Treball/projecte de fi de màster de recerca

Spotting *Translationese* in two Corpora of Original and Translated Catalan Texts: an Empirical Approach

Pau Giménez Flores

Màster: Màster Universitari en Estudis de Traducció
Edició: 2011-2012
Director/ora: Carme Colominas i Toni Badia
Any de defensa: 2012

Col·lecció: Treballs i projectes de fi de màster de recerca

Programa oficial de postgrau
"Comunicació lingüística i mediació multilingüe"

Spotting *Translationese* in two Corpora of Original and Translated Catalan Texts: an Empirical Approach

Pau Giménez Flores

Supervisors: Carme Colominas and Toni Badia

September 2012

Translation and Language Sciences Department

Universitat Pompeu Fabra

Index

Abstract

This research investigates the phenomenon of translationese in two monolingual comparable corpora of original and translated Catalan texts. Translationese has been defined as the dialect, sub-language or code of translated language. This study aims at giving empirical evidence of translation universals regardless the source language.

Traditionally, research conducted on translation strategies has been mainly intuition-based. Computational Linguistics and Natural Language Processing techniques provide reliable information of lexical frequencies, morphological and syntactical distribution in corpora. Therefore, they have been applied to observe which translation strategies occur in these corpora.

Results seem to prove the simplification, interference and explicitation hypotheses, whereas no sign of normalization has been detected with the methodology used.

The data collected and the resources created for identifying lexical, morphological and syntactic patterns of translations can be useful for Translation Studies teachers, scholars and students: teachers will have more tools to help students avoid the reproduction of translationese patterns. Resources developed will help in detecting non-genuine or inadequate structures in the target language. This fact may imply an improvement in stylistic quality in translations. Translation professionals can also take advantage of these resources to improve their translation quality.

Keywords: Translationese, Translation Universals, Comparable Corpora, POS tagging

Resumen

Este trabajo trata el fenómeno del translationese en dos corpus monolingües comparables de textos catalanes originales y traducidos. El translationese se ha definido como el dialecto, sublengua o código de la lengua traducida. Este estudio tiene como objetivo proporcionar pruebas empíricas de los universales de traducción independientemente de la lengua de origen.

Tradicionalmente, la mayoría de las investigaciones realizadas sobre estrategias de traducción principalmente se han basado en intuiciones. La Lingüística Computacional y el Procesamiento del Lenguaje Natural proporcionan técnicas para extraer información fiable de frecuencias léxicas, distribuciones morfológicas y sintácticas a partir de corpus. Por ello, se han aplicado estas técnicas para observar qué estrategias de traducción se producen en los corpus.

Los resultados parecen confirmar las hipótesis de simplificación, interferencia y explicitación, mientras que no hay indicios de normalización con la metodología utilizada.

Los datos recogidos y los recursos creados para la identificación de patrones léxicos, morfológicos y sintácticos de las traducciones pueden ser de gran utilidad para investigadores, profesores y estudiantes de Traducción: los profesores tendrán más herramientas para ayudar a los estudiantes a que eviten la reproducción de patrones de translationese. Los recursos desarrollados contribuirán en la detección de estructuras que no sean genuinas o adecuadas en la lengua meta. Este hecho puede suponer una mejora en la calidad estilística de las traducciones. Asimismo, los traductores profesionales también pueden aprovechar estos recursos para mejorar la calidad de sus traducciones.

Introduction

Translated language and *translationese* as such have not been thoroughly investigated until the raise of Translation Studies as an independent discipline. In this section, we will explain how research on translation has shifted from a source or target language-dependent perspective into considering translated language a product itself with its own idiosyncratic properties. We will illustrate these changes with a general overview of the evolution of Contrastive Linguistics; the raise of Translation Studies, and the change of paradigm due to new resources and techniques provided by more recent disciplines such as Corpus Linguistics.

Scholars have distinguished two kinds of translationese: the one which is a product of the incompetence of the translator when "unusual distribution of features is clearly a result of the translator's inexperience or lack of competence in the target language" (Baker, 1993: 248), and the one which is produced by competent professionals but has some linguistic properties that differ from the source as well as the target language. Our research will try to confirm the hypothesis of the latter conception. Therefore, we will need to explain what are the so called Translation Universals and which research that has been done to prove or refute such universals of translation.

We will continue with a description of some of the research done more recently with corpus exploitation tools and more sophisticated techniques of computational linguistics. These studies are important because some of them give empirical evidence of how translationese may be detected automatically or semi-automatically, which is what we intend to achieve with our corpora.

In the following sections, we will define the theoretical framework of the study, its goals, hypotheses and methodology. Then, results will be described and conclusions drawn.

The Study of Translated Language

Translated language has not always been considered a language with specific properties that deserves a thorough study. In fact, the raise of Translation Studies as an independent discipline focused in translated language is rather recent.

The comparative study of languages has its roots in Contrastive Linguistics, which studies contrasts between languages and proves these contrasts with practical aims (Coseriu, 1987), mainly in foreign language learning. Contrastive grammar tries to find why two sections of two languages which show some kind of correspondence do not have analogous structure. Therefore, contrastive linguists intend to explain what in the source language (native language) and the target language (the foreign language) does not coincide in the surface. In other words:

> *"¿Con qué medios lingüísticos no análogos pueden tales y cuáles oraciones de la lengua A y de la lengua B expresar los mismos contenidos de pensamiento?". O bien, de forma más general: "¿Cuáles son los tipos de oraciones de la lengua A y de la lengua B que, siendo estructuradas de manera distinta, "dicen", sin embargo, "lo mismo", o sea, designan los mismos hechos extralingüísticos?".* […] "*¿Qué se dice efectivamente en la lengua B en una situación análoga o con referencia al mismo hecho?"* (Coseriu, 1987: 82-83)

This happens in translation praxis as well. Coseriu describes "empty" or "null" correspondences, which means that some structures in language A may have an "empty" or "null" correspondence in language B: for instance, the usual filler *remarquez* in

French (that could be translated into *guardi* in Italian) is not present in German in most cases. On the other hand, certain utterances that are grammatically structured in a way in one language may admit a correspondence in other domains of linguistic structure (in the lexicon and even in phonology). For instance, the grammatical differences in Spanish *tenía un hijo* - *tuvo un hijo*, correspond to a lexical difference in German: *hatte einen Sohn – bekam einen Sohn*. (Coseriu, 1987: 91)

Within the domain of Contrastive Studies, Vinay and Darbelnet (1958: 47), defined different technical processes of translation. One of them would be *transposition,* which consists in replacing a part of speech with a different one, without changing the meaning of the message: "dès son lever" is *transposed* into "As soon as he gets up". *Modulation* would be another process: a variation of the message due to a change of point of view. This is done when literal translation or even transposed translation provides a grammatical utterance but faced with the target language's essence: "It is not difficult to show…" would be *moduled* into "Il est facil de démontrer".

