate on the proper nouns and have longer ones in general (e.g., by adding in translations the titles or positions of officials). The three naming features are taken from Volansky *et al.* (2013).

3.1.2 Simplification features

TTR and lengths of linguistic units. Most of our lexical features are commonly used in stylometry studies (Grieve 2007): TTR, length of various linguistic units, etc. Here, we use the operationalization by Volansky *et al.* (2013) when possible so that our results are directly comparable to theirs for English. We borrow two variants of TTR and add a character-based TTR measure since in Chinese, word segmentation is nontrivial and characters almost always constitute morphemes themselves, which can also serve as the building block of the lexicon. Mean word and sentence lengths are straightforward, with the hypothesis that translations are shorter because they should be simpler.

Lexical density. This is measured by the number of content words divided by the number of all tokens. We expect translations to have a lower proportion of content words.

Mean character/word rank. The mean rank of all characters/words from ranked frequency lists. A lower mean rank means the text uses more frequent words rather than rare words since they have lower rank. Translations are expected to have a lower mean rank.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Most of the above features are directly comparable to Volansky *et al.* (2013). We also use several syntactic measures of complexity, inspired by previous research in translation (Ilisei *et al.* 2010), language acquisition (Chen *et al.* 2009; Lu 2010), and psycholinguistics (Lin 2011; Kwon *et al.* 2013):

Constituent tree depth. The mean depth of all constituent trees (Ilisei *et al.* 2010).

Number of complex NPs per clause. Complex NPs include NPs involving adjectives, quantifier phrases, classifier phrases, etc., as modifiers, as well as conjoined NPs[b]. This is inspired by work by Lu (2010) who measured the complexity of learners' English. Chinese originals are hypothesized to use more complex NPs.

Number of verb phrases (VPs) per clause. For VPs, it is difficult to operationalize the notion of complex VPs, thus we simply count all VPs, excluding the ones with copula 是 (Eng. *be*).

Number of relative clauses per clause. Previous studies have shown that translations in Chinese tend to use more complex relative clauses (Lin 2011; Lin and Hu 2018). Relative clauses in different languages are structurally distinctive, for example, pre-nominal in Chinese, Japanese, and Korean, post-nominal in many Indo-European languages (Kwon *et al.* 2013; Lin 2018). Hence, we expect translations and originals to use them differently.

Entropy of grammar. Apart from structure counts, we also investigate the entropy of grammar rules. Entropy measures the uncertainty of a random variable (Shannon 1948). Here, we use it as an approximation of how scattered, diverse, or complex a grammar is. It has been used to measure syntactic complexity of child language (Chen *et al.* 2009). For context-free grammar (CFG) rules, we calculate the entropy of the rules headed by each phrase (NP, VP, etc.)[c]:

$$E(\mathrm{XP}) = -\sum_{r\in\{\text{rules headed by XP}\}} p(r)\cdot\log\left(p(r)\right)$$

A high entropy for NP rules roughly means that there are diverse and complex ways of forming an NP. Note that there are 28 phrasal tags in the Chinese Penn Treebank (Xue *et al.* 2005), on which our parser model is trained. Thus, we have 28 entropy features. In a related setting, we use the number of unique *types* of each XP-headed rule as features, which also gives us 28 features.

Previous studies only investigated certain syntactic features such as the *ba* and *bei* structures (Xiao and Hu 2015), without looking at syntactic complexity in a more general and principled manner. We are filling this gap by measuring the complexity of the grammar as a whole.

3.1.3 Normalization features

Repetitions. This feature counts the number of content words (nouns, verbs, adjectives, and adverbs) that occur more than once in a text chunk (Volansky *et al.* 2013). We also use a variant that counts only the number of re-occurring nouns. This feature operationalizes the tendency for translations to avoid repetitions (Ben-Ari 1998).

Average pointwise mutual information (PMI). This is the mean PMI across all word bigrams (following Volansky *et al.* 2013). Note that the greater the PMI, the more likely it is that the bigram is a collocation. The *normalization* hypothesis states that translations tend to conform to conventions in the language, leading us to expect more collocations or fixed phrases. If the hypothesis is true, we would expect higher average PMI for translations.

We add several features that capture the unique linguistic properties in Chinese since translations tend to "exaggerate . . . typical grammatical structures, punctuation and collocational patterns or clichés" (Baker 1996).

[b] We use Tregex patterns (Levy and Andrew 2006) to extract the relevant structures. For a list of the patterns, see Appendix C.

[c] Similarly, we have used a dependency grammar. Since the results are similar, we do not report them.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Aspect markers. Chinese lacks tense markers but does have three unique *aspect markers* (了 *le*, 过 *guo* and 着 *zhe*) which mark the perfective, experiential, and imperfective aspects, respectively. We use the normalized counts of these markers as features.

Measure words/classifiers. A measure word is required for all nouns in Chinese, and it is placed between a quantifier or demonstrative and the noun. We exclude measure words that are related to metric systems, such as 千克 *kilogram*, 磅 *pound*, since they differ between countries and would thus be predicitive in the machine learning task, but not with regard to translationese.

Sentence-final particles. Sentence-final particles (啊 *a*, 吗 *ma*, 吧 *ba*, etc.) express the emotion of speakers or the mood of the sentence (e.g., 吗 *ma* indicates interrogative mood). We again use the normalized counts of these particles as features.

4-character idioms. A unique type of idioms called *chengyu* (成语) is commonly used in Chinese. They mostly consist of four characters, usually with some historical reference. Assuming that they are highly discriminative for original Chinese, we use the normalized counts of the idioms from Xinhua Dictionary[d].

***ba*-structure.** The *ba*-structure is used to front the object and make it the focus of the sentence.

***bei*-structure.** The *bei*-structure is the Chinese passive form. It has been shown that translated Chinese uses more *bei*-structure than the originals (Xiao and Hu 2015).

3.1.4 Interference features

***n*-grams.** Word and character *n*-grams are used as upper bound in previous studies; POS *n*-grams have been shown to be effective in machine learning approaches (Volansky *et al.* 2013; Hu *et al.* 2018).

We also use a set of syntactic features based on constituent parse trees. These features have been tested in our previous work (Hu *et al.* 2018) and in Native Language Identification tasks (Swanson and Charniak 2012; Bykh and Meurers 2014; Malmasi and Dras 2018). We assume that they capture translationese structures.

CFG rules. Counts of CFG rules.

3.2 Dataset

3.2.1 Sub-corpus of translated Chinese

Our translation corpus is based on the Chinese newspaper 《参考消息》 (Eng.: *Reference News*), published by the official news agency of the Chinese government, Xinhua News Agency. The newspaper was launched in the 1950s, initially meant to be an internal publication for government officials to read (translated) foreign news. Therefore, it only contains news reports and commentaries translated from non-Chinese sources, for example, *The New York Times*, *Der Spiegel*, *Le Figaro*. A large number of articles are translated from German, French, Russian, Japanese, and other languages.

In the current study, we use articles translated from seven source languages: German (DE), American English (EN), Spanish (ES), French (FR), Japanese (JP), Korean (KO), and Russian (RU). We take articles from about 10 sources per language, in a time span of 20 years (1980 to 1999), totaling roughly 400,000 tokens for each language (except for Korean, for which we only have 372,000 tokens). For German, English, French, Japanese, and Russian, we use articles longer than 400 tokens to ensure a good representation.

We have manually checked information of each non-English newspaper to make sure that they do not have an English version, or their English version is launched after 2000, so that the texts we use are indeed translated from the language in question[e].

[d] https://github.com/pwxcoo/chinese-xinhua

[e] A distribution of articles for each source can be found in Appendix A.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

We take the 400,000 tokens for each source language and split them into 200 chunks, each with roughly 2000 tokens, in order to match the chunk size in the corpora used by Volansky *et al.* (2013) and Xiao (2010). Article boundaries are discarded. That is, one chunk of text in English might include different articles, but all chunks are roughly equal in length.

3.2.2 Sub-corpus of original Chinese

To find a comparable corpus of original Chinese, we turn to the Chinese news published by the same news agency that publishes *Reference News*: Xinhua. The Chinese Gigaword project contains news articles from Xinhua originally written in Chinese, from 1991 to 2006 (Graff 2007). We select an equal amount of tokens from each year between 1991 and 1999 to build our corpus of original Chinese. Only news of the type "story" are included to ensure that we have standard narrative texts. Other types include short summaries of news or advertisement.

We preprocess these texts in the same way as the translated Chinese. That is, we obtain 2,800,000 tokens of Xinhua News in total, which is roughly equivalent to number of all translations combined. They are then split into chunks of 2000 words, resulting in 1400 chunks. Hence, we have roughly the same number of chunks and words per chunk for translations and originals.

We have decided on these two sub-corpora for several reasons. First, both are in the news domain, which not only makes them comparable but also easier to process using a current NLP pipeline (segmentation, POS tagging, parsing etc.), since most models are trained on news data. In fact, the data in the Chinese Penn Treebank (Xue *et al.* 2005), which is generally used to train the segmenters, POS taggers, and parsers, consist mostly of news texts from Xinhua. Second, as data from both sub-corpora are published by Xinhua, we assume that they have gone through similar editing processes. In this way, we exclude differences caused by editing policies as far as possible. Third, the translations in *Reference News* are of high quality. *Reference News* remains one of the most widely circulated newspapers in China today with a very good reputation. It has a rigid system of selecting translators, and the translations undergo thorough proofreading before publication[f].

3.3 Experimental setup

For the machine learning experiments, we use a binary task, classifying text chunks into translated or original Chinese. We use the support vector machine classifier (SMO) in WEKA (Hall *et al.* 2009) with its default settings and 10-fold cross-validation (following Volansky *et al.* (2013)). For feature extraction, we extend the implementation by Lutzky and Star[g]. The Stanford CoreNLP toolkit (Manning *et al.* 2014) and their default models are used for segmentation, POS tagging, and parsing of Chinese.

3.4 Visualization of results

Throughout the analysis section, we will make extensive use of the violin plot (e.g., Figure 1(a)). In such plots, the x-axis represents translations from the seven source languages (DE: German, EN: English, ES: Spanish, FR: French, JP: Japan, KO: Korean, and RU: Russian) and the original text in Chinese (XIN, short for Xinhua News Agency). The y-axis shows the (normalized) value of the feature in question. The dots in the plot indicate the mean of each value, and the shape of the violin visualizes the full distribution of the texts for that feature, where the thicker part shows that the values in that section have higher frequency, and the thinner part shows lower frequencies. This makes this representation more informative than box plots.

