Linguistic Fingerprints on Translation's Lens

J.D. Porter, Yulia Ilchuk, Quinn Dombrowski

Journal of Data Mining and Digital Humanities

Linguistic Fingerprints on Translation's Lens

J.D. Porter[1], Yulia Ilchuk[2], Quinn Dombrowski[2*]
1 University of Pennsylvania, USA
2 Stanford University, USA

*Corresponding author: Quinn Dombrowski qad@stanford.edu

Abstract
What happens to the language fingerprints of a work when it is translated into another language? While translation studies has often prioritized concepts of equivalence (of form and function), and of textual function, digital humanities methodologies can provide a new analytical lens onto ways that stylistic traces of a text's source language can persist in a translated text.

This paper presents initial findings of a project undertaken by the Stanford Literary Lab, which has identified distinctive grammatical features in short stories that have been translated into English. While the phenomenon of "translationese" has been well established particularly in corpus translation studies, we argue that digital humanities methods can be valuable for identifying specific traits for a vision of a world atlas of literary style.

keywords
digital humanities, translation studies, syntax

INTRODUCTION

A central question in translation studies over the past two hundred years is whether there is anything German about German when we read it in English. This question preoccupied Friedrich Schleiermacher, who in his 1813 essay "On the Different Methods of Translation" argues that a good translator of literature must "keep the tone of the language foreign," with the "aim that a foreign spirit should blow towards the reader" (79-80). What's more, this is not just some vague "foreignness", no mere sense that the text is not from around here: The experienced reader ought to be able to "distinguish between works of Greek and Roman origin, or Italian and Spanish" (80). In other words, if our copy of his essay is any good, it ought to be recognizably German even if we read it in English. As Schleiermacher has it, a well-translated library should give readers something like a world atlas of literary style and linguistic possibility.

Schleiermacher argues that translation *should* follow this model, and even today theories often treat the question in rather ethical terms. Lawrence Venuti, for instance, suggests that a Schleiermachian concept of "foreignization" can help a translator resist ethnocentrism, racism, and imperialism (16). We leave these questions aside, to focus not on what translators *ought* to do, but on what they have, in fact, done. Foreignization as an inevitable outcome of translation has been the focus of recent corpus translation studies. Investigations of "translationese," per Gellerstam and others, show that it is often possible to tell whether a text originated in its present language or was translated, using only such features as the

frequencies of words, parts of speech, and syntactic structures. It seems that in many or, perhaps, most translated works of literature, computational methods can now detect the foreign spirit.

Though this presence of some general foreignness in translations is well established, the world atlas of linguistic style that Schleiermacher envisioned has yet to be produced. Until recently such a thing scarcely seemed possible. André Lefevere argued in the 1970's that the "prospect of the application of his method on a massive scale, so that all different shades of foreignness might be represented in the translations, is… clearly utopian" (67). Schleiermacher himself seems to have had a similar idea, arguing that a reader's ability to detect specific language origins will "only be possible if he is able to make massive comparisons," and that barring this "even the most educated readers can achieve only a very deficient knowledge of what is foreign" (80). Anyone familiar with debates in the Digital Humanities about scale, comparisons, and individual readings can see that Schleiermacher might as well have had "distant reading" somewhere in his subtitle. To arrive at a comprehensive account of every feature of every language rendered in every other language probably remains impossible at present, but we now have the computational tools and the digital corpora to begin the task.

In this essay we take a step towards that vision with a specific focus on literary criticism. Beginning with a brief account of the current state of corpus translation studies, we then turn to a small experiment of our own, demonstrating what may need to happen for English or comparative literature departments to get in on the action. We show how information about the traces of particular languages across the barrier of translation can serve as a "textual grid" (to borrow a phrase from Bassnet and Lefevere) that both clarifies and illuminates the ideas of global reading. Finally, we propose a method by which translators (or even authors of texts that have not been translated) could calibrate their use of the stylistic features indicative of translation from a particular language to achieve the desired effect on the reader. In short, we aim to shed a little new light on the affordances of translation[1].

I A BRIEF SURVEY OF DIGITAL APPROACHES TO TRANSLATION STUDIES

Compared with many other areas of contemporary DH, the use of quantitative methods to determine whether a text has been translated has a long history. In the 1986 essay that first popularized the term "translationese", Martin Gellerstam examines word frequencies in novels written in Swedish (for clarity's sake, we will often use "L1" to refer to a text's original language) versus those translated into Swedish from English (Swedish L2). By the early 1990s, scholars working in corpus translation studies had theorized a number of detectable "translation universals," or features that tend to be present in a translated text irrespective of the particular languages involved. Many of these features revolve around clarity, including, to use the list Mona Baker compiled in 1993, "explicitation" (i.e., preferring explicit options even where ambiguous ones are available), simplification, conventional grammaticality, and avoidance of repetition. Baker also observes that translation might exaggerate certain characteristic features of the source language (243-245)[2]. The upshot is that translated texts contain the information

[1] We use the term "affordances" in the sense suggested by Levine.