The idea of Translation Studies as an independent discipline comes from Holmes. Translation Studies should describe "the phenomena of translating and translation(s)" and establish "general principles by means of which these phenomena can be explaining to and predicted." (Holmes, 1972 in Venutti, 2004: 90). However, it is not until the raise of Descriptive Translation Studies (DTS) that researchers focus on the observation of existing translations (i.e. on translation products), considered as facts of target cultures, rather than on the translation process. The primary objective is the reconstruction of the norms governing translations in a given culture during a specific period, using both extratextual sources, such as statements made by translators and

publishers, and textual sources (the translated texts), which are compared with their source texts (Toury 1980, 1995). Translation Studies "shifted emphasis from the relationship between source and target text to the role and position of the target text as relatively autonomous from the source text, with the translator's role through time potentially more visible and more powerful as mediator of perspectives and ideology." (Olohan, 2004: 10). Therefore, translation began to be seen as a modification or adaptation of a source text to the constraints and norms imposed by the target literary system, but in a positive sense (see e.g. Hermans, 1985).

Large bilingual corpora gave Contrastive Linguistics specialists a much more solid empirical basis than had ever been available. Previous research, as in Vinay & Darbelnet (1958 and 1995) and Malblanc (1968), had been largely intuition-based. But intuitions lack an empirical basis, so a new paradigm in Translation Studies was provided by the raise of Corpus Linguistics.

Translated language began to be seen as a sub-language and "source-text-oriented Translation Studies shifted to descriptive, target-text and target-system-oriented studies" (Puurtinen, 2003). As Baker (1995: 224) states: "this reflects a move away from conceptual and formal representations of language, which have not proved very helpful in the past, to addressing natural language in all its variety and infiniteness". It was precisely under Mona Baker that the domain of Translation Studies underwent this corpus-based trend in the early 90s. She laid down the agenda for corpus-based Translation Studies and started collecting corpora of translated texts with a view to uncovering the distinctive patterns of translation. Her investigations brought to light a

number of potential ‘translation universals’ (Baker, 1993), that we will examine further on.

In this new theoretical framework, translated texts started to be considered as texts in their own right, which were analysed in order to “understand what translation is and how it works” (Baker 1993: 243). Thus, the focus of research in Translation Studies made a radical change of perspective in comparing translation with text production:

> *we need to effect a shift in the focus of theoretical research in the discipline, a shift away from comparing ST with TT or language A with language B to comparing text production per se with translation. In other words, we need to explore how text produced in relative freedom from an individual script in another language differs from text produced under the normal conditions which pertain in translation, where a fully developed and coherent text exists in language A and requires recoding in language B. This shift in focus may be facilitated by access to comparable corpora*. (Baker, 1995: 233)

And so, this is where the term *translationese* comes into play. This term was first used by Gellerstam in a study where he described vocabulary differences between original Swedish texts and Swedish texts translated from English. He saw a “systematic influence on target language (TL) from source language (SL), or at least generalizations of some kind based on such influence” (Gellerstam, 1986: 88). Gellerstam argued that there are syntactic fingerprints in translations; a set of linguistic features of translated texts which are different both from the source language and the target language.

Scholars have distinguished two kinds of translationese: the one which is a product of the incompetence of the translator when "unusual distribution of features is clearly a result of the translator's inexperience or lack of competence in the target language" (Baker, 1993: 248), and the one which is produced by competent professionals but has some linguistic properties that differ from the source as well as the target language. The first is due to translation errors, while the second is described as translation universals (Baker: 1993). For our research, we will primarily focus on the second one, although the border between both may not be clear in some cases.

This idea of sub-language or third code had also been described by other authors like William Frawley: "the translation itself […] is essentially a third code which arises out of the bilateral consideration of the matrix and target codes: it is, in a sense, a sub-code of each of the codes involved" (Frawley, 1984: 168). This *third code* is a result of the confrontation of the source and target codes and which distinguishes a translation from both source texts and original target texts at the same time.” (Baker, 1993)

Translation is therefore seen as product, independent of SL. Thus, “the term *translationese* is used in a neutral sense, meaning the translation-specific language or translation dialect, without any negative connotation”. (Øverås, 1998: 557-570)

This research will be based on the assumption of the existence of translationese. We will therefore try to validate the hypothesis of translationese and define the properties of translated Catalan.

Translation Universals

If we consider translated language as a sub-language, dialect or code *per se*, we must consider that there has to be some general features that apply to translated text, no matter which is the source or target language. Mona Baker (1993: 243) identifies these universal features as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems".

Gellerstam (1986: 91), for instance, compared texts in original Swedish and translated Swedish from English. He noticed an increase in English loanwords, a decrease in Swedish colloquialism (or dialectal features), a lexical choice of "standard translation in Swedish although the Swedish translation differs stylistically from the English original", "words from the international vocabulary (e.g. classical loanwords from Latin and Greek) are used in translations with new shades of meaning taken from the English lexeme (*drastisk, local, massiv*)", and a reorganization of semantic fields suiting the source language when SL and TL differed, e.g. with evaluative adjectives, or verbs of feeling.

These generalizations are strongly related with the translation universals described by other authors (see Baker 1993, 1995, 1996 and Laviosa 1996, 1997, 1998a. Olohan, 2001). Among the tendencies that would be caused by translation are greater explicitness, simplification, normalization, and an overall conservatism, although authors differ on the name and number of translations universals.

The *explicitation* or *explicitness* hypothesis implies that translations tend to be more explicit than source texts, less ambiguous. In other words, the translation process tends to add information and linguistic material. Translators may tend to repeat redundant grammatical items, such as prepositions, and overuse lexical repetition, which in turn results in a lower frequency of pronouns (hypotheses of Laviosa, 1996). Interestingly, *avoidance of repetition*, has also been suggested as a universal (Toury, 1995). Vanderauwera's (1985) and Blum Kulka's (1986) studies, which were not based on electronic corpora, showed that translators tend to explicate source text material, e.g. by using repetitions and cohesion markers. Øverås's (1998) findings on an English-Norwegian parallel corpus showed that added connectives and replacement of connectives with more explicit ones are forms of cohesive explicitation in translations. Olohan and Baker's (2000) corpus-based study found that the optional *that*-connective is more frequent in reported speech in translated than in original English. Puurtinen's (2004) observations on the use of connectives in Finnish children's literature partly contradict, although partly also support, the explicitation hypothesis.