[f] Unfortunately, we do not have information on the translators for each individual text. However, the first author has interned at the newspaper and has witnessed the rigorous translation and proofreading procedure, performed by a team of expert Chinese translators.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

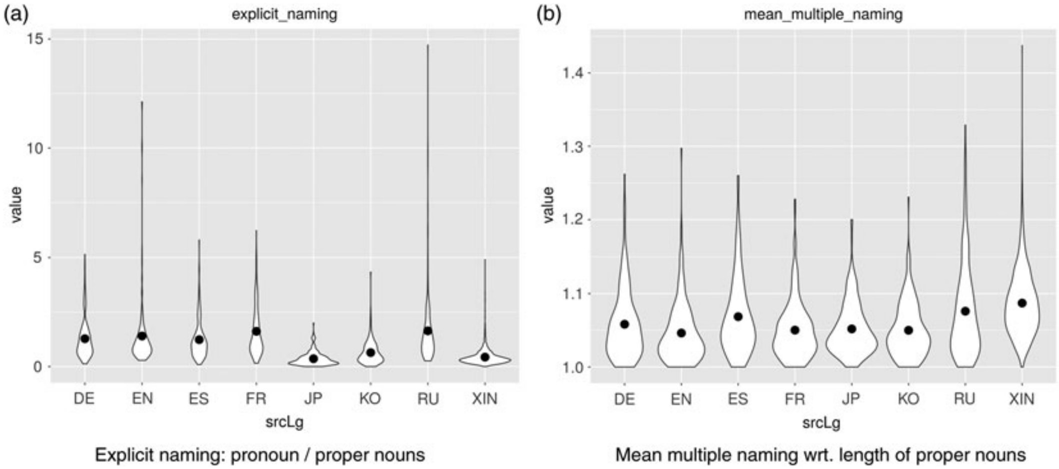


Figure 1. Explicitation features in translations and original Chinese (XIN). Black dots indicate the mean.

4. Results and discussion

We use machine learning to determine which types of linguistic information are highly distinctive for original Chinese or translations, that is, if we use certain features and reach a high accuracy, we can deduce that these features are distinctive. Our previous work (Hu *et al.* 2018) has shown that only using word *n*-grams, we reach a perfect accuracy in classifying texts into Chinese originals or translations. For many other features, our classifier is able to distinguish translations from originals with high accuracy, even if the features have worked poorly for English as target language (Volansky *et al.* 2013). In the curent work, we see that translations are distributed based on their typological traits for a number of features. This means that the Japanese and Korean translations often pattern together, and translations from Indo-European languages are closer in many respects.

The following analysis of our results is structured along the four translations universals that also served as the basic structuring principle in the work by Volansky *et al.* (2013). If we find features that allow a successful separation of translations and originals, we can assume that these features belong to the translation universals. If we find features that work for some languages but not others, we have candidates for the law of interference.

4.1 The explicitation hypothesis

The first hypothesis we look at is explicitation: translations are more explicit than the originals. We report the classification results in Table 1.

Naming features. Figure 1(a) shows the value of *explicit naming* for all seven translations, as well as the originals (XIN). As this feature is operationalized as the ratio of personal pronouns to proper nouns, the higher the value, the more personal pronouns are used compared to proper nouns. The 69.20% accuracy (from Table 1) indicates that this feature is discriminative to a certain extent. Crucially, this argues against the explicitation assumption, which would expect translations to be more explicit, that is, to use more proper nouns.

Figure 1(a) shows that translations from all Indo-European languages use more pronouns than translations from Japanese and Korean. Translations from the latter languages use fewer personal

[g] https://github.com/lutzky/translationese

Table 1. Classification accuracy for our Chinese experiments in comparison to English (Volansky *et al.* 2013): Explicitation features

Features	Chinese	English
Explicit naming	69.20	58
Single naming	57.61	56
Mean multiple naming	65.25	54
Cohesive markers: all	96.55	81
Cohesive markers: top 5	92.00	NA
Cohesive markers: counts	90.24	NA
Cohesive markers: TTR	90.63	NA
Cohesive markers: entropy	87.87	NA

pronouns: their means (indicated by the dot) are lower, and the mass of their violin plots is skewed toward 0. We assume that this behavior is based on similar characteristics of the source and target language rather than resulting from explicitation: Japanese and Korean are similar to Chinese in that their pronouns can be dropped—given enough context—in both subject and object positions. Indo-European languages, in contrast, generally do not allow dropped pronouns, thus it would require a higher cognitive load to decide when to correctly drop the pronouns than to simply keep them in the translations from Indo-European languages[h].

The *single naming* feature achieves an accuracy just above chance (57.61%). *Mean multiple naming* performs better, with 65.25%. Figure 1(b) shows that the original Chinese texts have the longest proper NPs in general. This again argues against explicitation, which assumes that the translations will add additional information about proper nouns, resulting in longer proper name phrases. This may be attributed to the writing style of Xinhua Chinese news, which tends to add titles or positions of officials in the news, for example, 中共中央总书记、国家主席、中央军委主席江泽民 (Eng.: Jiang Zemin, General Secretary of the Central Committee of the Chinese Communist Party, President of P. R. China, and the Chairman of the Central Military Committee). This is an extreme example, but similar cases are not uncommon in Xinhua news.

Cohesive features. These are usually the main features in translation studies on explicitation. If we use all 160 cohesive markers used by Chen (2006: Appendix), we reach an accuracy of 96.55%. If we only use the five most discriminative features, selected via Gain Ratio, we reach 92.00%. The high accuracy shows that cohesive markers are very discriminative. Figure 2 shows the top 10 markers, and we see that translations use them more frequently than originals. This is in line with previous studies on Chinese (e.g., Xiao 2010), which found an overuse of cohesive markers across all genres.

There are other interesting patterns with respect to individual translations. For example, translations from English use 但是 (Eng.: but) most frequently, whereas translations from Japanese overuse 据说 (Eng.: it is said that). Korean and Japanese translations overuse 所以 (Eng.: so/therefore) but seem to have fewer 因为 (Eng.: because) compared to other translations.

[h] Note that some Indo-European languages like Spanish also exhibit pro-drop, but it is limited to subject positions, with surface morphological cues in the conjugations of verbs. This is different from pro-drop in Chinese, Japanese, and Korean, where no surface lexical cues are available.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

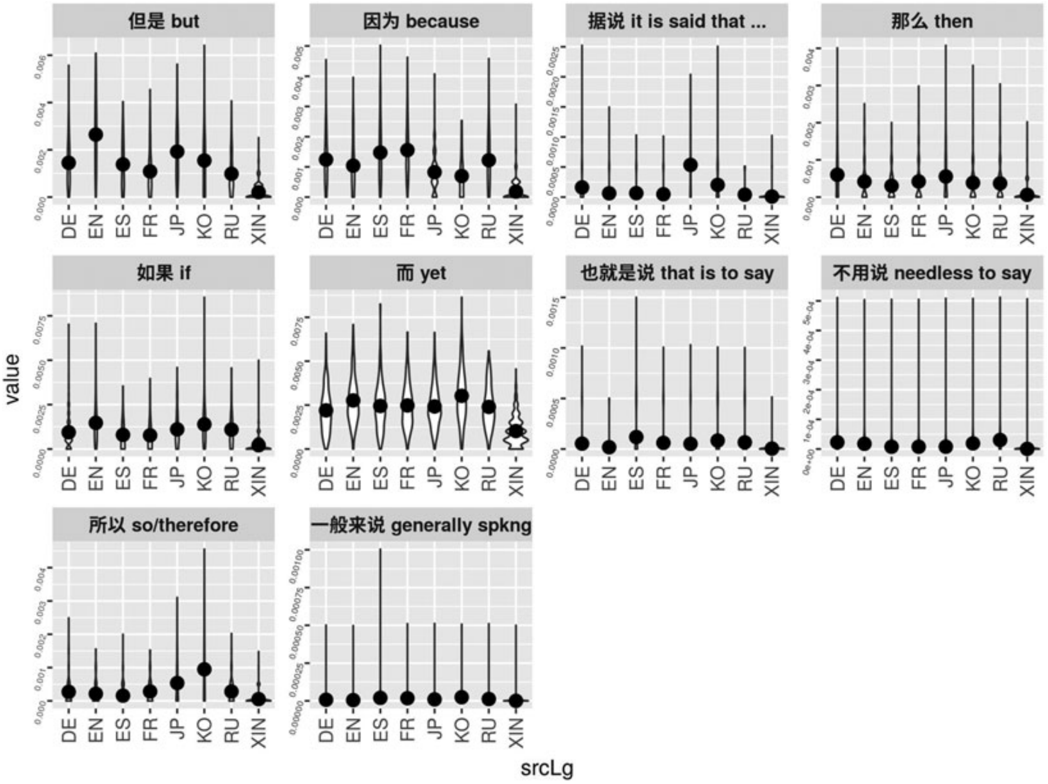


Figure 2. Normalized counts of the 10 most discriminating cohesive markers in translations and original Chinese (XIN).

We now have a closer look at adversative conjunctions: when we use all adversative conjunctions in Chinese (但, 但是, 然而, 可是, 不过) (Eng.: but, yet, however) as features, the accuracy is 86.86%. The analysis, however, shows no consistent pattern of English translations overusing all adversative conjunctions, but overall, translations do use more adversatives than the originals. As for the extremely frequent use of 据 说 (Eng.: it is said that) in Japanese translations, this may be caused by interference from the commonly used Japanese hearsay marker (そうです), which indicates that the speaker has heard the information from other sources and is not confident of its validity[i].

We also look at other operationalizations of cohesive markers: total counts, TTR, and the entropy of these markers. The high accuracies (see Table 1) suggest that they are all good indicators of translations. A closer look shows that translations use about 50 cohesive markers per 2000-token chunk, whereas the originals use only half, about 25 per chunk.

When we compare the results of cohesive markers between Chinese in our work and English (Volansky *et al*. 2013) in Table 1, we observe that using only five Chinese markers performs much better than a list of 40 in English (92% vs. 81%). One possible reason is the difference in experimental settings: while we use translations from seven source languages, the English study uses 10, possibly making it more difficult to find shared features among all 10 translations into Chinese. Additionally, all the languages in the English study are Indo-European languages, whereas in our experiment, Chinese is very different from any of the source languages, which makes it easier to identify the translations. We see the discrepancy in classification accuracy throughout our study.

[i] Based on a suggestion by a Japanese-to-Chinese translator (p.c. Zhu).

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Another major reason is that all source and target languages in the study by Volansky *et al.* (2013) belong to the Indo-European family; since they are more closely related, they will not differ as much in their ways of realizing cohesion. However, Chinese has a very different mechanism to express cohesion, which in most cases depends solely on the context with no surface cohesive markers at all. This has been noticed early on by Wang (1943, 1944) and is a major topic in all translation courses in Chinese. To give an example (from Zhu (1985)), 买不起别买- (Eng. literally: cannot afford don't buy; meaning: if you cannot afford it, don't buy it) is perfectly grammatical and natural in Chinese, but would be ungrammatical in English unless the dropped pronouns and subjunction *if* are present in the sentence. Therefore, the overuse of cohesive markers in translations in this study may have two sources: one is the tendency to make translations more explicit, which is also the case for English as target language (Volansky *et al.* 2013) and potentially other languages (Pápai 2004), confirming the explicitation universal. The other source is interference from the source languages which use more cohesive markers than Chinese, and as a result are likely to have more such markers in their translations.

Examining this interpretation more thoroughly would require parallel corpora where the translation *process* can be studied, which we leave for future work. Yet, as described by Blum-Kulka (1986: 34), to examine *explicitation* properly, we need to tease apart the two sources of overused cohesive markers. In an ideal setting, the source language and target language should have similar preferences for the explicitness of cohesion, for example, in close cousins such as Spanish and Portuguese, so that differences in frequency can only be the result of the translation process, thus corroborating explicitation.

To summarize, if we measure the explicitness of a text by the frequency of cohesive markers, then translations into Chinese are generally more explicit than the Chinese originals. In fact, five cohesive markers provide enough information for the classifier to distinguish between originals and translations correctly in more than 90% of the cases. However, the use of pronouns and proper nouns in our case is potentially affected by the properties of the source and target languages and by writing styles. Consequently, we find interference rather than explicitation.