[2] The notion of "translation universals" is related to the unique language emerging at the meeting point of the source and target texts. An important characteristic of the notion of translation universals is that they are independent of the languages involved, although their expression is necessarily related to the source and target languages. Like a transfer of grammatical structures, the translation of lexical items is also governed by a

necessary to determine that they are translated. Whatever the translators may have intended, statistical comparisons with untranslated work in a similar genre reveal foreignization at work.

More specific applications of these principles are relatively easy to come by. Sessions of Canada's parliament, for example, contain English and French in both L1 and L2 guises; Kurokawa, Goutte, and Isabelle detect whether given transcripts are translated—their program guesses correctly 85% of the time when using parts of speech (POS), and 90% when using word bigrams[3]. Santos applies Gellerstam's translationese concept in a Portuguese context, finding that, although the quality of results depends on translation quality and language proximity (e.g., Portuguese is genealogically closer to Spanish than to Russian), it is still possible to identify translatedness using grammatical features. Xiao and Yue find consistent differences between Chinese L1 texts and Chinese L2 texts across a number of features, such as sentence length and various markers of explicitation. And among the many projects focusing on texts translated into English, Olohan has demonstrated particular features that go a long way toward producing a foreignization effect, including lower frequencies of contractions and higher frequencies of the word "that" used as a complementizer. We have zeroed in on these examples mainly to highlight a diversity of genres and languages, but the list of experiments like these is long and growing.

Notwithstanding these advances, the Schleiermachean vision of a translated library that retains linguistic specificity remains a bit utopian. In the first place, experiments that identify language-specific markers of translation are, obviously, less numerous than those which identify translatedness per se, or what we might call general foreignness. Of course, these can be deduced by reading, say, Xiao and Yue for Chinese, Santos for Portuguese, etc., but the point is that the work remains somewhat piecemeal. The application of these methods to literary analysis in particular is also rarer. When we began this experiment a few years ago, we were able to find just one paper that did what we had in mind: Lynch and Vogel successfully identified the source languages of novels translated into English from Russian, German, and French. Their findings are quite suggestive, but their experiment is limited by their tiny corpus—just five novels for each source language.

More important, we believe that literary criticism has not yet taken full advantage of the new technical findings as a context for and aid to reading and analysis. For anyone working with literature in translation, this kind of work provides invaluable context about why the form of a text is the way it is. The application of digital research methods in translation studies can generate new insights for a broad range of research questions: what is the role of the author's or translator's "fingerprint" in transferring the meaning from the source to the target language? Does the style of the original text impose constraints on the translator? Are there linguistic or textual features associated with each language that affect the way that language's texts work in English? Whichever view one adopts in these discussions, it is clear that translation leaves a stylistic trace on a text. Of course, literary scholars, especially within the comparatist tradition, have long thought through the effects of translation, and accordingly we want to proceed with caution and humility about the novelty of our methods here. Our hope is simply that they may prove a useful example of bringing computational corpus translation studies into the literary critical realm.

---

need to retain aspects of the corresponding source invariants. This causes alien forms to appear in the target language, be it structures that are never encountered in the target language, or lexical items which belong to the cognates and foreign-language loans. For an additional study of the universals proposed by Baker, see Granger 2005.

[3] Even when examining single sentences, their method guesses correctly 77% of the time.

II CORPUS BUILDING

The greatest technical obstacle to literary critical engagement with corpus translation studies is the corpus itself. The development of a standard, widely used corpus of translated texts remains a topic of contention and aspiration even among translation studies specialists[4]. For the sorts of questions we are considering here, we need as many texts as we can get from multiple languages, authors, and translators, and we need them to be roughly comparable across all of those vectors. That is, if we have 100 well-translated 19th-century novels written in French, we need something like 100 well-translated 19th-century novels written in Russian. The difficulties are immediately apparent—for English readers, *are* there 100 well-translated 19th century novels written in Russian?

Two other standard practices of literary DH raise the bar even higher, to the disadvantage of some existing corpora of translated texts. First, there is the usual attention given to the author signal. An author's works tend to be very statistically similar to each other across a number of features, including many commonly used in translation studies (especially word frequencies). For researchers studying some other phenomenon—say, the statistical features associated with being Portuguese L1—care must be taken to mitigate the author signal, often by limiting the corpus to one text per author or by having enough texts to drown out any one author[5]. Second, a similar, if far less pronounced, phenomenon may apply in rare cases to translators. Rybicki shows that translators do sometimes leave faint, detectable traces on the texts they translate[6]. Ideally, then, a corpus would contain only one text per author and one text per translator. However, within a specific time period, there tends to be a limited number of active translators for a given pair of languages. Permitting only one text per translator quickly narrows the pool of possible texts.