Simplification means that the language of translations is assumed to be lexically and syntactically simpler than that of non-translated target language texts. One indicator of simplification might be a narrower range of vocabulary, a measure of which is a lower type-token ratio (i.e. ratio of the number of different words to the number of running words in a text). Another indicator might be a lower level of information load in translations. As information load is likely to increase with the lexical density of a text (ratio of the number of lexical items, as opposed to grammatical items, to the number of running words in a text), a lower lexical density might be regarded as an aspect of

simplification in translations. Laviosa-Braithwaite's (1996) results support this hypothesis: her corpus of translational English had a lower lexical density than her corpus of non-translational English. Another indicator of simplification in Laviosa-Braithwaite's study was that the proportion of high frequency words vs. low frequency words was higher in translations. Another feature of simplification, according to Laviosa, is the texts sentence length mean: the lower it is, the simpler the text is. However, Corpas (2008:129) shows how Laviosa's results contradict the simplification universal: although the sentence length mean is lower in translated journalistic prose, it is higher in translated narrative. Lexical simplification has not always been confirmed. Jantunen (2001) findings on synonymous amplifiers in Finnish translations do not support the assumption that translations would have a smaller variety of lexical items.

Examples of syntactic simplification include translators' tendency to use finite constructions more than non-finite (see Vanderauwera, 1985), and coordination instead of subordination. Eskola's research (2002) on the use of particular non-finite constructions in Finnish literary translations partly supports the simplification hypothesis: translations show a narrower range of alternative syntactic structures than Finnish originals. Similarly, Borin and Püritz (2001) show an overuse of non-finite constructions in translated English instead of other possible syntactic structures due to the SL (Swedish) grammar influence.

Normalization (sometimes also *conventionalization*, *conservatism*, *normalcy* or *standarization*) refers to the exaggeration of typical features of the target language. Metaphors and idioms can be *conventionalized*, and dialectal and colloquial expressions less frequent. This universal would be clearly related with the decrease of dialectal

features and the lexical choice of "standard translation" that Gellerstam described. Baker (1995: 238), for instance, proves that if "the overall lexical density of a corpus of English translations is significantly lower than that of a comparable corpus of original English, we might want to argue that translators use this feature, consciously or subconsciously, to control information load and to make a translated text more accessible to its new readership". Translations are assumed to be in various ways more unmarked and conventional, less creative, than non-translations or source texts: translated literature normally occupies a peripheral position (Ever-Zohar, 1990, and consequently translations tend to follow established conventional norms of the dominant target language literature, even to the point of becoming a major factor of conservatism. Baker (1996) says that normalization can be manifest in grammaticality, typical punctuation, and collocational patterns. Kenny's (2000, 2001) corpus study found signs of normalization in literary translators' use of words and collocations. Mauranen's (2000) results, however, point to an opposite tendency in academic texts: collocations and other multi-word combinations are more varied and thus less conventional in Finnish translations than in originally Finnish texts. Puurtinen (1995, 1997, 2003) has also discovered unconventional use of syntactic structures in translated children's literature. In non-corpus based studies, evidence of normalization has been provided by Toury (1980, 1995), Vanderauwera (1985), and Malmkjaer (1998). The degree of normalization is likely to be influenced by the status of the source text; e.g. highly valued classical literature is unlikely to be normalized to the same extent as popular literature or instruction manuals.

In addition, *interference* from the source text and language has been proposed as a translation universal (Toury, 1995; Mauranen, 2000). Interference can occur on all

levels from the morphological and lexical level to syntax, information structure, and quantitative features, for instance.

Tirkkonen-Condit makes an interesting contribution with her *unique items hypothesis*, according to which translated texts "manifest lower frequencies of linguistic elements that lack linguistic counterparts in the source languages such that these could also be used as translation equivalents" (Tirkkonen-Condit, 2002). These unique items or unique elements are not untranslatable, and they can be normal phenomena in the language, but they are unique in their translational potential due to they are not similarly manifested in other languages, or at least not similarly manifested in the source languages of the translations.

However, Tirkkonen-Condit thinks that translations are not readily distinguishable from original writing on account of their linguistic features, because they follow the norms of the original text language and therefore contradict the assumption that translation tends to be "clumsy, unidiomatic, and tends to sound like translations". In her study of translated Finnish, the so-called Translation Universals would be identified by human translators as patterns of a non-translated text rather than of a translation, where normalcy, for instance, tends to be attributed to non-translated Finnish rather than to translated Finnish. This is why Tirkkonen-Condit suggests that the hypothesis of translationese is, at least, controversial, whereas the *unique items hypothesis* can describe in a better way the translated or non-translated nature of a text. Puurtinen is also sceptical about the existence of translation universals:

> *The as yet relatively small amount of research into potential translation universals has produced contradictory results, which seems to suggest that a search for real, 'unrestricted' universals in the field of translation might turn out to be unsuccessful.* Puurtinen (2003: 403)

Regarding this hypothesis, Toury says that "the requirement to communicate in translated utterances may impose behavioural patterns on its own" (1995: 225), so if we consider this unique items hypothesis valid, we would have to assume then that translations in this sense are "less normal" than original texts. However, high-frequency items have more bearing on normalcy than low frequency ones, according to Tirkkonen-Condit. This hypothesis is, at least, arguable: these *unique items* may be just another kind of translation universal or proof that, for instance, standarisation or normalisation exists and that these *unique items* are just a way in which this process takes shape.

Empirical Methods and Translation Studies

The increasing use of corpora and computational methods in Translation Studies has provided scholars with data and tools to obtain empirical evidence of translationese. Intuitive or non-corpus based methodology may induce to error, but if we use empirical methods in corpora in such a way that that we can predict if a text is translated or not, we will definitely conclude that translationese exists, and that it is automatically or semi-automatically detectable.

Mona Baker was the pioneer in using comparable corpora to describe translated language, but other interesting studies have been done in this field. Laviosa (Laviosa-Braithwaiste, 1996) investigated the linguistic nature of English translated text in a sub-section of the English Comparable Corpus (ECC), consisting of two collections of narrative prose in English: one made up of translations from a variety of source languages and the other with original English texts produced during a similar time span. The study revealed four patterns of lexical use in translated versus original texts: a relatively lower proportion of lexical words versus grammatical words, a relatively higher proportion of high-frequency versus low-frequency words, relatively greater repetition of the most frequent words, and less variety in the words most frequently used.

Øverås (1998) investigation of explicitation in translational English and translational Norwegian aimed to describe the specificity of the language of translation regardless of the contrastive differences existing between the two languages in contact.