4.2 The simplification hypothesis

To test the simplification hypothesis, we perform classification tasks first using commonly used *lexical* features and then *syntactic* features, which have rarely been explored before. The results are shown in Table 2.

4.2.1 Lexical simplification

Word-based TTR reaches an accuracy of 72.15% (see Table 2). Figure 3 shows that all translations have lower TTR rates compared to the Chinese originals. Japanese and Korean again pattern together, with the lowest TTR among all translations. This supports the simplification hypothesis in that translations have lower lexical diversity. However, translations and originals are nearly indistinguishable when using character-based TTR, and the results are close to chance.

Figure 3 also shows that *lexical density* is lower in translations from Indo-European languages, that is, they use more function words. It is interesting that Japanese and Korean translations pattern with Chinese originals, sharing a higher lexical density. Thus, lexical density seems to be dependent on the source language, which refutes the claim of a universal feature. Again, these results argue for interference rather than for simplification.

Other measures such as *mean word length* and *mean sentence length* suggest that translations use shorter words and sentences in general, lending support to the simplification hypothesis. Their accuracies are all above 70% (see Table 2). In fact, the Chinese originals have longer words and sentences than all translations, without exception (see Figure 3). This is different from the study on English (Volansky *et al.* 2013), where the English originals rank 8th in sentence length among

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Table 2. Classification accuracy for our Chinese experiments in comparison to English (Volansky *et al.* 2013): Simplification features

Lexical features	Chinese	English
TTR 1	72.33	72
TTR 2	72.15	72
TTR 1 char	54.95	NA
TTR 2 char	56.28	NA
Lexical density	76.85	53
Mean word length	78.46	66
Mean sentence length	77.75	65
Mean word rank	77.49	69
Mean character rank	75.88	NA
Five most frequent characters	93.25	64
Mean tree depth	53.27	NA
Syntactic features		
Complex NP/Cl	73.35	NA
VP/Cl	83.05	NA
RC/Cl	69.22	NA
CFG rule entropies	89.27	NA
CFG rule types	93.25	NA
CFG rule entropies + types	93.72	NA

the 10 translations. The result for *mean character rank* suggests that the Chinese originals are more likely to have rare characters, that is, characters that are ranked very low on a frequency list. This result is in line with results by Xiao and Hu (2015), who found the same tendency for all four genres they investigated.

Therefore, apart from character-based TTR, all features for lexical complexity and length of linguistic units do support the simplification hypothesis. For many features, we see that the accuracy in our experiment is roughly 10% higher than in the English study.

Feature values of *N most frequent characters* are shown in Figure 4. Note that the ranking is based on a Chinese corpus (Da 2004) of 200 million characters[j]. It is interesting that using only the top five characters reaches above 93% accuracy (Table 2). We see that for 的 (particle DE), 是 (copula, Eng.: be), and 不 (Eng.: not/no), the translations consistently use them more often than the originals. This is in line with previous studies (He 2008; Xiao 2010; Hu *et al.* 2018) and can be interpreted as translations preferring very frequent words, that is, they try to conform to the norms of the target language (Baker 1996).

For the character 一 (Eng.: one), we observe that Japanese and Korean pattern together, with similar values to original Chinese. This is another case of source language interference, resulting most likely from the fact that 一 is the standard translation for indefinite articles from Indo-European languages. Chinese, as well as Japanese and Korean, does not have articles. While

[j] The list is available at http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

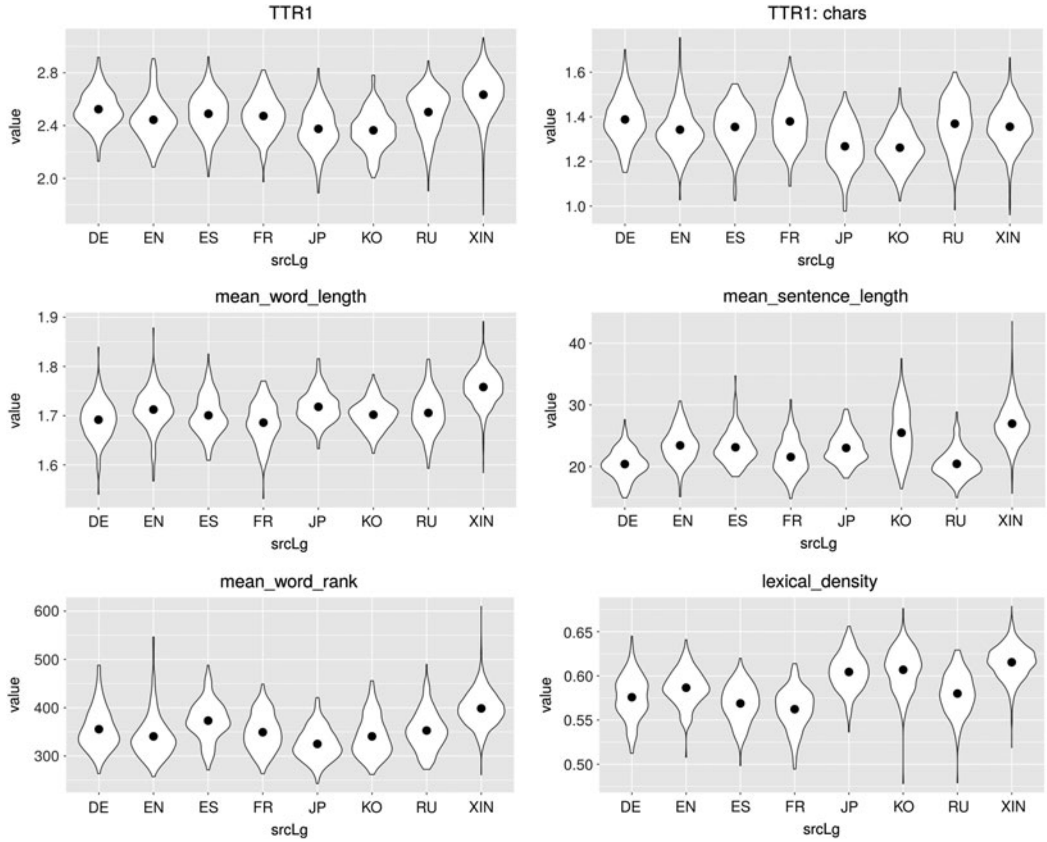


Figure 3. Lexical simplification features in translations and original Chinese (XIN).

it is possible to use a pseudo-indefinite article (一), this is dispreferred in original Chinese. For instance, 他是好人 (Eng.: He is nice guy) is preferred over 他是 一个 好人 (Eng.: He is one + classifier nice guy). However, translators tend to opt for the cognitively easier solution to keep the articles. In fact, having such redundant articles has long been identified as an influence from Indo-European languages (Wang 1943, 1958) and as a characteristic of so-called Europeanized or translation-flavored Chinese (He 2008). Our results provide solid empirical evidence that Europeanized Chinese is a more appropriate description since not all translations overuse the articles. Translations from languages that also lack articles do not overuse them[k].

4.2.2 Syntactic simplification

The results of the syntactic simplification features are shown in the second part of Table 2.

Count-based features. Recall that the first three features are simple counts of complex NPs, VPs, and relative clauses per clause. Among them, the count of VPs performs the best, achieving 83.05% accuracy. The other two reach accuracies of around 70%. These results demonstrate that the three features have some discriminative power in distinguishing translations and originals. A closer look at their distribution in Figure 5 shows that overall, Chinese originals use more complex NPs (excluding relative clauses) than the translations, suggesting greater complexity. This would support the simplification hypothesis. However, if we look at relative clauses, which inevitably form complex NPs, the reverse is true. That is, translations from all the source languages tend

[k] Russian is a surprise here since it lacks articles but patterns with the other European languages.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

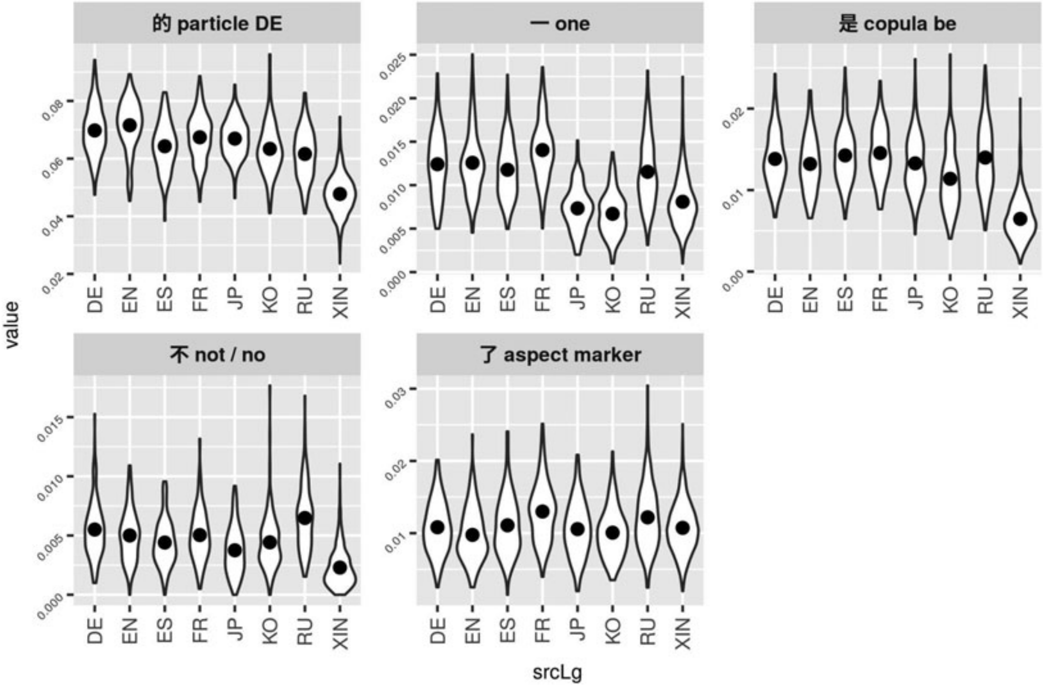


Figure 4. Five most frequent characters in translations and original Chinese (XIN).

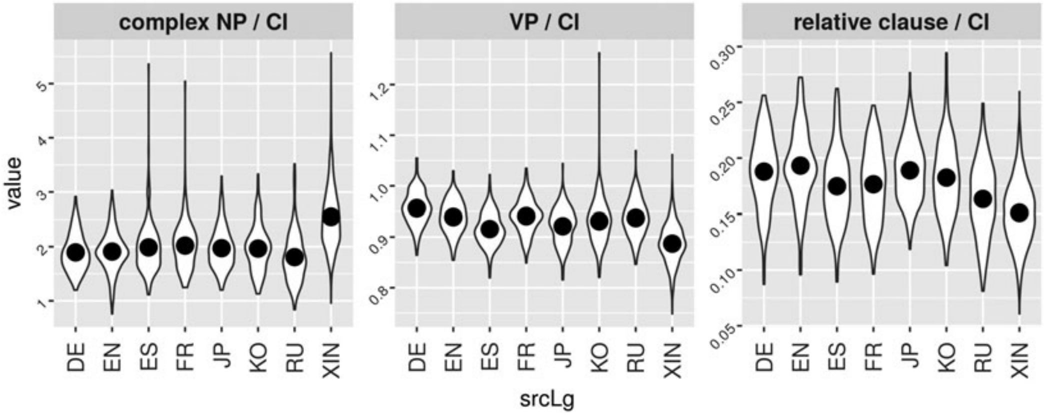


Figure 5. Syntactic features for simplification in translations and original Chinese (XIN).

to have more relative clauses. This seems to be truly universal despite obvious differences in the source languages. One explanation could be that this is a case of explicitation, assuming that the translators make complex NPs more explicit by "spelling them out" in the translation. An example of this would be the complex NP "a very complex idea" being translated into "an idea that is very complex." However, to test this hypothesis, we would need access to the original texts in the source languages.