Many commonly used translation corpora are highly limited if researchers want to control for these author and translator signals. Consider the well-known Translated English Corpus (TEC; Olohan uses it in her 2001 and 2003 work). The TEC fiction corpus contains dozens of short stories. Yet within particular languages, the corpora have substantial author and translator repetition. The metadata for the Portuguese corpus, for instance, shows just nine text files; four of these are by José Saramago, and six are translated by Giovanni Pontiero. For

[4] Cf. Ustaszewski.

[5] Note, however, that the choice of style in translation is not entirely motivated by the source text, because the stylistic features in the source text do not always reflect the author's conscious choices and because the author's intention is not easily discovered for the reader. The "authors' statements and suggestions could be seen as simply another source of information, like dictionaries, background reading, or the opinions of other readers" (Boase-Beier, 50). Therefore, a translator does not reconstruct what is inferred as the author's intention; their inference derives from evidence in the style of the text. In translation terms, style is not just a choice of words or syntactic structures; stylistic choices "reflect a speaker's (subjective) choice of a given conceptualization" (Tabakowska 1993:7), and thus are always a reflection of different content rather than just different expression (Leech & Short 1981:15ff.; see also Pilkington 1996:159).

[6] Rybicki actually emphasizes the near invisibility of translators, "an unexpected corroboration of Venuti's observation on translator's invisibility. Indeed, it is adding insult to injury". In Rybicki's tests, the author signal nearly always appears to overpower the translator signal. However, he does show at least one case where a particularly free translator leaves a trace (W.S. Kuniczak's translations of Henryk Sienkiewicz). Would this be true of someone like the prolific translator of Russian, Constance Garnett? (Joseph Brodsky supposedly once complained that "The reason English-speaking readers can barely tell the difference between Tolstoy and Dostoevsky is that they aren't reading the prose of either one. They're reading Constance Garnett." (quoted in Remnick)). We prefer to err on the side of eliminating these "translator signals".

a researcher who seeks to avoid duplicate authors and translators, there are only four usable files in total. As Olohan's work shows, the TEC has proven its quality and usefulness in other respects; the issue is simply that this is a difficult problem to tackle. Consider the example of the Russian novels again—if we allow just one appearance per author and just one Constance Garnett translation, the total stack of public domain English titles might well fit in a single backpack—and the student wearing it could probably read them all in a semester.

For these reasons, we decided to build our own corpus for this experiment. We collected short stories from six different language origins as well as a comparison corpus of L1 English stories, as detailed in Table 1[7]. For the purposes of this experiment, we did not need the texts of the stories in their original languages; we only have them in English. They come from sources like the *New Yorker*, a few specialized anthologies, and the *Akashic Noir* series, which compiles detective stories set in single cities, e.g. *Buenos Aires Noir,* that are typically translated into English from a language spoken in that city. The *Akashic* stories are useful in that they are all new, meaning the dates for originals and translations are easy to find, but they do tilt our corpus toward detective fiction. Across the corpus as a whole, the authors range from well-known, high-prestige writers (Colson Whitehead, Isabel Allende, Clarice Lispector) to fairly obscure (or, in some cases, early career) writers. Each author and translator appears only once.

Original Language	Number of stories	Total tokens
English	20	106,963
Indonesian	12	41,561
Italian	20	73,314
Korean	20	165,375
Portuguese	20	74,008
Russian	20	160,686
Spanish	20	65,714

Table 1. The corpora used in this study.

This corpus is still far from ideal. Some language corpora proved more difficult to put together than others; we cannot be certain that they are all equivalent when it comes to quality of translation, and some corpora have much longer stories, on average, than others. It also proved difficult to gather stories from a tight date range: our English corpus is more recent than the others, dating mostly from the 2010s; the Portuguese corpus is the oldest, with some texts ranging as far back as the turn of the 20th century. Though most texts are from the past 50 years—by the standards of literary periodization, it is fair to describe this as a mostly contemporary corpus, with a few outliers—the results should be understood in the context of this suboptimal date spread[8]. Still, this corpus should prove to be useful for computational translation studies, particularly given the non-repetition of authors and translators.

III FORMS OF DISTINCTION

In simple terms, we want to find features of our texts that distinguish them on the basis of their original language. The most obvious place to look is at the words. Unfortunately, if we let every word in, they distinguish the texts so well and in such obvious form that the results

[7] Antonio Lenzo, Shana Hadi, and Eunji Lee were instrumental in building the corpus. Initially, we collected some stories in German and French as well, but since many of these were substantially older, we have excluded them from the analysis reported here.