Kenny's research (1999) aimed to develop a methodology for the investigation of lexical creativity and lexical normalization in a parallel corpus of contemporary experimental German literary texts and their translations into English. The research found that forty-four per cent of creative forms were normalized in English translations.

Olohan and Baker's (2000) investigation tested the explicitation hypothesis, based on the linguistic, comparative analysis of the omission and inclusion of the reporting *that* in translational and original English. The authors discovered a preference for the use of *that* with reporting verbs in translated versus non-translated English.

Borin and Prütz (2001) compared original newspaper articles in British and American English with articles translated from Swedish into English. They searched part-of-speech n-grams and found an over-representation of adverbs, infinitives, pronouns, sentence-initial verbs, and sentence-initial prepositions in translation, which could be interpretable in terms of translationese effects.

Puurtinen (2003) aimed to find potential features of translationese in a corpus of Finnish translations of children's books. She detected a lack of colloquial words in translated texts, high frequencies of nonfinite constructions and specific uses of certain conjunctions in Finnish children's literature.

Similarly, Rayson et al. (2008) showed how translationese could be detected fully automatically by comparing the frequencies of words and phrases in three corpora of the Information and Communication Technology domain. Studying translation revision carried out by native speakers of English might offer one way in to study

Chinese-to-English translationese. The authors carried out a systematic comparison of texts translated from Chinese to English by Chinese translators with the same texts subsequently edited by native speakers of English. They used corpus-based techniques such as keywords, corpus annotation and n-gram extraction tools. The results show that techniques from Corpus Linguistics could be used to assist in a quantitative study of translation revision.

An interesting counter-study to the translationese hypothesis is *Translationese – a myth of an empirical fact?*, where Tirkkonen (2002) explained how human translators did not identify well if a text was translated or not. Their performance was poor: only 63.1% of correct choices. However, she admitted that it might be due to the fact that human translators have negative perceptions of translated language. Thus, the absence of colloquialisms, for instance, made them think that a text was originally written rather than translated.

Baroni and Bernardini's research (2006) is a very innovative study: they apply supervised machine learning techniques to corpora in order to give empirical evidence of translationese. They apply the machine learning algorithm SVM (Support Vector Machines) on two monolingual comparable corpora of translated and original Italian texts. Their results describe how a machine performs better than competent professional translators in detecting translated versus non-translated text. They conclude that SVM is the algorithm which gives better results, due the fact that it allows to use a large amount of linguistic-poor features. The combination of functional words is the feature that gives a higher accuracy, precision and recall in the detection of translated texts.

Baroni's research is the first to apply, as far as we know, machine learning techniques into translationese detection. Previously, this methodology had been used in related fields such as authorship attribution (Kindermann et al., 2003) or text categorization (Sebastiani, 2002).

Theoretical Framework

Taking into account the research previously explained, this study finds itself at the crossroad of Corpus Linguistics, Translation Studies and Computational Linguistics.

This is a Corpus Linguistics study in the sense that it uses corpora as the main source of data to try to build a theory regarding translated language. We also use compiled and pre-processed texts as a source of hypotheses in order to confirm our previous suppositions, namely, the existence of translationese. However, this is not a Contrastive Linguistics study which uses parallel or multilingual corpora, as Olohan (2004: 10) explains: "the comparative model is well established as a framework within which to study translation, generally involving analysis of certain features of the source language or text and comparison with their translation into a target language or text". On the contrary, we aim to study translated text from monolingual comparable corpora. By monolingual comparable corpora, we mean two sets of texts in the same language, translated and non-translated, with the same genre, style and years of production

Regarding this confluence of Corpus Linguistics and Translation Studies, Olohan (2004: 16) describes very well how Corpus Linguistics aims, aspirations and applications can be applied to Translation Studies research, and gives orientations for Translation Studies research, namely: "an interest in the descriptive study of translations as they exist"; "an interest in language as it is used in the translation product, as opposed to the study of language in a contrastive linguistic, i.e. system-oriented, sense"; "an interest in uncovering what is probable and typical in translation, and through this, in interpreting what is unusual"; "a combining of quantitative and qualitative corpus-

based analysis in the description, which can focus on (a combination of) lexis, syntax and discoursal features"; and "application of the methodology to different types of translation, i.e. translation in different socio-cultural settings, modes, etc."

Translation Studies represent a broad field that has been developed, under this denomination, since 1972, when Holmes proposed the notion of *Translation Studies* to the discipline that "describes the phenomena of translating and translation(s)" and establishes "general principles by means of which these phenomena can be explaining to and predicted." (Holmes, 1972 in Venutti, 2004: 90). Therefore, our research can be considered a traductological study because it tries to validate or refute the existence of translationese and to define the linguistic properties of translated language as a product.

Moreover, this is a Computational Linguistics research since Computational Linguistic and Natural Language Processing techniques are used: information extraction procedures are applied in order to identify linguistic patterns of translationese; taggers and parsers are used, and scripts *ad hoc* have been programmed in order to extract information (n-grams, POS, etc.) from the corpora.

Hence, if corpora are the source of data and hypotheses, and Translation Studies a broader theoretical framework where this research lies, the domain of Computational Linguistics is extremely important since it provides empirical techniques to implement the experiments.

Goals

The main goals of this research are two: first, to validate the hypothesis of the existence of *translationese* empirically; second, to capture the linguistic properties of translationese in observable and refutable facts.

The first goal is achieved by comparing statistical measures that may shed light on which translation processes take place. Also, lexical and syntactical patterns are described to demonstrate how translationese takes shape in our corpora.

In a second stage, different phenomena of translationese are classified in relation with the universals of translation, that is to say, in which structures simplification, explicitation or normalization take place.

Secondary goals are the implementation of NLP applications such as n-gram, passive structures and POS tag frequency extractors, and the development of a consistent and extrapolative methodology for translation universals research.

Hypotheses

The hypotheses we aim to validate are two: first, translationese exists and it is observable in explicitation, simplification and normalization strategies in translated texts. These universals can be manifested in different morphological, syntactical and lexical patterns.

Second, translationese can be described with empirical methods applied to corpora. A variety of statistical parameters can measure lexical richness, for instance, to prove the simplification hypothesis. Moreover, the extraction of POS frequencies provide data to demonstrate processes such as explicitation or interference.

If the hypotheses are confirmed, the methodology will be extrapolated into other languages to prove to what extent translationese is language-dependent. Otherwise, the hypotheses and methodology will need to be reviewed in order to be able to obtain consistent conclusions with other languages.