We also see in Figure 5 that translations tend to use more VPs than the originals. Note that the VPs in this operationalization do not include the copula.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Table 3. Comparison of entropies of CFG rules. Upper half: H(O) < H(T); lower half H(O) > H(T). ***: $p < 0.001$. O: originals. T: translations

Phrase head	H(O)	H(T)	sig.
ADJP: adjectival phrase	0.337	0.455	***
ADVP: adverbial phrase	0.216	0.321	***
CP: complementizer phrase	1.177	1.499	***
DNP: “XP + DEG” phrase	1.568	1.681	***
DP: determiner phrase	1.316	1.393	***
PP: prepositional phrase	1.572	1.62	***
VP: verb phrase	4.636	4.682	***
CLP: classifier phrase	0.056	0.036	***
LCP: localizer phrase	1.237	1.204	***
NP: noun phrase	4.031	3.834	***
QP: quantifier phrase	1.973	1.888	***
Sum over all heads	24.750	25.456	***

Entropy-based features. First, we see that entropy-based features work quite well. If the entropies of CFG rules headed by each phrase (NP, VP, etc.) are used as features, we achieve 89.27% accuracy. If we only use the number of types of CFG rules for each phrase, we achieve 93.25%. Our assumption is that if originals have higher entropy and more types than translations, we can say that originals are syntactically more complex than translations. However, when we look into CFG rules headed by different phrases, the opposite is true. For most of the major phrases, such as ADJP, CP, DP, PP, and VP, the originals exhibit a significantly lower entropy (see left of Table 3). The distributions in the number of rule types are the same as their entropies, except for LCP. Only four phrases exhibit a higher entropy in the originals: CLP, LCP, NP, and QP. Two of them are, in fact, structures specific to Chinese: CLP is the classifier phrase unique to Chinese (as well as Japanese and Korean). LCP is a localizer phrase, which is a discontinuous prepositional phrase of the form “P + NP + localizer”, also unique to Chinese.

The sum of entropies of CFG rules over all phrasal heads in Table 3 is also higher in translations (25.456 > 24.750). Based on this entropy, it seems reasonable to conclude that translations are generally more complex and diverse in terms of syntactic structure, contrary to the simplification hypothesis.

Therefore, our results based on entropy-related measures suggest that the simplification hypothesis may be too simplified to capture the whole picture. A case-by-case analysis is required here. Originals have more diverse ways of creating specific phrases such as NP and QP. Other phrases such as ADJP, ADVP, PP, and VP, however, have a more complex distribution in the grammar of translations. If we look at the bigger picture, in most cases, translation have a more complex and more unpredictable grammar, thus contradicting the simplification hypothesis.

To summarize, we find that many of the simplification features work well in our classification tasks. A closer look at the *lexical* features reveals that Chinese originals are indeed more complex than the translations from multiple source languages. Our *syntactic* features, however, suggest a more complicated picture. There is no conclusive evidence that all translations are syntactically simpler than originals. Originals have more complex NPs, but their grammar is generally less

Table 4. Classification accuracy for our Chinese experiments in comparison to English (Volansky *et al*. 2013): Normalization features

Features	Chinese	English
Repetitions (content words)	68.66	55
Repetitions (nouns only)	74.48	NA
Average PMI	85.79	52
Aspectual markers	66.19	NA
Measure words	91.13	NA
Sentence-final particles	70.67	NA
ba-structure	54.81	NA
bei-structure	52.76	NA
Idioms: summed counts	54.45	NA
Idioms: individual	81.30	NA

complex than that of translations. Future studies need to find more fine-grained ways of looking at syntactic complexity.

4.3 The normalization hypothesis

The normalization hypothesis states that translations are likely to exaggerate characteristics unique in the target language.

The *repetitions* feature is able to distinguish translations from originals 68.66% of the time (see Table 4), which is considerably higher than the corresponding 55% in the English setting. Figure 6 shows the distribution of two normalization features. We see once again that Japanese and Korean pattern together, close to the Chinese originals. Thus avoiding repetitions, as discussed by Ben-Ari (1998), is likely to be source language specific rather than universal since translations from Japanese and Korean clearly use similar numbers of repetition as the originals. If only the repetitions of *nouns* are considered, then the accuracy reaches 74.48%, with original Chinese having the highest number of noun repetitions. This supports the repetition avoidance assumption in translations, but there is potentially a confound, which is the particular anaphora reference strategy in Chinese. Traditionally, when the antecedent is an *inanimate* NP, Chinese speakers would prefer to repeat the whole NP rather than to use a pronoun "it" to refer to it (Lv 1942; Wang 1943; Chen 1987). Hence, the more frequent noun repetition in original Chinese. However, while translating from Indo-European languages where "it" is routinely used to refer to inanimate objects, translators gradually start to follow suit, which results in fewer noun repetitions and an overuse of "it" (see (He 2008, ch. 4) for a detailed corpus analysis and discussion). Considering the prevalence of NP repetition as a strategy for anaphora reference in Chinese, as reported in previous studies, we assume that this feature is more likely to model a specific feature of the target language, rather than showing normalization.

The *average PMI* measures the degree of collocation usage in one text chunk. A higher average PMI indicates more collocations, with the initial assumption that translations tend to have more collocations, conforming to the conventions of the target language. However, this is not supported by our results, as shown in Figure 6. The originals have a higher average PMI than all translations,

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

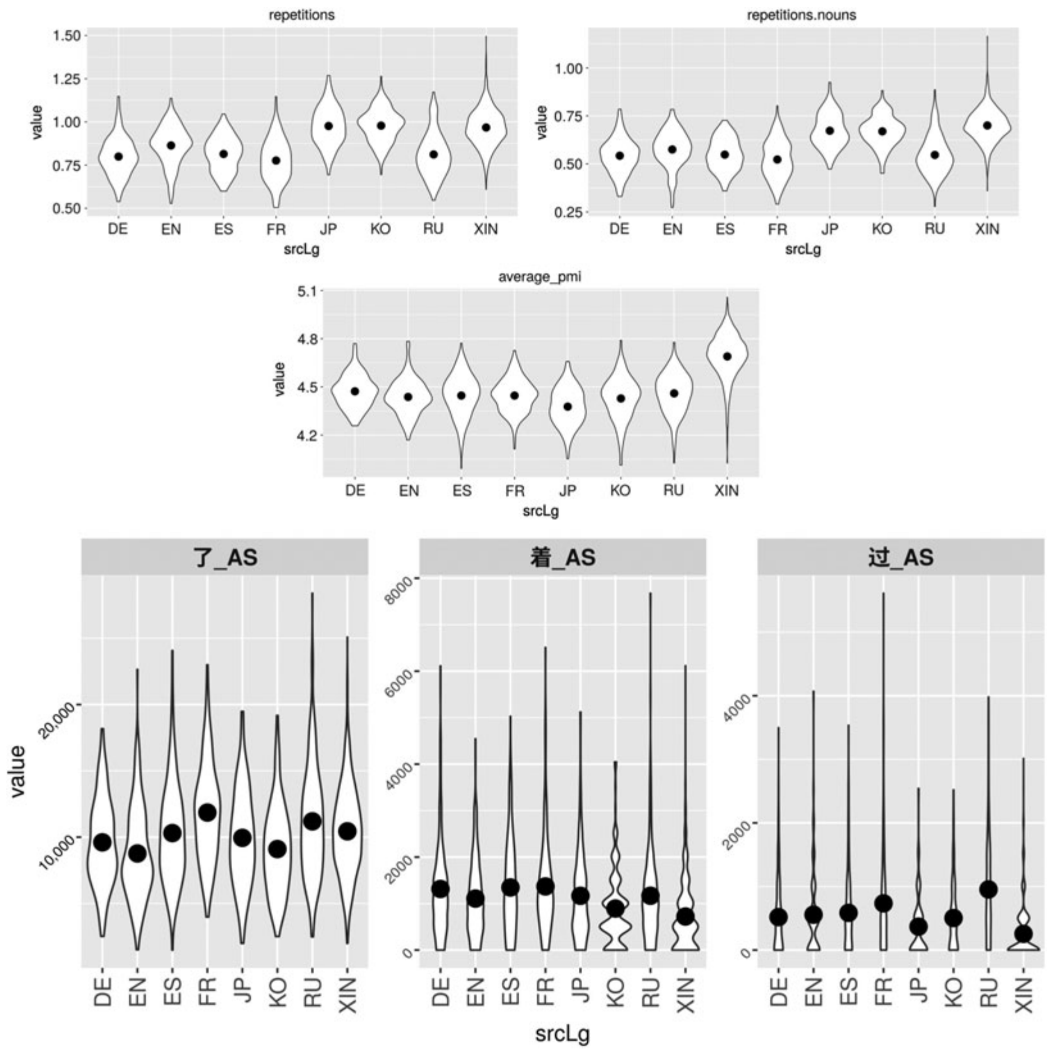


Figure 6. Normalization features in translations and original Chinese (XIN).

which is in line with the English study where originals are found to have "far more associated bigrams" (Volansky *et al.* 2013).

The three *aspect markers* reach an accuracy of 66.19%. The perfective marker 了, most frequent and hardest to acquire for language learners, seems to be used differently in all translations, with no clear pattern. For the imperfective marker 着, we observe that the Korean translations and the Chinese originals have the lowest counts. For the experiential marker 过, the Japanese translation and the Chinese original use it least frequently. Again, the results are contrary to the normalization hypothesis: for a target language-specific grammatical property, translations do not overuse or "exaggerate" it.

Using 114 *measure words* as features reaches a very high accuracy of 91.13%. Figure 7 shows the distribution of the 10 highest ranked measure words. We observe that the translations and the originals use them rather differently, which is why the experiments reach such a high accuracy. For instance, Japanese and Korean translations use the lowest number of 个 (Eng.: item); the Chinese originals have the lowest number of 种 (Eng.: type/kind), but the highest number of 项

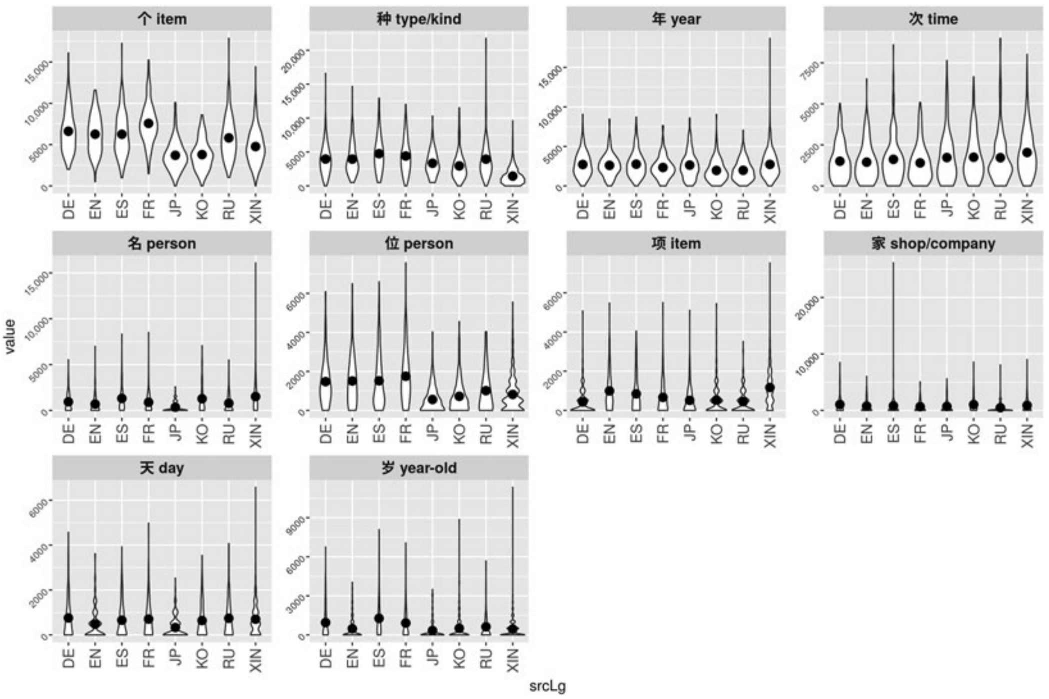


Figure 7. The 10 most frequent measure words (classifiers) plus their noun types (e.g., 个 (Eng.: item) means that "5个apple" is grammatical since apple is an item).