[8] When we could not find dates of initial publication, we gathered author birth and death dates.

are difficult to use. For instance, we ran a Most Distinctive Words (MDW) analysis of our corpus, looking for words that show up statistically significantly more often in each corpus than they do in the English L1 corpus[9]. Some of the results are thought-provoking, if potentially side effects of the fairly small corpus. For instance, why do Spanish, Portuguese, and Russian (but not Korean, Italian, or Indonesian) L1 texts use "love" more often than English? Yet more often, we find things like the highly distinctive uses of "petersburg" and "moscow" in Russian L1 stories, of "euros" in Italian L1 and "won" in Korean L1, and even of "indonesian" in stories originally written in Indonesian. Analysis of this kind seems to lead us fruitlessly into sociology, elevating differences of cultural context rather than linguistic style. Although culture and language are certainly linked, our interest lies less in the "aboutness" captured by this capacious semantic analysis, and more in a kind of style that can be traced to language, no matter the currency characters use to buy biscotti or bibimbap. In other words, we are interested in the detectable linguistic traces of Russian that withstand translation; we are less interested in discovering that stories written in Russian are more apt to mention "petersburg".

To avoid the petersburg problem, we turned away from words as a whole, in two ways. First, we examined the grammar of the texts, tagging them for parts of speech (POS) as well as dependencies[10]. Once again we checked these features for distinctiveness. Second, we examined stop words. Also known as function words, they are sometimes described as the alternative to "content words": these are the articles, prepositions, pronouns, conjunctions, and other workhorse words that sit at the top of most frequency lists for English texts—in our corpus, roughly one out of every 18 words is "the". We counted the frequency of every word included in the list of stop words that comes with the Natural Language Toolkit Python library and once again looked for distinctiveness by L1 corpus. To be clear, this filter still leaves a fairly high percentage of each text intact; about 49% of all words in our corpus are stop words[11].

In general, we believe that stop words are undertheorized in literary DH. They clearly bear some sort of important relationship to style. For instance, they are extremely useful in authorship attribution, and it is often possible to make reasonably good assessments of authorship using only stop words. For this to be true, they have to be doing a lot of formal work. Understanding the formal or aesthetic role of these words is quite difficult; most of the time they are virtually invisible, and it would be no small feat to construct a general theory of the aesthetics of words that seem to carry zero semantic content. Yet, in combination with grammatical features, we have found them quite useful for examining the affordances of foreignization.

III READING THE RESULTS

[9] We calculate MDW by getting word scaled frequencies for each corpus, assigning an expected rate of scaled frequency based on the frequency of each word in the English L1 corpus, and using a Fisher's exact test to determine the significance of the difference between the observed and expected frequencies. For example, if the English L1 corpus has a word count of 100,000 and says "furthermore" 20 times, the expected rate of "furthermore" for all corpora is .0002. If the Portuguese L1 corpus says "furthermore" 35 times in 85,000 words, then our test would be (using the Python scipy.stats format): fisher_exact([[a,b],[c,d]]) where a = 35, b = 84,965 (i.e. the number of times "furthermore" was *not* observed), c = 17 (or .0002 * 85,000) and d = 84,983 (the number of times "furthermore" was expected not to appear). The test returns a p-value, which describes the likelihood that the observed difference is attributable to chance. In this example, "furthermore" would appear statistically significantly more often in Portuguese, with p = .009.

[10] We used the Spacy library for Python for both part of speech and dependency tagging.

[11] Per corpora, the percent stop words ranges from 47.6% in English L1 to 50.7% in Spanish L1.

In this section we focus on one finding, a confirmation of existing translation studies work, though it is an observation so small that an ordinary reader might miss it in the course of reading a story. We noticed early on that the word "that" is more common in all of the translated texts than it is in the English L1 texts (Table 2). This fact poses a substantial interpretive challenge, not just because of the usual ineffability of stop words, but because *that* is particularly multifunctional: It can be an adverb ("not *that* bad"), a determiner ("toss me *that* bag"), or a complementizer ("I see *that* your bag is bad"). The POS tags and dependency parses helped us to clarify the situation (Table 3): The word "that" is more common in the translated stories across all of these uses, but the effect is especially pronounced when it comes to complementizers[12]. Narrowing the focus to this usage shows an even more pronounced effect across all of the L1 corpora.

Corpus	Times used	p-value	Observed/expected uses[13]
Spanish	946	7.28E-16	1.50
Indonesian	553	2.90E-07	1.39
Portuguese	957	8.90E-10	1.34
Italian	931	1.06E-08	1.32
Korean	2029	1.01E-13	1.28
Russian	1682	0.014	1.08

Table 2. Distinct uses of the word "that"

Corpus	Times used	p-value	Observed/expected uses
Spanish	397	3.39E-21	2.27
Indonesian	395	5.85E-17	2.03
Portuguese	226	3.90E-10	2.00
Italian	351	2.95E-10	1.71
Korean	777	5.43E-20	1.69
Russian	712	3.11E-15	1.59

Table 3. Distinct uses of the word "that" as a subordinating conjunction

Syntactic differences in L1 and L2 govern the variability of the translation of *that*. Unlike in many other languages, the use of *that* in English compound sentences is optional. It is optional if it is used to reproduce somebody else's speech (indirect speech); or one's own thoughts and beliefs (after verbs believe, think, know, hope, etc.); or in the constructions "to be+ adj. (sure, certain, right, important, afraid, pleased, sorry, surprised, worried)+ that" to express feelings, opinions, beliefs; or in the constructions "be+noun (belief, fact, hope, idea, possibility, suggestion, statement, claim, comment, argument)+that" to express somebody else's beliefs, opinions, attitudes.