Methodology

We will use Computational Linguistics techniques applied to monolingual comparable corpora. Monolingual comparable corpora consist of a set of texts in the original language and translated language. Baker, the first scholar to use these corpora, defined them as "something which is a cross between parallel and multilingual corpora" (Baker 1995: 234). This is not equivalent to parallel corpora, where original texts and their translations are aligned, or multilingual corpora. On the contrary, both texts are in the same language but some are originally written in that language and the other are translated into that language. Using comparable corpora allows to "identify patterning which is specific to translated texts, irrespective of the source or target languages involved" (Baker, 1995: 234). Multilingual corpora may provide how 'equivalent' items and structures behave in various languages and this can be useful for training translators, but "research based on, or limited to, multilingual corpora cannot provide answers to theoretical issues which lie at the heart of the discipline and cannot explain the phenomenon of translation per se" (Baker 1995: 233*)*. In other words, "access to comparable corpora should allow us to capture patterns which are either restricted to translated text or which occur with a significantly higher or lower frequency in translated text than they do in original" (Baker 1995: 235).

Baker makes a clear distinction from contrastive linguistics methodology, which uses parallel or multilingual corpora, according to Baker's terminology. As Olohan (2004: 10) explains, "the comparative model is well established as a framework within which to study translation, generally involving analysis of certain features of the source language or text and comparison with their translation into a target language or text."

In this preliminary study, we use two monolingual comparable corpora of original and translated texts in Catalan in the domain of arts and architecture. It is very important that, in order to have comparable corpora, the domain has to be restricted to the same topic and time span. We have compiled a corpus of art and architecture from biographies, guides and other texts about the work of Dalí, Miró, Picasso and Gaudí. The choice of these four artists was due to the need to have texts in Catalan. Since these are some of the most international Catalan artists or with a close relationship with Catalonia, it was more likely to find literature in Catalan, both original and translated. The source languages (SL) are English and German, in a proportion of 50%. English and German translations are mixed with no way of identifying which part belongs to each SL. The compiled corpus originally had 228.476 tokens in original Catalan and 265.824 in translated Catalan.

First, the corpora have been tokenized, lemmatized, tagged and parsed with CatCG (Alsina, Badia *et al.,* 2002). CATCG is a shallow parser for Catalan. It uses the Constraint Grammar formalism and contains three basic tools: a morphological analyzer, a POS tagger and a shallow parser. See below an example of the output of running the original corpus through CatCG. The different columns show the wordform, lemma, POS and grammatical function. There are also opening and closing sentence markers (<s id="1">, </s>) which provide useful information to calculate the sentence length mean:

```
<s id="1">
Biografia          biografia_biografiar          Nom_Verb          N5-FS_VDR3S-  NA,O,S
CD_CN>_Subj
<enty>
Salvador           Salvador          Nom     N46MS  INDEF  CD
</enty>
Dalí       dalí       Nom     N5-MS   INDEF   <NN
va         anar       Verb    VDR3S-NA,P,SS   VAux>
néixer     néixer     Verb    VI----  NA,SS   <C_Advl_VPrin
a          a          Prep    P       INDEF   Advl
Figueres   figuera    Nom     N5-FP   INDEF   <P
el         el         Det     EA--MS  INDEF   DN>
1904       1904       Nom     N5-MS   INDEF   Subj
i          i          Conj    CC      INDEF   Conj
va         anar       Verb    VDR3S-NA,P,SS   VAux>
morir      morir      Verb    VI----  NA,P,SS <C_Advl_VPrin
a          a          Prep    P       INDEF   Advl
la         el         Det     EA--FS  INDEF   DN>
Torre      torre      Nom     N5-FS   INDEF   <P
<enty>
Galatea    Galatea    Nom     N46FS   INDEF   <NN
</enty>
el         el         Det     EA--MS  INDEF   DN>
1989       1989       Nom     N5-MS   INDEF   CD
.          .          PT      .       0       PT
</s>
```

Different tools for implementing experiments have been used. Wordsmith Tools software has proven to be useful in obtaining data in previous corpus studies. Frequency lists of each of the two corpora have been extracted in order to compare statistics that may reflect a sign of translationese, namely sentence length mean, type/token ratio, lexical richness, lexical density or information load.

However, although Wordsmith seems quite useful for extracting concordance lists of particular tokens or type frequency lists, there is no lemmatization and POS information. Hence, we have implemented Python scripts in order to extract lists of lemmas' frequencies from the tagged corpus (see Annex). POS tags allow the extraction of frequencies of POS in each corpus, which is essential to calculate the number of lexical and functional words and, consequently, the information load, lexical richness,

lexical density, and mean sentence length measures. These may be indicators of the simplification universal.

Regarding how Laviosa (1998a) uses Wordsmith and lexical density as an indicator of simplification, Corpas argues:

> *Esta forma de calcular la densidad léxica considera las variantes morfológicas de cada palabra como palabras distintas, lo cual distorsiona los resultados sobre la variedad léxica de los textos analizados. Merece la pena señalar que, en cualquier caso, la autora no hubiera podido computar la densidad léxica en función de los lemas, ya que utiliza Wordsmith Tools, un programa de gestión de corpus que no ofrece la opción de análisis morfológico.* Corpas (2008: 128)

This is why we have developed scripts with the programming language Python using the tagged corpora as an input in order to extract lemmas, POS, and n-grams frequency lists (Benjamin Kolz and Pau Giménez, 2012). See below some examples of the outputs:

Lemmas frequency list

```
1#el#27551
2#de#18251
3#,#12813
4#.#7134
5#a#6630
6#i#6375
7#que#4716
8#un#4538
9#indef#4042
10#en#3297
```