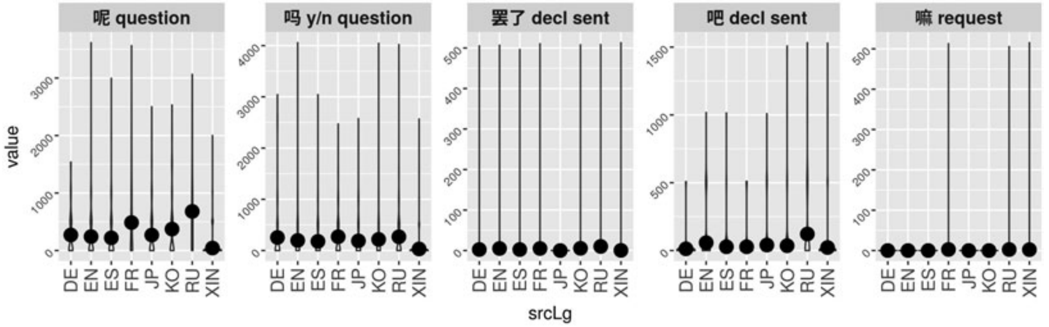


Figure 8. The top 5 sentence-final particle features with their functions.

(Eng.: item); and Japanese, Korean, and Russian translations use fewer 位 (Eng.: person). Thus, it seems that translations overuse some of the measure words, but not all of them. The characteristics of the source language also play a role here, as we see the different distributions in different translations. Compared with translations from seven source languages individually, originals have more measure words than six of them. In fact, only translations from Spanish use more measure words than originals (68.20 vs. 65.60 per chunk). This shows that most translations are underusing measure words rather than exaggerating this feature.

With 17 *sentence-final particles* as features, the accuracy is well above chance level, at 70.76%. Again, we examine the top features, illustrated in Figure 8. It is clear that all translations make more use of particles that denote questions: 呢 (ne) and 吗 (ma). This is an interesting discovery,

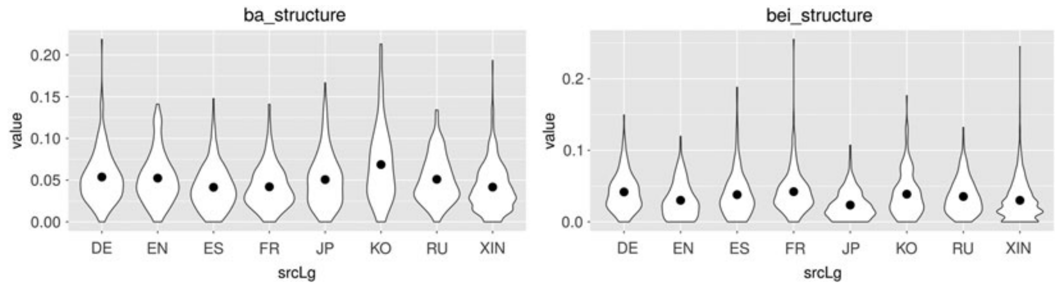


Figure 9. *ba-* and *bei*-structures.

with two possible scenarios: (1) The source texts of all translations use more questions than the originals; (2) In the translation process, declarative sentences in the source text are rendered as rhetorical questions. An example for the second scenario would be: "This is a nice book" translated as "Isn't this a nice book?."

After a quick search in a randomly chosen sample of 200 sentences with 呢 (out of 1020 in total in all translations), we found that only 39 out of the 200 sentences with 呢 can count as rhetorical questions. That is, the particle 呢 is indeed used in questions most of the time. This seems to suggest that translations tend to have more questions than the originals, potentially due to the style of news writing in different languages.

Next, we look at *ba and bei structures*. These two features have similar classification accuracy, only slightly above chance. The feature distributions in Figure 9 demonstrate no clear pattern. However, independent-samples t-tests show that for the *ba*-structure, the normalized frequency in the originals ($M = 0.0417$) is significantly lower than that of the average of all translations combined ($M = 0.0512$); $t(2737.2) = -8.0933$, $p = 8.639e{-16}$. The same is true for the *bei*-structure, where the difference between the originals ($M = 0.0300$) and the translations ($M = 0.0356$) is again significant; $t(2776.1) = -5.3462$, $p = 9.711e{-08}$.

As a result, it becomes challenging to interpret the classification results and the t-tests results meaningfully. The results from the t-tests indicate that it is extremely unlikely that the translations and the originals are drawn from the same population. It cannot tell us whether we can reliably distinguish the two classes based solely on one feature. In this sense, the classification accuracy can be considered a more "conservative" measure for the features since a much greater difference is needed to achieve good classification accuracy. Consequently, we need to keep in mind that the significant features in previous t-test-based studies may not be good features in a classification task. Also, a classifier is able to make use of the complicated distribution of multiple features so as to classify new data, which is beyond the capability of statistical tests.

When looking at the Chinese idioms *chengyu*, the total count of all idioms only gives an accuracy of 54.45%, slightly above chance. The t-test again shows a significant difference between translations ($M = 0.0039$) and originals ($M = 0.0033$), $p = 5.847e{-12}$. Thus, we conclude that translations use significantly more idioms, supporting the normalization hypothesis, but the difference is not large enough for a classifier to reliably distinguish translations from originals.

Nevertheless, translations and originals do use the idioms quite differently, indicated by the 81.30% accuracy of treating each idiom as a single feature. Appendix D shows the 20 most frequent idioms and their frequencies in the two varieties. For example, in the originals, the most frequent two idioms are political slogans of the government: 艰苦奋斗 (Eng.: work hard) and 实事求是 (Eng.: seek truth from facts). The most frequent idioms in the translations are 引人注目 (Eng.: eye-catching), 成千上万 (Eng.: thousands of), 无论如何 (Eng.: anyways/no matter what), 众所周知 (Eng.: as is known to all), and 有朝一日 (Eng.: some day). When we look at the idiom

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Table 5. Classification accuracy for our Chinese experiments in comparison to English (Volansky *et al.* 2013): Interference features

Features	Chinese	English
Character unigram	100	100
Word unigram	100	100
POS trigram	99.46	98
CFG rules all	100	NA
CFG rules top 100	99.43	NA
CFG rules top 50	98.85	NA
CFG rules top 20	97.42	NA
CFG rules top 5	94.11	NA
NP-headed CFG rules all	96.95	NA
NP-headed CFG rulestop 30	95.90	NA

usage in specific translations, we again find that it is dependent on the source language. For example, Japanese translations use more 引人注目 (Eng.: eye-catching) than any other translations or the originals. French translations overuse 无论如何 (Eng.: anyways); Spanish translations overuse 成千上万 (Eng.: thousands of). These are interesting observations to be further explored when we have access to the source texts.

4.4 *The interference hypothesis*

The classification results of interference features are presented in Table 5.

4.4.1 n-grams

Character and word *n*-grams work extremely well in this task, in line with previous studies (Volansky *et al.* 2013; Hu *et al.* 2018). However, looking at the 100 features ranked highest by Gain Ratio, we notice that many of these features are content characters/words. For example, out of the 10 most discriminating characters for the originals, 2 are Chinese family names (*Liu* and *Yang*). Among the top word unigrams, many are names of Chinese governmental bodies. POS trigrams also perform very well, close to character and word unigrams. Since many of the interesting structures shown in POS trigrams are also captured by CFG rule features, we will leave feature analysis to the next section.

4.4.2 CFG rules

CFG rules as features capture the syntactic structures of the texts, without any topic-related information, unlike word or character *n*-grams. As shown in Table 5, with only the five highest ranked features, we reach an accuracy of 94.11%. Note that the translations are based on seven source languages, which means that our syntactic features must be consistent across all translations to reach such a high accuracy.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Before we present a detailed feature analysis, we want to point out that a comparison between the features in this study and in our previous work (Hu *et al.* 2018) can be very fruitful: 99% of the translations examined in Hu *et al.* (2018) are translated from English, but in multiple genres: news, fiction, general prose, and science. Therefore, if a translation feature is found to be important in both studies, there is a good chance that it is a robust feature across source languages and genres. Such features will be truly "universal" for Chinese translations, and we will focus on analyzing these features, in addition to high ranking features.

Rules typical for originals. We start with the three features that are typical for originals, which were the most important in this and our previous study (Hu *et al.* 2018: Table 5). They are underlined in Table 6a, with graphical illustrations in Appendix E.

First, there are two VP rules: VP → VP PU VP PU VP (3 VPs) and VP → VP PU VP PU VP PU VP (4 VPs). After searching for sentences with such structures, we found multiple predicates connected by either the comma , or "、", the unique punctuation in Chinese[1]. Indeed, parallel VPs within one sentence is characteristic of Chinese writing. We found one example with five VP predicates in a Chinese original: 各级人民法院要 ($_{VP}$ 认真学习), ($_{VP}$ 联系实际), ($_{VP}$ 制定措施), ($_{VP}$ 落到实处), ($_{VP}$ 抓出成效)" (Eng.: courts at all levels must ($_{VP}$ study . . . carefully), ($_{VP}$ integrate theory with practice), ($_{VP}$ make plans), ($_{VP}$ take actions), ($_{VP}$ be effective)). Note that these predicates do not have to be in chronological order in Chinese or have an explicit conjunction.

The same is true for NPs. For example, we have discussed the rule NP → NN PU NN in Table 6a previously (Hu *et al.* 2018). Here it should suffice to say that Chinese originals prefer using the punctuation (PU) "、" to conjoin common nouns (NN), whereas the translations prefer the conjunction 和 (Eng.: and).

In summary, these three rules are consistent for the Chinese originals, but they are clearly not typical of Chinese translationese. If structures that are carried over from source languages to the target language showcase *interference*, then these rules show the opposite.