If we look at Russian as a specific example of usage outside of English, the complementizer *that* is mandatory in all kinds of compound sentences. It is associated with the pragmatics of

[12] The information in this table is based on the uses of "that" tagged by Spacy as subordinating conjunctions and given the "mark" designation. Our hand tags showed that Spacy does not always correctly identify the uses we want here; nonetheless, it seems accurate far more often than not, and we thought it worth showing what a common automated tagger finds.

[13] This column shows how many times the word "that" appeared, compared to the number of times we'd have expected it to appear if it was used at the same rate as in the English L1 corpus. For example, the Spanish corpus uses the word "that" 1.5 times as often as English, if we adjust for word count.

communication and reflects the speaker's attitude to the topic of communication, listener, or the utterance itself. The obligatory use of *that* makes Russian speech more emotional, emphatic, but also more structured. In the use of *that* to reproduce indirect speech ("he said that he won't come home"), the grammatical subject "he" can be omitted («он сказал, что не придет домой») because "that" ("что") can substitute it functioning as a grammatical subject. The retention of *that* in discursive terms also helps eliminate an ambiguity in Russian, that the double bind creates. In Russian, a sentence like "Peter said he would marry her" is ambiguous, as it remains unclear who "he" refers to – Peter or somebody else. We believe that the complementizer *that* partially clarifies the context.

In other words, the findings here replicate a well-known example of explicitation; Olohan's 2001 work shows the same overrepresentation of complementizer "that" in translated texts (425-426) This usage indicates a widespread translator preference for the more explicit option in a grammatical scenario that allows (and, at times, might even call) for a more ambiguous choice. In English, optional nature of the complementizer "that" means that the sentences "I know *that* the sky is blue" and "I know the sky is blue" are equally valid. As we have seen with Russian, however, many languages treat the complementizer as requisite. In Spanish, for instance, the word "que" must be included in the equivalent sentence, "Se *que* el cielo es azul." Translators have the option of eliminating the *que* when it turns into a *that*, and sometimes doing so may produce more natural or colloquial sentences, but it seems that they generally keep the word around. The reasons are not clear. It may be that, seeing the "que", they are more apt to reach for a one-to-one lexical translation, or maybe they just prefer to choose the clearest possible option (this is after all the theory behind explicitation). Schleiermacher might argue that retaining the complementizer is a subtle way to mimic the feel or a small taste of the *weltanschauung* of the original speaker's linguistic context.

When it comes to the text, however, there is less need for speculation. Whatever their origin story, the more frequent presence of the complementizer "that" in these translated texts is simply the fact of the matter. What is more, the uses of this stop word point clearly to a specific conceptual space. Following a hunch that developed over the course of creating many different sample sentences to explain what the complementizer form of *that* looks like, we set out to determine which verbs tend to govern its use. Given the complexity of the phenomenon, the best way to do this seemed to be hand-tagging—that is, reading the sentences and writing down what we thought was going on. We isolated the roughly 5,000 sentences in our corpus that, according to a Spacy parse, contained approximately the correct usage of "that", and hand-tagged six hundred of them, randomly selected[14]. We then identified the lemma form of the governing verb and, following our hunch, classified it as a verb of communication (like "say" or "tell"), cognition (like "know" or "realize"), perception (like "see" or "feel") or something else[15]. Of our 600 sentences, 447 turned out to have the kind of usage of "that" we wanted (many had other constructions, like "so angry that he choked" or "it is true that", and either slipped through our filters or were mistagged by Spacy). Table 4 shows the breakdown of the verbs according to our categories[16].

[14] We were still using our French corpus at this stage of the experiment, even though we removed it in the rest of the experiments detailed in this paper. However, since the purpose of this tagging was simply to gather many examples of the complementizer "that" in action, we believe that the results are still worth reporting.
[15] A lemma is a standardized form of a word that can appear in multiple forms, e.g. "say" for "said", "saying", "says", and so on.
[16] The count in Table 4 refers to total (lemmatized) token appearances within a given category, rather than unique types (that is, a lemma like "say" is counted multiple times for the "communication" category, since it appears in many sentences).

Category	Count	Percent of all
Communication	161	36%
Cognition	187	42%
Perception	75	17%
Other	24	5%

Table 4. What sorts of verbs govern the complementizer "that"?

These are, overwhelmingly, verbs that have to do with the relationship between an agent and some situation, whether reported, pondered, or sensed. In other words, they are verbs designed to treat propositional content; it is tempting to say that they represent facts (or at least, potential facts). The distinction between a fact and the ordinary objects of apparently transitive verbs is a bit slippery, but it is more or less the difference between "a red apple" and "that the apple is red". The former is clearly not a fact so much as an object, albeit a complex one that contains multiple words and even parts of speech. The latter is a fact, and can only be subordinated to verbs that can handle facts. You can eat or throw "a red apple" but cannot eat "that the apple is red"—you can say, think, or see it.