POS frequency list

```
1#Nom#53344
2#Det#35831
3#Prep#35037
```

4#PT#24941
5#Verb#24675
6#Adj#15550
7#Pron#9168
8#Conj#8948
9#Adv#7574

Trigram frequency list

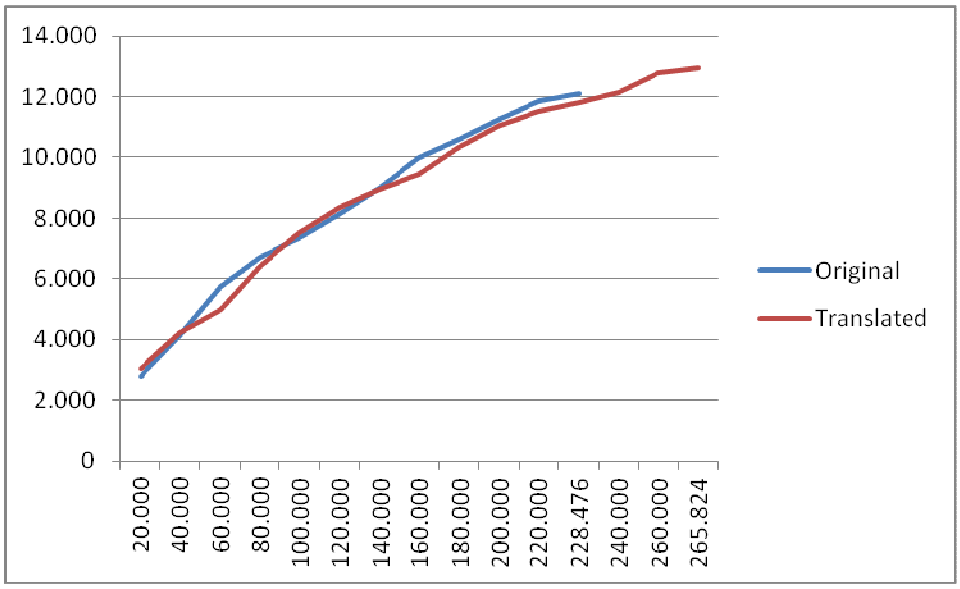
1#de la seva #198
2#la Sagrada Família #118
3#de el seu #109
4#per a la #108
5#de la ciutat #105
6#un de els #104
7#, a el #95
8#, a la #92
9#de l' edifici #92
10#de la Sagrada #90

Lists of n-grams (bigrams and trigrams) frequencies have been extracted because they may provide information of simplification or explicitation, for instance, of certain phrases.

Regarding the size of the corpora, they have a different number of tokens: the original corpus has 228,476 whereas the translated one has 265,824. In a first approach we thought that it was enough to normalize the data by obtaining a normalization factor with a simple formula:

N= Tokens Translated / Tokens Original → 265.824/228.476=**1,1634**

To be able to compare the lemmas, we normalized the amount of lemmas in the original corpus by applying the normalization factor. Nevertheless, the increase of tokens is not proportional with the increase of lemmas. This means that if we divide a corpus in different subsections and we calculate the accumulated amount of lemmas in each subsection, it tends to stabilize the bigger the corpus is. In other words, the bigger the corpus is, the fewer new lemmas will appear in each new subsection:



To sort out this problem, the translated corpus has been cut in order to have the exact same number of tokens as the original one. This way, the normalization factor has been discarded since all the data computed with this new corpus is now comparable. One may think that cutting this corpus does not seem a fair scientific methodology. However, it has to be taken into account that both corpora have been built from different sources such as art guides or architecture books, so there is not thematic progression which may be broken. That is to say, it is assumed that there is not a different distribution of data in the segment discarded. The results obtained seem to corroborate this assumption.

Apart from calculating lemmas and POS globally, a script has been developed to count the new lemmas for each new 20,000 tokens, as well as the accumulated lemmas for each subsection of the corpus. This means that the original corpus has 11 subsections of 20,000 tokens each plus the remaining subsection of 8,476 tokens. Similarly, the translated corpus has been divided into these subsections plus three more in order to achieve the 265,824 tokens. These last three sections have been considered with the purpose of having a broader scope to observe the progression of lemmas.

The abovementioned parameters have been then calculated for the original, the translated and the shortened translated corpus (*translated-cut*): number of tokens, types, lemmas, type-token ratio, sentences, sentence length mean, lexical density, lexical richness, information load, and POS frequencies. As Baker (1995) and Laviosa (1998, 2002) explain in their research, these measures may show different frequencies in original and translated corpora. Lower lexical density in translated texts, for instance, is generally described as an indicator of simplification.

Results

The types calculated of the original corpus are 22255, 20138 for the translated-cut and 22399 for the translated corpus:

	Original	Translated-cut	Translated
Tokens	228476	228476	265824
Types	22255	20138	22399

Table 1. Types and Tokens

Table 1 shows how the number of types is significantly higher in the original than in the translated-cut corpora: 2117 more tokens. Actually, the whole translated corpus has just 144 more types than the original while it has 37348 more tokens. At a first glance, this difference of the number of types may confirm Laviosa's hypotheses (1998a): translated texts would present a lower lexical density. However, as Corpas notes,

> *"si bien estos indicadores de simplicidad resultan aceptables en líneas generales, calcular la variedad léxica en función de las palabras tipo, sin tener cuenta los lemas, es, desde nuestro punto de vista, incorrecto. En otras palabras, desde tal perspectiva el estilo de un autor que utiliza palabras como* cantar, cantaría, cantante *y* cantantes *se consideraría igual de rico que el de un autor que* utilizase actuar, representa, artista *y* cantantes*".* (Corpas 2008:129),

This is why the tagged corpus has been the main source of data in order to have more reliable parameters of lexical density. While most of previous studies have only focused in studying the raw text file of the corpus using software as Wordsmith, we have designed programs *ad hoc* to extract statistics from types, lemmas, POS, n-grams and even particular syntactic structures such as passive forms (Benjamin Kolz and Pau

Giménez, 2012). This flexibility in the management of data is not possible with Wordsmith, which does not lemmatize nor allow the use of POS tags.

Therefore, the number of lemmas in each corpus can be counted and compared to have a more precise indicator of lexical distribution:

	Original	Translated-cut	Translated
Lemmas	12094	11800	12949

Table 2. Lemmas

As it can be seen in Table 2, the difference of lemmas between the original and translated-cut is significantly lower than the difference of types, 294 lemmas less in the translated corpus for 2117 types less. Proportionally, the amount of types in translated-cut is the 90% of the original corpus whereas the lemmas proportion is 98%. Following Corpas argument, the difference of types might be regarded as significant, and therefore indicator of a bigger lexical richness in the original text. However, the small difference of lemmas cannot lead to categorical conclusions about lexical richness. Table 3 shows the proportion between lemmas and types in each corpus and between corpora.

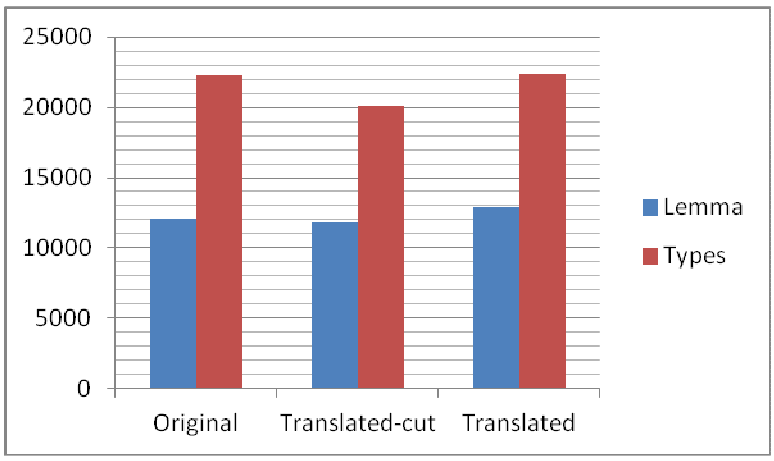


Table 3. Lemmas and types in each corpus.