Next, we turn to the top rule in the originals, VP → VV NP QP and VP. We find structures such as:

- ($_{VP}$($_{VV}$ 产) ($_{NP}$钢) ($_{QP}$ 771.05 万吨))
  ($_{VP}$($_{VV}$ produced) ($_{NP}$ steel) ($_{QP}$ 7.7105 million tons)
  Eng.: produced 7.7105 million tons of steel
- ($_{VP}$($_{VV}$ 节约) ($_{NP}$ 运输成本) ($_{QP}$ 40多万元))
  ($_{VP}$($_{VV}$ saved) ($_{NP}$ transportation costs) ($_{QP}$ 400,000 more RMB)
  Eng.: saved more than 400,000 RMB transportation costs

In English, German, Spanish, and French, the NP and QP are usually ordered differently, as VV QP NP. In fact, the "VV QP NP" order is also acceptable in Chinese for the second example above, but Chinese news emphasizing the amount of savings would prefer the "VV NP QP" order to put the final emphasis on QP, giving the flavor of idiomatic Chinese. To our knowledge, this has not been reported in any previous studies. Note that since this rule abstracts to the phrase level, no *n*-gram features will be able to detect it.

Rules typical for translations. Again, we first look for features that occur across all languages, in this study and in our previous work (Hu *et al.* 2018) (underlined in Table 6a). ADVP → CS indicates a higher frequency of subordinating conjunctions in translations. CP → ADVP IP in effect captures the same structure, as a search in the trees shows that 98% of the ADVPs serve as subordinating conjunctions, making the CP a subordinating clause. NP → DNP NP matches an

[1] It is usually used to delimit items in a list.

Table 6. CFG rules ranked by Gain Ratio. O: original. T: translation.

Rank	CFG rule	Predicts
1	VP → VV NP QP	*O*
2	VP → NP PP VP	*O*
4	VP → VP PU VP PU VP	*O*
5	NP → NN PU NN	*O*
6	ROOT → FRAG	*O*
7	NP → NN NN NN NN	*O*
9	NP → NP NP	*O*
12	NP → NP QP NP	*O*
13	NP → NP PU NP	*O*
16	NP → NP NP NP	*O*
18	VP → VP PU VP PU VP PU VP	*O*
19	NP → NP ADJP NP	*O*
3	ADVP → AD	*T*
8	ADVP → CS	*T*
10	CP → ADVP IP	*T*
11	CP → IP SP	*T*
14	ROOT → CP	*T*
15	VP → VC VP	*T*
17	NP → DNP NP	*T*
20	VP → VV VP	*T*

(a) Top 20 CFG rules, underlined rules correspond to rules by Hu *et al.* (2018)

Rank	CFG rule	Predicts
1	NP → NN PU NN	*O*
2	NP → NN NN NN NN	*O*
3	NP → NP NP	*O*
4	NP → NP QP NP	*O*
5	NP → NP PU NP	*O*
6	NP → NP NP NP	*O*
8	NP → NP ADJP NP	*O*
9	NP → NP QP ADJP NP	*O*
10	NP → NP NP ADJP NP	*O*
11	NP → NN PU NN PU NN	*O*
12	NP → NN NN NN	*O*
13	NP → NN PU NN NN	*O*
16	NP → NN NN	*O*
18	NP → NR PRN	*O*
20	NP → NN NN CC NN NN	*O*
7	NP → DNP NP	*T*
14	NP → PN PN	*T*
15	NP → CP	*T*
17	NP → PN	*T*
19	NP → QP DNP NP	*T*
23	NP → QP CP NP	*T*
30	NP → DP CP NP	*T*

(b) Top rules headed by NP

NP with a preceding noun modifier DNP (which is of the form "XP + particle DE 的"), more frequently used in translations. This suggests that NPs in translations are more likely to be modified by nouns, with the particle 的 as the marker (for a detailed discussion see (Hu *et al.* 2018)).

We can also focus on a subset of CFG rules, for example, those headed by *NP*. Using only NP-headed rules as features means that the classifier will only have information about the *internal structure* of NPs to distinguish Chinese originals and translations. We reach an accuracy of over 95% using only 30 features. This suggests that the structure of NPs alone is enough for the classifier to make reliable decisions. The top ranking features are shown in Table 6b. Here, we focus on the rules typical for translations, with their distributions illustrated in Figure 10. We first

Table 7. Quantifier in: ($_{NP}$ ($_{QP}$ quantifier classifier) (DNP/CP) (NP)). DNP: an NP modifier. CP: relative clause. XIN: original Chinese

Raw counts	DE	EN	ES	FR	JP	KO	RU	XIN
Quantifier = "one"	761	879	812	1071	229	249	607	2586
All QP	1017	1136	1058	1287	337	403	831	3988
Percentage (%)	75.05	77.38	76.75	83.22	67.95	61.79	73.04	64.86

see that for five out of the seven rules, Japanese and Korean translations again pattern together with the originals, suggesting that the Japanese and Korean translations tend to have NPs that are structured similarly to original Chinese texts.

One interesting rule in Figure 10 is NP → PN PN, where a NP is composed of two consecutive pronouns. A search in the parse trees finds mostly a personal pronoun followed by a reflexive pronoun, for example, 我们 自己 (literally: we self; Eng.: we ourselves), 他 本人 (literally: he self, Eng.: he himself). This has not been reported in previous literature and indicates that either Indo-European texts favor such a structure, or translators overuse it in the translation process, which can be verified in the future with a parallel corpus.

The three features at the bottom of Figure 10 are very similar in that they all have two modifiers preceding the head NP, where the second modifier (DNP/CP) ends with the particle DE 的. When searching for NP → QP DNP NP and NP → QP CP NP, we find that the quantifier in QP is very often the word "one," forming the structure ($_{NP}$($_{QP}$ "one" *classifier*) (DNP/CP) (NP)), for example, ($_{NP}$($_{QP}$一 种) ($_{DNP}$新型的) ($_{NP}$关系)) (Eng.: ($_{NP}$($_{QP}$ one *classifier*) ($_{DNP}$ new) ($_{NP}$ relationship))). This is parallel to the "articles" in Chinese (briefly discussed in Section 4.2.1). Recall that Japanese, Korean, Russian, and Chinese do not have articles and that indefinite articles are usually translated into Chinese as "one + classifier + NP." If there were an interference phenomenon, we would expect translations from article-less languages to have fewer such structures with "one" as the quantifier; translations from languages with articles should have more "one + modifier + NP". To verify this, we calculate the proportion of structures where the quantifier is "one" to those with other quantifiers ("two" or "many", or others)[m], with results shown in Table 7.

It is obvious that for languages that do have articles, that is, German, English, Spanish, and French, there are more structures using "one" as the quantifier. The opposite is true for Japanese, Korean, and Russian (underlined in Table 7), suggesting source language interference. Our results confirm previous observations (Wang 1958; He 2008) by not only looking at frequencies of "one" (Figure 4) but also at its frequencies in complex NPs that involve another modifier. Most importantly, the multiple source languages used in this study allow us to connect the observation to typological differences among all source languages, that is, languages with articles versus languages without articles. This clearly demonstrates the interference effect and indicates that it may be traced back to the typological properties of the source languages.

In summary, with regard to interference, we have seen in previous sections that each source language has its distinctive, favored cohesive markers and idioms. In this section, we discover features in typologically distinct language groups, for example, languages with indefinite articles. Finally, there are also structures with higher frequencies across all translations such as the rule `NP → DNP NP`. In this sense, interference is at work simultaneously at multiple levels, sometimes for individual source languages, sometimes for language groups, and others for translationese as a whole.

[m] The patterns we use can be found in Appendix C.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

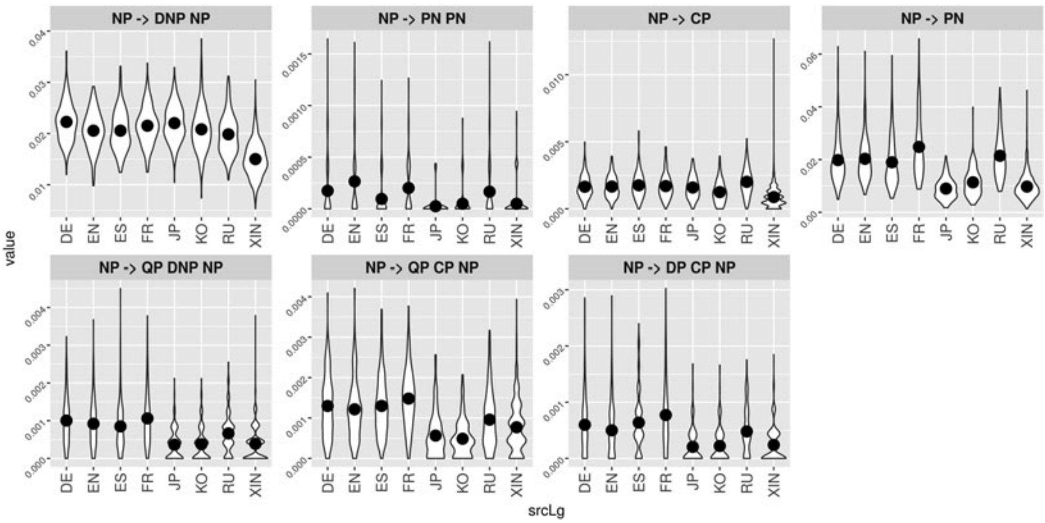


Figure 10. Top translation-dominant CFG rules headed by NP.

5. Conclusion

In the work reported here, we have investigated the characteristics of translations into Chinese along with text originally written in Chinese. Our corpus consists of original Chinese texts from Xinhua and translations from seven languages into Chinese, also published by Xinhua.

Our methodology is based on a machine learning approach that classifies texts into originals and translations. Looking at the classifier's accuracy when given specific characteristics allows us to determine how discriminative these characteristics are. We have shown that translated Chinese texts differ from non-translated ones, thus constituting a variety that is different from original Chinese. Some of these differences originate from *universal* tendencies in translations while others are a result of source language *interference*. The latter phenomenon creates different "dialects" of translated Chinese, depending on the properties of the source language.

Our approach is novel in several aspects: we look at translations into Chinese rather than into English, adding valuable empirical results from a non-Indo-European target language; we look at a typologically diverse group of source languages, complementing previous studies on Chinese. Additionally, we use a finer grained metric than pure counts or frequencies: we use a machine learning approach, which allows us to integrate syntactic features. Since the number of these features is high, a purely count-based approach cannot accommodate such features very well.

In terms of the four specific hypotheses, we find evidence supporting the explicitation hypothesis, as shown in the discriminative power of cohesive markers and their overuse in translations. Simplification is generally supported by the lexical features, different from the findings for English (Volansky *et al*. 2013). However, the grammar of translations has a higher entropy than the grammar for originals, which contradicts the simplification hypothesis. As for normalization, we do not find an overall tendency of translations "exaggerating" uniquely Chinese features. In fact, translations have a lower average PMI, indicating fewer collocations. They do use significantly more *ba-*/*bei-* structures and idioms, but these features only yield at-chance accuracy in classification. Finally, interference is manifested in many places throughout our study. For example, (1) the overuse of pronouns in translations from non-pro-drop languages, (2) the higher frequency of

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

quantifier phrases involving "one" in translations from languages with articles, and (3) the patterning together of Korean and Japanese translations. These are all cases of interference from specific source languages. Some of these effects are at the lexical level while others are syntactic in nature. Thus, we can confirm that interference is a major force, and it creates multiple dialects of translationese. Another important finding is that our accuracies on Chinese are in most cases (considerably) higher than those on English (Volansky *et al.* 2013), except for single naming and TTR. Since we have very similar experimental settings and an almost identical implementation, this is unlikely due to chance. It suggests that how well a feature performs in classification is closely related to the distance between the source languages and the target language. When they are very distant as in our experiment, classification becomes an easier task. We therefore need empirical evidence from a wider variety of languages before deciding which features are "universal" among all translations. Our analysis of Chinese is an initial step in this direction toward building general theories of translations, and our method can be easily extended to other languages[n].