Of course, this is really a way of restating the meaning of the relevant linguistic categories. A complementizer is designed to subordinate a clause, and a clause is a proposition, in the sense that it must have a verb and a noun. We belabor the distinction only because it appears to raise some thorny issues about the objects of belief, knowledge, discussion, and so on. As Quine says, in order to name *what* we believe:

> … we enlist a sentence as a subordinate clause. For example, we speak of the belief *that* Hannibal crossed the Alps, and *that* Neptune is a planet. We use a sentence, with "that" prefixed, as a name of the "thing" believed. Now what manner of thing is this believe thing—*that* Hannibal crossed the Alps? (4, emphasis in original)

For Quine, the answer is basically "don't worry about it"[17]; he thinks it is fine to consider sentences to be the basic units of belief. For our purposes, the upshot is that this seemingly miniscule side effect of different linguistic requirements—the presence or absence of a word that would be filtered out of many DH experiments—gets us into some weighty territory. If the contents of thought and communication are generated by prefixing the word 'that' to a subordinate sentence, then the empirical observation that these subordinate sentences are handled consistently and significantly differently in translated texts means that foreignization colors the treatment of a basic feature not just of language but of existence, propositions, or facts in general. If we took Quine literally, we would have to argue that translated Italian L1 texts, by prefixing the word 'that' to more subordinate sentences, have notably more "contents of belief" than English L1 texts.

This obviously goes too far. We doubt that Quine would place as much importance on the "that" as on the subordinated sentence. Still, the fact that "subordinated sentences" are more clearly demarcated, separated from the subject and verb by an extra word—four additional letters and a space—is a good example of the kind of background information that we believe

[17] He does not literally say "don't worry about it", although he comes close: "This, like various other philosophical questions, is better deflected than met head on" (4). Also apt is his observation that, "Foreign speakers, after all, are said to share the belief that Hannibal crossed the Alps, even when they do not understand the English sentence" (4).

this kind of corpus translation studies can offer literary critics. A scholar ought to go into a reading of a story translated from Spanish with knowledge of the differences attributable to the fact that it is translated from Spanish. They form the textual grid over which a reading can be performed.

As a small example, consider the story "Sophie and the Angel", written by Dora Alonso and translated by Beatriz Teleki. Sophie is a pious and lonely Catholic octagenarian who starts to receive regular visits from a beautiful angel who plays electric guitar and canasta, converses with her over lemonade, and gives her a unicorn. The erotic tension of these visits eventually grows unbearable for Sophie, and on the advice of the local priests, she apparently gets the angel to leave her alone. Alonso makes the diegetic reality of the angel ambiguous: the priests view the whole thing as alarming and unhealthy "illusions," but a twist ending describes "a long white feather of unquestionably angelic origin poking out luridly—and inexcusably—from beneath the sheets of her bed" (22). As for Sophie, "She often thought that she was living a delightful dream" (21). Or, as we have just primed ourselves to see it, she often thought a subordinate sentence beginning with the prefix "that."

If we were naive close readers, we might describe the inclusion of the optional "that" as a subtle tactic Alonso deploys to underline the propositional nature of "she was living a delightful dream," to establish that Sophie is engaging with the hypothetical nature of what is going on, as if Sophie is saying this sentence to herself. In our view, there is a slightly different sense (or, if you prefer, feel) to the phrasing "She often thought she was living a delightful dream", a bit of an elision of the space between the agent and the proposition, as though Alonso is giving us the subordinate sentence, but Sophie is experiencing its contents. It is admittedly a very small distinction, like an accent, the difference between pronouncing it "She often thought *that* she was living a delightful dream" and "She often *thought* she was living a delightful dream". Grammatically speaking, these are the same, but that does not mean readers do, or have to, experience them the same way; the existence of the option makes possible a separation. So, if Alonso includes the "that", perhaps she intends to place just a bit of distance between Sophie and the dream, "that" as a form of metacognition.

In the context of our corpus translation results, however, we have good reason to believe that Alonso did not make this choice, that in a practical sense she probably *could* not make this choice. Indeed, the sentence in the original Spanish is "En muchos momentos ella pensaba *que* vivía un delicioso sueño" (182, emphasis added). It is just as expected: Alonso includes the "que", because in Spanish a sentence like this must include it. The person who made a decision in the English text is Teleki, the translator. She would have been perfectly justified in leaving out the "that"; it might even have been a nice match for the simple, witty, light style of the story. The critic is also perfectly justified in observing the effect of the "that" in the story; everything we argued in the previous paragraph is true with respect to the English text that we have. The critic just needs to understand that the "that" is present because, in some order, 1) translated texts in general have more explicitation and specifically more complementizer "that" instances than English L1 texts, and 2) Spanish in particular requires the presence of the complementizer in sentences like these, and 3) Teleki made a choice (conscious or not) to include it. The subtle emphasis on severing agent from proposition, the little extra light on the sentential nature of objects of thought and speech, is an effect in this story, but it is also part of the textual grid of translation in general and translation from Spanish in particular[18]. These factors contextualize without erasing the formal features of the

[18] It is probably saying too much to suggest that this sort of effect has an influence on the English-speaking-world's relationship to Latin American magical realism (the anthology containing the Alonso and several other

English text. A slight extra sharpness to the contents of thinking, perceiving, and saying may be one of the affordances of translation.