The explanation of these differences is that the proportion between lemmas and types is higher in the original than in the translated-cut corpus. For example, if the lemma *anar* may have the types *va, anava, anirà, van*, etc., a higher frequency of the forms of this verb may appear in the original corpus as a sign of a broader lexical variety. On the other hand, in English there is less morphological variety in verbs and less tenses than in Catalan, so a parameter that calculates the correlation between lemmas and types may be useful to study lexical variety and the influence of the SL morphological and syntactical structures on translated language. Therefore, the same way there is a standard measure to measure lexical density called type/token ratio, we have created another indicator of lexical diversity which indicates the correspondence between lemmas and types:

Lemma/type ratio= Number of types/ Number of lemmas

Similarly, a relation between lemmas and tokens will provide more accurate information about lexical richness than type/token ratio:

Lemma/token ratio= Number of tokens/Number of lemmas

Table 4 shows this information:

	Original	**Translated-cut**	**Translated**
Tokens	228476	228476	265824
Types	22255	20,138	22399
Lemmas	12094	11800	12949
Type/token ratio	10.27	11.35	11.87

Lemma/type ratio	1.84	1.70	1.73
Lemma/token ratio	18.9	19.3	20.53

Table 4. Type/token, lemma/type and lemma/token ratios.

As we can see, there are some differences between the type/token, lemma/type and lemma/token ratios. Type/token ratio shows an average of how many tokens correspond to each type. It is slightly lower in the original than in the translated-cut corpus. On the other hand, lemma/type ratio shows that there is a bigger proportion of types for each lemma in the original corpus. Thus, lemma/token ratio is slightly lower in the original corpus.

Although these figures do not provide conclusive information due to the small differences, they are consistent and show in all parameters a slightly higher lexical richness in the original corpus.

So far, it has been proved the importance of working with lemmas rather than types. Lexical richness -the inverse calculation of the abovementioned lemma/token ratio- (Corpas, 2008), sentence length, lexical density and information load (Laviosa 1998, 2002) may provide more information:

	Original	Translated-cut	Translated
Lexical Richness (lemmas/tokens)	0.053	0.051	0.049
Sentence Length Mean	30.01	31.29	30.16
Information Load (lexical words/tokens)	0.53	0.54	0.54
Lexical density (lexical words/ functional words)	0.88	0.85	0.85

Table 5. Lexical richness, sentence length mean, information load, lexical density

Table 5 shows again a small difference in the figures. They seem to confirm previous results and they are consistent in all measures but in one case: information load is one hundredth higher in translated-cut. Obviously, it is a very small difference by all means, but it strikes that it is the only parameter that shows a contrary result. Lexical richness is higher in the original, sentences are slightly longer in translated-cut –which may be a hint of redundant information- , and the proportion of lexical and functional words is also higher in the original corpus.

However, taking a deeper look in the results, the deviance of the information load figures can be explained.

POS	Original	Translated-cut	Translated
Nom	53344	51543	59890
Det	35831	34420	40032
Prep	35037	34984	40376
PT	24941	25529	30016
Verb	24675	26555	30981
Adj	15550	15000	17554
Pron	9168	9626	11139
Conj	8948	8750	10187
Adv	7574	9734	11404
Unknown	4042	3077	3582
ambiguous	9366	9258	10663
Total	**228476**	**228476**	**265824**

Table 6. POS tags.

CatCG morphological tagger has a precision of 92% and recall of 98%, which means that there are certain tokens that are not identified ("unknown") or whose POS cannot be disambiguated ("ambiguous"). Table 6 shows the POS information that has been extracted from the tagged corpora. There are 4042 unknown and 9366 ambiguous tags and in the original corpus, as opposed to the 3077 unknown and 9258 in the translated-cut corpus. It means that there are more than 1000 tokens in the original corpus than in the translated-cut one that cannot be classified as lexical or functional words. Most of the unknown words are proper nouns or verbs that have not been found in the CatCG dictionary, so in these 1000 tokens there is a proportion of lexical words that surpasses the hundredth of difference of information load. On the other hand, there is a very significant difference of verbs in both corpora. Translated-cut corpus has almost 2000 more verbs. The calculation of lexical and functional words was done without separating auxiliary verbs, which are not lexical words, from main verbs, which are. The assumption was that the proportion of auxiliary verbs would be similar in both corpora, so it would not distort the relative figures of lexical and functional words. However, since the difference was so marked, we decided to observe more thoroughly the verbs distribution and focus on the auxiliary verbs. The hypothesis was that the translated texts could have more auxiliary verbs.

POS	Original	Translated-cut	Translated
Aux Verbs	2729	3604	4275

Table 7. Auxiliary Verbs.

Indeed, as Table 7 shows, there is a high difference of auxiliary verbs: 875 more in the translated-cut corpus. Therefore, at least 875 more words of the translated-cut

corpus would have been counted as lexical instead of functional words. This would have been another factor of distortion in the information load calculation.

Such differences in auxiliary verbs frequencies have led us to study passive structures. It has been assumed that the high number of auxiliary verbs in translations could be caused by the presence of more passive structures than in the original. The hypothesis was that there could be more passives in the translated corpus due to the influence of the source languages, German and English, since they use more passive structures than Catalan.

A Python script has been developed for extracting passive structures based on the morphological and syntactical tags in the corpora.

Results show that the original corpus contains 575 passive structures and the translation-cut corpus, 704. Again, the hypothesis is validated. This deviance (18%: 129 more passive structures in translated-cut) is either a proof of translationese or a proof of the influence of the source languages on the target language. A more thorough study should be carried out to observe if such differences are due to translationese effects or to a poor translating performance.