Finally, our study suggests a number of new challenges. The first concerns the question of how to reconcile results from different methodologies such as *t*-tests and classification accuracies. Is a statistically significant difference between translations and originals enough to identify translationese features? If classification accuracy is considered, what accuracy is required for a feature to be accepted as discriminative? Second, after the advent of large scale corpora, translation studies have gradually shifted their focus from the relation between translations and their source text (e.g., Blum-Kulka 1986) to textual, stylistic properties of translations *per se* (Baker 1993). In our study, we see a need to go back to the source texts again to uncover the ultimate *source* of translationese. In an ideal setting, a corpus needs to be composed of three types of texts: source texts, target texts, and comparable texts originally written in the target language. This structure guarantees that we can gain a deeper understanding of translationese and the source from which it originates.

Acknowledgements. The authors thank Ruoze Huang, Chien-Jer Charles Lin, Jingyi Guo, Yueqi Zhu, and the computational linguistics colloquium at Indiana University for helpful discussions. The authors are grateful for the comments and suggestions from Aini Li and our two anonymous reviewers. The authors also thank Noam Ordan for providing information about the implementation for the study on English. The first author is partly funded by China Scholarship Council.

References

Baker M. (1993). Corpus linguistics and translation studies: Implications and applications. In Baker M., Francis G. and Tognini-Bonelli E.(eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, pp. 233–250.

Baker M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, **7**(2), 223–243.

Baker M. (1996). Corpus-based translation studies: The challenges that lie ahead. In Somers H.(ed), *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, vol. 18. Amsterdam and Philadelphia: Benjamins, pp. 175–186.

Baroni M. and Bernardini S. (2005). 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.

Becher V. (2011). *Explicitation and Implicitation in Translation. A Corpus-Based Study of English-German and German-English Translations of Business Texts*. PhD Thesis, University of Hamburg.

Ben-Ari N. (1998). The ambivalent case of repetitions in literary translation. Avoiding repetitions: A "universal" of translation? *Meta: Journal des Traducteurs/Meta: Translators' Journal*, **43**(1), 68–78.

Blum-Kulka S. (1986). Shifts of cohesion and coherence in translation. In House J. and Blum-Kulka S. (eds), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Gunter Narr, Tuebingen, Germany, pp. 17–35.

Bykh S. and Meurers D. (2014). Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, Dublin, Ireland, pp. 1962–1973.

[n] Our code is available at: https://github.com/huhailinguist/translationese.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Cappelle B. and Loock R. (2017). Typological differences shining through: The case of phrasal verbs in translated English. In De Sutter G., Lefer M.-A. and Delaere I. (eds), *Empirical Translation Studies: New Theoretical and Methodological Traditions*. Walter de Gruyter, Berlin, Germany, pp. 235–263.

Cartoni B., Zufferey S., Meyer T. and Popescu-Belis A. (2011). How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland, OR, pp. 78–86.

Chen J.W. (2006). *Explicitation Through the Use of Connectives in Translated Chinese: A Corpus-Based Study*. PhD Thesis, The University of Manchester.

Chen P. (1987). Discourse analysis of zero anaphora in Chinese. *Chinese Philology* (In Chinese). **5**, 363–378

Chen Z., Boston M.F. and Hale J.T. (2009). Using entropy to evaluate child language performance. In *The 22nd CUNY Conference on Human Sentence Processing*, Davis, CA.

Church K.W. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* **16**(1), 22–29.

Da J. (2004). A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction. In *Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese*, Beijing, China, pp. 501–511.

De Sutter G., Lefer M.-A. and Delaere I. (eds). (2017). *Empirical Translation Studies: New Methodological and Theoretical Traditions*, vol. 300. Walter de Gruyter, Berlin, Germnay.

Evert S. and Neumann S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In De Sutter G., Lefer M.-A. and Delaere I. (eds), *Empirical Translation Studies: New Theoretical and Methodological Traditions*. Walter de Gruyter, Berlin, Germany, pp. 47–80.

Ferraresi A. and Miličević M. (2017). 5 phraseological patterns in interpreting and translation. Similar or different? In De Sutter G., Lefer M.-A. and Delaere I. (eds), *Empirical Translation Studies: New Theoretical and Methodological Traditions*. Walter de Gruyter, Berlin, Germany, pp. 157–182.

Frawley W. (1984). Prolegomenon to a theory of translation. In Frawley W. (ed), *Translation: Literary, Linguistic and Philosophical Perspectives*. Associated University Press, London, pp. 159–175.

Gellerstam M. (1986). Translationese in Swedish novels translated from English. In Wollin L. and Lindquist H. (eds), *Translation Studies in Scandinavia*, vol. 1. CWK Gleerup, pp. 88–95.

Graff D. (2007). Chinese Gigaword, 3rd Edn. LDC Catalog No.: LDC2007T38, ISBN: 1-58563-455-7.

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

Hale J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass* **10**(9), 397–412.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I.H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* **11**(1), 10–18.

He Y. (2008). *A Study of Grammatical Features in Europeanized Chinese*. Commercial Press (In Chinese), Beijing.

Hu H., Li W. and Kübler S. (2018). Detecting syntactic features of translated Chinese. In *Proceedings of the 2nd Workshop on Stylistic Variations at NAACL-HLT 2018*, New Orleans, LA, pp. 20–28.

Hu X., Xiao R. and Hardie A. (2016). How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis. *Corpus Linguistics and Linguistic Theory* **15**(2), 347–382.

Ilisei I. and Inkpen D. (2011). Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications* **2**(1–2), 319–32.

Ilisei I., Inkpen D., Pastor G.C. and Mitkov R. (2010). Identification of translationese: A machine learning approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, Iasi, Romania, pp. 503–511.

Ke F. (2005). Fanyi zhong de xian he yin (implicitation and explicitation in translations). *Foreign Language Teaching and Research* **37**(4), 303–307 (In Chinese).

Koehn P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Koppel M. and Ordan N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, pp. 1318–1326.

Kunilovskaya M. and Kutuzov A. (2017). Testing target text fluency: A machine learning approach to detecting syntactic translationese in English-Russian translation. In Menzel K., Lapshinova-Koltunski E. and Kunz K. (eds), *New Perspectives on Cohesion and Coherence*. Language Science Press, Berlin, pp. 75–104.

Kwon N., Kluender R., Kutas M. and Polinsky M. (2013). Subject/object processing asymmetries in Korean relative clauses: Evidence from ERP data. *Language* **89**(3), 537.

Laviosa-Braithwaite S. (1996). *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. PhD Thesis, University of Manchester.

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

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

Levy R. and Andrew G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 2231–2234.

Lin C.-J.C. (2011). Chinese and English relative clauses: Processing constraints and typological consequences. In *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, Eugene, OR.

Lin C.-J.C. (2018). Subject prominence and processing filler-gap dependencies in prenominal relative clauses: The comprehension of possessive relative clauses and adjunct relative clauses in Mandarin Chinese. *Language* **94**, 758–797.

Lin C.-J.C. and Hu H. (2018). Syntactic complexity as a measure of linguistic authenticity in modern Chinese. In *26th Annual Conference of International Association of Chinese Linguistics and the 20th International Conference on Chinese Language and Culture*, Madison, WI.

Lu X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* Beijing, **15**(4), 474–496.

Lv S. (1942). *A Sketch of Chinese Grammar*. Commercial Press (In Chinese).

Malmasi S. and Dras M. (2018). Native language identification with classifier stacking and ensembles. *Computational Linguistics* **44**(3), 403–446.

Manning C.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J. and McClosky D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, pp. 55–60.

Mauranen A. and Kujamäki P. (eds) (2004). *Translation Universals: Do they Exist?*, vol. 48. John Benjamins, Amsterdam.

Meyer T. and Webber B. (2013). Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pp. 19–26.

Olohan M. and Baker M. (2000). Reporting that in translated English. Evidence for subconscious processes of explicitation? *Across Languages and Cultures* **1**(2), 141–158.

Pápai V. (2004). Explicitation: A universal of translated text? In Mauranen A. and Kujamäki P. (eds), *Translation Universals: Do they exist?* John Benjamins, pp. 143–164.

Puurtinen T. (2004). Explicitation of clausal relations: A corpus-based analysis of clause connectives in translated and non-translated Finnish children's literature. In Mauranen A. and Kujamäki, P. (eds), *Translation Universals: Do they exist?* John Benjamins, pp. 165–176.

Rabinovich E., Nisioi S., Ordan N. and Wintner S. (2016). On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 1870–1881.

Rabinovich E. and Wintner S. (2015). Unsupervised identification of translationese. *Transactions of the Association of Computational Linguistics* **3**(1), 419–432.

Rubino R., Lapshinova-Koltunski E. and van Genabith J. (2016). Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, pp. 960–970.

Shannon C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**(3), 379–423.

Swanson B. and Charniak E. (2012). Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, South Korea pp. 193–197.

Teich E. (2003). *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Walter de Gruyter Berlin.

Toury G. (1978). The nature and role of norms in translation. In Holmes J., Lambert J. and van den Broeck R. (eds), *Literature and Translation: New Perspectives in Literary Studies*. Acco, Leuven.

Toury G. (1995). *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam.

Volansky V., Ordan N. and Wintner S. (2013). On the features of translationese. *Digital Scholarship in the Humanities* **30**(1), 98–118.

Wang L. (1943). *Contemporary Grammar of Chinese*. Commercial Press (In Chinese) Beijing.

Wang L. (1944). *Theory of Chinese Grammar*. Commercial Press (In Chinese) Beijing.

Wang L. (1958). *History of the Chinese Language*. Zhonghua Book Company (In Chinese) Beijing.

Xiao R. (2010). How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistics* **15**(1), 5–35.

Xiao R. and Hu X. (2015). *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. Springer Berlin.

Xue N., Xia F., Chiou F.-D. and Palmer M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering* **11**(2), 207–238.

Zhu D. (1985). *Dialogues in Grammar*. Commercial Press (In Chinese) Beijing.