IV APPLICATIONS

The results from this initial study are suggestive of the potential for translators and writers to evaluate and shape the use of translationese features in their text. With support from digital tools, backed by data from larger and more nuanced corpora, translationese features could potentially take on a more significant role in shaping literary style, both for new translations and in original text meant to invoke a sense of foreign-ness.

For computationally-oriented digital humanists, there are already relatively easy-to-use tools for using the findings described here as benchmarks for comparing new text. As a proof-of-concept, we developed a Jupyter notebook (an interface for juxtaposing executable code with human-readable text) that takes an English-language text as the input, and compares its use of 'that' to the results we identified for Italian, Portuguese, Russian, and Spanish. The resulting reports reveal traits of the text that it may be difficult for the translator or writer to see, and provide suggestions for how to increase or decrease the score for each linguistic feature. The writer or translator could, by adjusting the text and re-running the Jupyter notebook for updated scores, achieve a fine-grained level of control over the extent of translationese manifested within the text.

While Jupyter Notebooks may be a familiar, relatively "easy" environment for DH scholars who are used to working with code, they are still far outside the mainstream tool set for writers and translators. That said, both of these groups have increasingly welcomed computationally-driven tools into their workflows over the past five years, but for such tools to be successful, they must be incorporated into word processing environments. Companies such as Grammarly have built upon advances in machine learning and NLP to develop plugins for web browsers as well as Microsoft Office that offer suggestions around tone and style, in addition to spelling and grammar. PerfectIt is a similar tool offering "professional" proofreading, including locating undefined abbreviations and enforcing style manual conventions.

The large English-language data sets that serve as training data for existing commercial tools are relatively easy to obtain. To develop an effective tool for benchmarking translationese, one would first need to amass corpora much larger than the ones described here. Ideally, those corpora would reflect a range of different genres and time periods, to capture, for instance, differences in style between 19th century crime fiction and 20th century magical realism. At the same time, more variation and nuance inevitably leads to smaller corpora, potentially to a point where the "benchmark" for a particular combination of genre and language is just the sum of the noise caused by a small set of peculiar exemplars. Nonetheless, developing the kind of corpora necessary to develop a tool usable for writers and translators to modulate their use of translationese features is increasingly feasible. Based in the United States, HathiTrust provides access to scanned, OCRed versions of millions of books, including a large number of literary texts in translation. Texts that are still protected by copyright may be accessible for computational analysis via the HathiTrust "data capsule". Outputs of this analysis are vetted to ensure that they honor agreements about what kinds of derivatives can be extracted from the text, but calculations similar to those described here (relating to frequency of syntactic

---

of our Spanish L1 stories is subtitled *The Magic and the Real*). It may just be a coincidence. Still, as many magical realists could tell you, coincidences count.

structures, rather than content) should be permissible. To expand beyond the strictly-literary, there are also large shared-interest communities online who post translations of other forms of cultural production, such as song lyrics and fan fiction. These could potentially serve as another source of data for translationese features and frequency when developing a tool for writers and translators.

Conclusion

In this paper we have laid out a very ambitious scheme for a world atlas of the features of translation as such, as well as decidedly modest examples of the kinds of findings such an atlas might contain, and the interpretive possibilities it may generate. The first goal is, of course, far beyond the scope of any one paper. It may even be impossible. Still, the advances in libraries, linguistics and NLP departments, and translation studies and DH labs have made the project seem a lot less utopian than it did when Lefevere levied that critique. We want to emphasize that the digital findings of this paper have largely replicated the work already done by specialists in the more technical side of this field (broadly conceived). Confirmation has value, but the point is that we have not attempted to overthrow any existing wisdom on that front. Rather, we believe that findings like these can be deployed in new and exciting ways. Translators may be able to take advantage of these findings to assert greater control over how their translation reflects the source language. Moreover, these findings may open the door to extensive work by literary scholars. Knowledge of the ambiguities, limitations, quirks, and felicities of translation will open the way to more sophisticated and accurate engagement with texts from a wider variety of language origins. Ideally, it will empower more scholars to work, responsibly, in languages they don't know, drawing the world's scholarly literary language communities closer together. Those literary critics who have ideas about texts in languages where they lack fluency may find in this set of methods a way to justify and structure engaging those texts, to the benefit of all. Small findings and readings like those we have produced here may eventually add up to this utopian goal. The massive number of comparisons that Schleiermacher imagined may well tear down the distinction between we subjects and the delightful dream.