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

A Text distribution in our multi-source language corpus

Data sources for each source language. Only showing sources with more than 50 texts for ES. All KO newspapers are included, but they are not listed here because there are too many

Source (in Chinese)	Source (in original language; *in English*)	Language	# Articles
德国《商报》	Handelsblatt; *Commerce Paper*	DE	75
德国《明星》	Stern; *Star*	DE	61
德国《明镜》	Der Spiegel; *The Mirror*	DE	57
德国《星期日图片报》	Bild am Sonntag; *Picture on Sunday*	DE	73
德国《法兰克福汇报》	Frankfurter Allgemeine Zeitung; *Frankfurt General Newspaper* (Germany)	DE	69
德国《经济周刊》	Wirtschaftswoche; *Economy Weekly*	DE	63
西德《法兰克福汇报》	Frankfurter Allgemeine Zeitung; *Frankfurt General Newspaper* (West Germany)	DE	63
德国《时代》	Die Zeit; *The Time*	DE	64
西德《世界报》	Die Welt; *The World* (West Germany)	DE	68
德国《世界报》	Die Welt; *The World* (Germany)	DE	72
美国《新闻周刊》	*Newsweek*	EN	61
美国《纽约时报》	*The New York Times*	EN	66
美国《读者文摘》	*Readers Digest*	EN	44
美国《基督教科学箴言报》	*The Christian Science Monitor*	EN	75
美国《商业日报》	*Investor's Business Daily*	EN	79
美国《商业周刊》	*Bloomberg Businessweek*	EN	65
美国《华盛顿邮报》	*The Washington Post*	EN	70
美国《华尔街日报》	*The Wall Street Journal*	EN	63
美国《时代》	*Time*	EN	50
美国《洛杉矶时报》	*Los Angeles Times*	EN	64
西班牙《改革十六》	Cambio 16; *Change 16*	ES	83
墨西哥《至上报》	Excélsior; *Excelsior*	ES	242
阿根廷《画报》	El Graáfico; *The Graphic*	ES	84
阿根廷《民族报》	La Nación; *The Nation*	ES	109
阿根廷《号角报》	Clarín; *Bugle*	ES	77
西班牙《趣味》	Muy Interesante; *Very Interesting*	ES	79
西班牙《论坛》	La Tribuna; *The Forum*	ES	94
西班牙《终极日报》	Ya; *Already*	ES	48
法国《青年非洲》	Jeune Afrique; *Young Africa*	FR	63
法国《论坛报》	La Tribune; *The Gallery*	FR	74
法国《费加罗报》	Le Figaro; *The Figaro*	FR	73

Source (in Chinese)	Source (in original language; *in English*)	Language	# Articles
法国《新观察家》	Le Nouvel Observateur; *The New Observer*	FR	59
法国《快报》	L'Express; *The Express*	FR	58
法国《巴黎竞赛画报》	Paris Match; *Paris Match*	FR	54
法国《回声报》	Les échos; *The Echoes*	FR	76
法国《周末三日》	We Demain; *Weekend Tomorrow*	FR	60
法国《世界报》	Le Monde; *The World*	FR	73
法国《解放报》	Libération; *Liberation*	FR	70
《日本工业新闻》	*The Nikkan Kogyo Shimbun*	JP	68
《日本经济新闻》	*The Nihon Keizai Shimbun*	JP	70
日本《东京新闻》	*Tokyo Shimbun*	JP	76
日本《日刊工业新闻》	*The Nikkan Kogyo Shimbun*	JP	74
日本《日经产业新闻》	*Nikkei Sangyo Shimbun*	JP	70
日本《朝日新闻》	*The Asashi Shimbun*	JP	72
日本《读卖新闻》	*Yomiuri Shimbun*	JP	74
日本《产经新闻》	*Sankei Shimbun*	JP	75
日本《每日新闻》	*Mainichi Shimbun*	JP	76
俄罗斯《共青团真理报》	Комсомольская правда; *Komsomol Truth*	RU	71
俄罗斯《劳动报》	Труд; *Labor*	RU	80
俄罗斯《旅伴》	Путешественник; *Traveler*	RU	63
苏联《消息报》	Известия; *Izvestia* (USSR)	RU	41
俄罗斯《独立报》	Независимая газета; *Independent Newspaper*	RU	65
俄罗斯《真理报》	Правда; *Truth*	RU	57
俄罗斯《红星报》	КраснаяЗвезда; *Red Star*	RU	67
俄罗斯《消息报》	Известия; *Izvestia* (Russia)	RU	75
《俄罗斯报》	Российскаягазета; *Russian Gazette*	RU	75

B Summary of features

Feature	Operationalization
Explicitation	
Explicit naming	Personal pronouns/proper nouns × 3
Single naming	Raw counts of single-token proper nouns

Feature	Operationalization
Mean multiple naming	Average num tokens of proper nouns × 3
Cohesive markers	Normalized counts from Chen’s (2006) list
Simplification	
TTR1	$V/N \times 6$, where $V =$ # types, $N =$ # tokens
TTR2	$\log V/\log N \times 6$
TTR1 and 2: char-based	TTR with characters as basic units
Mean word length	Mean word length in characters
Lexical density	# content words/# tokens
Mean sentence length	Mean sentence length in words
Mean character/word rank	Mean ranks of characters/words in a frequency list
5 most frequent chars	freq(word)/# tokens for each of the 5 chars
Mean tree depth	Mean of parse tree depths
complex NP/Cl	Number of complex nouns per clause
VP/Cl	Number of verb phrases per clause
RC/Cl	Number of relative clauses per clause
CFG rule *entropy*	Entropy of context-free grammar rules
CFG rule *types*	# types of context-free grammar rules
CFG rule *entropy* + *types*	Above two features combined
Normalization	
Repetitions	# content words that occur > 1/# tokens
Repetitions (only nouns)	# nouns that occur > 1/# tokens
Average PMI	Average PMI (Church and Hanks 1990) of all bigrams
Aspect markers	Normalized frequency of aspect markers
Measure words	Normalized frequency of measure words
Sentence-final particles	Normalized frequency of sent-final particles
4-character idioms: *Chengyu*	0: count of each idiom; 1: total counts
ba-structure	# of *ba*-structure/# sents
bei-structure	# of *bei*-structure/# sents
Interference	
Char *n*-gram	Character unigram
Word *n*-gram	Word unigram
POS *n*-gram	POS trigram
Context-free grammar rules	Normalized counts of rules, for example, S → NP VP

C `Tregex` patterns

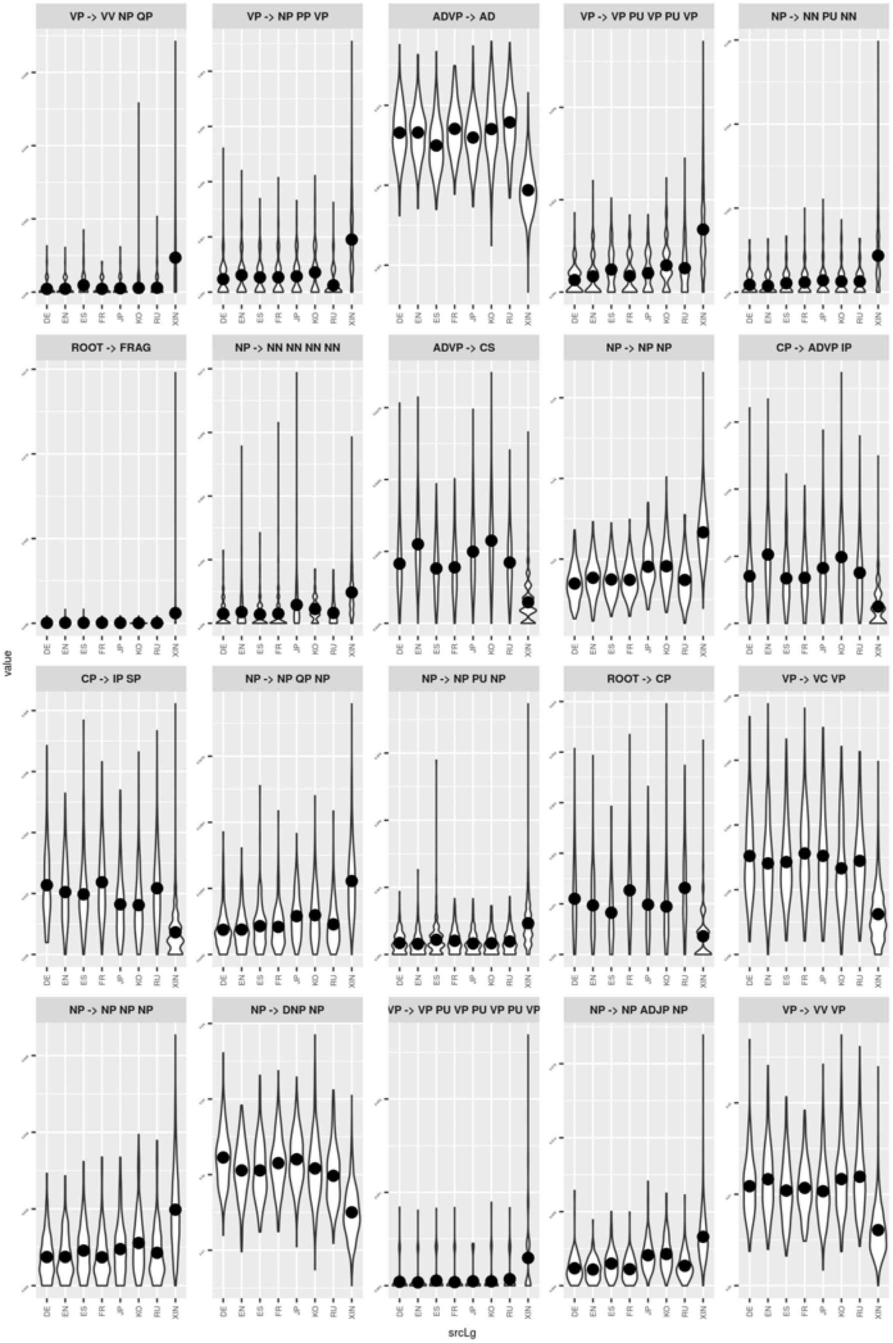
- Complex NP: `NP [ <<JJ|QP|ADJP|CLP|DNP|DP|<< (NP $++ NP !$+ CC) ]`
- VP: `VP > IP !< VC`
- Relative clause: `CP>(NP|QP|DP|PRN|LCP)<(DEC|DEG)`
- “one” + classifier + DNP/CP + NP: `NP <1 (QP<1(CD<-/[1——]/)) <2 (DNP|CP) <3 NP`
- all QP + DNP/CP + NP: `NP <1 QP <2 (DNP|CP) <3 NP`

D Most frequent idioms (Chengyu)

Rank	freq	Originals	freq	Translations
1	145	艰苦奋斗 *work hard*	248	引人注目 *eye-catching*
2	93	实事求是 *seek truth from facts*	85	成千上万 *thousands of*
3	90	千方百计 *leave no stone unturned*	79	无论如何 *anyways/no matter what*
4	86	东道主 *host*	77	众所周知 *as is known to all*
5	72	扭亏增盈 *reduce loss and increase profits*	70	有朝一日*some day*
6	70	源源不断 *endless*	63	雄心勃勃 *ambitious*
7	70	络绎不绝 *endless*	63	前所未有 *unprecedented*
8	63	引人注目 *eye-catching*	60	自给自足 *self-reliant*
9	63	坚定不移 *unwavering*	54	难以置信 *unbelievable*
10	62	全心全意 *whole-heartedly*	53	无家可归 *homeless*
11	57	见义勇为 *to be the good Samaritan*	50	停滞不前 *stagnating*
12	57	千家万户 *thousands of households*	46	理所当然 *naturally*
13	52	不正之风 *bad practices*	45	微不足道 *a drop in the bucket*
14	51	丰富多彩 *colorful*	43	显而易见 *obvious*
15	51	前所未有 *unprecedented*	40	轻而易举 *extremely easy*
16	50	因地制宜 *tailor to local needs*	39	消息灵通 *well-informed*
17	49	脱颖而出 *stand out*	39	名副其实 *be worthy of the name*
18	49	供不应求 *in short supply*	38	司空见惯 *be accustomed to*
19	47	一年一度 *annual*	38	不可思议 *unbelievable*
20	46	无家可归 *homeless*	38	犹豫不决 *indecisive*

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press

E Top 20 syntax features



Cite this article: Hu H and Kübler S (2021). Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering* **27**, 339–372. https://doi.org/10.1017/S1351324920000182

https://doi.org/10.1017/S1351324920000182 Published online by Cambridge University Press