References

Alonso, D. "Sofía y el Ángel". *Cuentos Cubanos del Siglo XX*. Ed. Luis Rafael. Madrid: Editorial Verbum, 2019: 179-184.

Alonso, D. "Sophie and the Angel". Trans. Beatriz Teleki. *Short Stories by Latin American Women: The Magic and the Real*. Ed. Celia Correas de Zapata. Houston: Arte Publico Press, 1990: 18-22.

Baker, M. "Corpus linguistics and translation studies: Implications and applications." *Text and Technology: In Honour of John Sinclair* 233 (1993): 233-50.

_____. "Towards a methodology for investigating the style of a literary translator." *Target. International Journal of Translation Studies* 12.2 (2000): 241-66.

Bassnett, S., and André Lefevere. "Where Are We In Translation Studies?" *Constructing Cultures: Essays on Literary Translation*. Vol. 11. Multilingual Matters, 1998: 1-11.

Boase-Beier, J. "Stylistics and translation." *The Routledge Handbook of Stylistics* (2014): 393-407.

Fawcett, P. "Presupposition and translation." *The Pragmatics of Translation*. Ed. Leo Hickey. Multilingual Matters LTD, 1998. 114-140.

Gellerstam, M. "Translationese in Swedish novels translated from English." *Translation Studies in Scandinavia* 1 (1986): 88-95.

Granger, S. "A lexical bundle approach to comparing languages: Stems in English and French." *Languages in Contrast* 14.1 (2014): 58-72.

Koller, W. *Einführung in die Übersetzungswissenschaft*. Heidelberg: Quelle & Meyer, 1979.

Kurokawa, D., Cyril Goutte, and Pierre Isabelle. "Automatic detection of translated text and its impact on machine translation." *Proceedings of MT-Summit XII* (2009): 81-88.

Leech, G., and H. Michael. "Short." *Style in Fiction* 235 (1981).

Lefevere, A. *Translating Literature: The German Tradition from Luther to Rosenzweig*. Assen: Van Gorcum, 1977.

Levine, C. *Forms: Whole, Rhythm, Hierarchy, Network*. Princeton University Press, 2017.

Lynch, G., and Carl Vogel. "Towards the Automatic Detection of the Source Language of a Literary Translation." *Proceedings of COLING 2012: Posters*. 2012.
Mikhailov, M., and Robert Cooper. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. Routledge, 2016.
Moretti, F. *Distant Reading*. New York: Verso Books, 2013.
Munday, J. "Using Systemic Functional Linguistics as an Aid to Translation between Spanish." *Revista Canaria de Estudios Ingleses* 40 (2000): 37-58.
Olohan, M. "How Frequent Are the Contractions? A Study of Contracted Forms in the Translational English Corpus." *Target* 15:1 (2003): 59-89.
————————. "Spelling Out the Optionals in Translation: A Corpus Study." *UCREL Technical Papers* 13 (2001): 423-432.
Pilkington, A. "Introduction: Relevance Theory and Literary Style." *Language and Literature* 5.3 (1996): 157-162.
Quine, W. (Willard Van Orman), and J. S. Ullian. *The Web of Belief*. New York: Random House, 1970.
Remnick, D. "The Translation Wars: How the Race to Translate Tolstoy and Dostoyevsky Continues to Spark Feuds, End Friendships, and Create Small Fortunes". *The New Yorker*. October 31, 2005. Online.
Rybicki, J. "The Great Mystery of the (Almost) Invisible Translator." *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*. 231 (2012). 231-48.
Santos, D., and Cristina Mota. "Experiments in Human-Computer Cooperation for the Semantic Annotation of Portuguese Corpora." *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010) (Valletta 17-23 May de 2010) European Language Resources Association*. European Language Resources Association, 2010.
Schleiermacher, F. "On the Different Methods of Translation." Ed. André Lefevere. *Translating Literature: The German Tradition from Luther to Rosenzweig*. Assen: Van Gorcum, 1977: 67-89.
Tabakowska, E. *Cognitive Linguistics and Poetics of Translation*. Vol. 9. Gunter Narr Verlag, 1993.
Ustaszewski, M. "Optimising the Europarl Corpus for Translation Studies with the EuroparlExtract Toolkit." *Perspectives* 27.1 (2019): 107-123.
Van Halteren, H. "Source Language Markers in EUROPARL Translations." *Proceedings of the 22nd International Conference on Computational Linguistics,* 2008.
Venuti, L. *The Translator's Invisibility: A History of Translation*. N.J.: Routledge, 2017.
Xiao, R., Lianzhen He, and Ming Yue. "In Pursuit of the Third Code: Using the ZJU Corpus of Translational Chinese in Translation Studies." *Using Corpora in Contrastive and Translation Studies* (2010): 182-214.