Table 8 shows the 20 more frequent passive structures in each corpus:

Passive structure	Frequency
està format	6
és considerat	5
ser realitzat	4
ser construïda	4
ser realitzada	4
estan fetes	4

Passive structure	Frequency
ser pintat	7
ser cridat	5
és donada	4
ser feta	4
és cridat	4
és difamat	4

està fet	3	és considerat	4
està dedicat	3	ser difamat	4
fou substituït	3	és sentit	3
és obert	3	és representat	3
ser declarat	3	ser comparat	3
ser reconegut	3	ser considerat	3
ser empresonat	2	són coronades	3
és marcada	2	està mancada	3
estan decorats	2	està formada	3
fou construïda	2	ser exposada	3
són representats	2	és ferit	3
estar protegit	2	està contingut	3
està feta	2	és nomenat	3
ser projectat	2	ser convidat	3

Table 8. Passive structures in original and translated-cut corpus

Regarding the use of pronouns, Table 6 shows a significant higher number of pronouns in translated-cut: 9626 versus 9168. Again, considering that in English and German subject is obligatory, but optional in Catalan, a calculation of the strong personal pronouns frequency (*pronoms personals forts*) may provide useful information about these different figures. Since there is a tag in the corpora that specifies this type of pronoun, we used a regular expression to extract only strong personal pronouns. Results can be observed in Table 9.

POS	Original	Translated-cut	Translated
Strong personal pronouns	320	630	686

Table 9. Strong personal pronouns in original and translated-cut corpora

Again, our assumption was confirmed: there is a proportion of 97% more strong personal pronouns (310) in translated-cut than in the original corpus. This fact reflects that there has been an explicitation process. As it happened with passive structures, further research would be needed to demonstrate if this phenomenon is a result of translationese or a poor translating performance.

The different numbers of adverbs (7574 vs. 9734) is surprisingly high. Tables 9.1 and 9.2 show the frequencies of the 30 most frequent adverbs.

1#com#1538
2#no#910
3#més#704
4#també#445
5#ja#330
6#molt#308
7#on#281
8#així#167
9#quan#163
10#només#148
11#encara#142
12#ara#122
13#sempre#105
14#tan#97
15#tant#73
16#mai#68
17#aleshores#68
18#abans#64
19#alhora#62
20#ben#61
21#aquí#55
22#bé#52
23#finalment#50
24#posteriorment#47
25#especialment#46
26#actualment#45
27#sobretot#45
28#totalment#45
29#gairebé#43
30#avui#43

Table 9.1 Adverbs frequency in original corpus

1#com#1187
2#no#1292
3#més#790
4#també#503
5#ja#299
6#quan#276
7#on#269
8#només#236
9#molt#233
10#ara#199
11#encara#199
12#tan#180
13#aquí#168
14#així#149
15#sempre#141
16#abans#117
17#gairebé#98
18#sovint#88
19#mai#85
20#tant#83
21#sobretot#81
22#ben#76
23#bé#73
24#aleshores#71
25#tanmateix#70
26#potser#56
27#especialment#55
28#pas#55
29#precisament#47
30#clarament#45

Table 9.2 Adverbs frequency in translated-cut corpus

Such difference may be explained by cohesive explicitation (Øverås, 1998), which means that connectives are added in translated language. The word *tanmateix*, for example, appears 70 times in the translated corpus and only 28 in the original. We also looked for other cohesive structures such as *fins i tot* in Wordsmith, and frequencies were higher in translated Catalan as well.

Conclusions

Regarding lexical variety, there is evidence that the translated-cut corpus has lower lexical richness. Although differences are not significant, our results are consistent: they suggest an increase of these differences in larger corpora. If so, the simplification universal is confirmed in these particular corpora: translated texts have a lower lexical richness than original texts. However, larger corpora will be needed in further research to corroborate these tendencies.

Important differences have been found in POS frequencies. The difference of pronouns is mainly caused by a much higher use of strong personal pronouns in the translated corpus. Since German and English have obligatory subject personal pronouns, there seems to be an influence of the source language in the translation process. Explicitation is therefore confirmed in the overuse of pronouns in Catalan translations. Nevertheless, it is difficult to know if this process is due to a universal translation process or a poor translating performance.

Likewise, a significant higher volume of passive structures have been found in the translated corpus. Since English and German use more passives than Catalan, there seems to be a clear syntactic interference with the source language.

Another proof of explicitation is the much higher frequency of adverbs in the translated corpus. Explicitation implies that translation process tends to add information and linguistic material. Frequencies observed show an overuse of certain connectives, which has been described in previous studies as cohesive explicitation.

None of the data has proved any normalization process. It is very difficult, if not impossible, to investigate any standarization process without having the source language corpus to compare, for instance, how certain idioms are translated.

In further experiments, source language corpora will be needed to be able to study the influence of source language.

Regarding the distribution of lemmas and tokens through the corpora, it is important to note that the frequency of new lemmas tends to stabilize the bigger the corpus is, so the ratio between new lemmas and new tokens decreases progressively. Larger corpora and more stylistic varieties will be needed to confirm this tendency. This data will be an interesting starting point for a more thorough study of information distribution through texts.

In short, for more conclusive results, it is necessary to compile larger corpora with more variety of styles and domains. Moreover, the methodology of using monolingual comparable corpora needs to be complemented with source language corpora: either parallel with the translated corpus, comparable, or both. It seems quite clear that without these resources there is no way of validating hypotheses such as the normalization universal. Also, new parameters will have to be taken into account, such as ambiguity and sentence complexity.

Finally, it is important to define a systematic methodology to find when the abovementioned phenomena are caused by the translation universals or by a bad translation performance. Limits are not clear, but if there are significant differences between translated and original texts, it means that an ideal translation should have the same or very similar parameters as original texts. An interesting challenge for the future will be to establish a methodology to automatically rate the quality of a translation based on translationese indicators.

Appendix

```
# -*- coding: utf-8 -*-

# Translationese Analyzer
# Benjamin Kolz and Pau Giménez
# 2012
# for Python 2.7
# @ Universitat Pompeu Fabra

import re,os,sys
import random

class Token:
	Counter_Tokens = 0

	def __init__(self):
		self.form=None
		self.lemma=None
		self.pos = None

		Token.Counter_Tokens += 1  # class member

class N_Gram(list):

	def __init__(self,n,token):
		                                                        # first token
		self.append(token)
		self.string="" # init

	def get_string(self):
		for token in self:
			if token != None:
				self.string+= token.form + " "
			else:
				self.string += "None"

	def join_to_string(self):
		print "join() in N_Gram started"
		string ="" # init
		for token in self:
			if token != None:
				string += token.form + " "
			else:
				string += "." # end symbol
		return string

class Corpus_Admin:

	def __init__(self):
		print "Corpus_Admin started"


		# build from original corpus
```

```
                or_token_list=self.read_tokens("Corpus_originals_iac.txt") # returns list of Token objects
                or_list_bigrams = self.build_n_gramas(2,or_token_list) # get bigrams
                or_list_trigrams= self.build_n_gramas(3,or_token_list) # get trigrams

                # build from translated corpus
                tr_token_list=self.read_tokens("Corpus_traduits_iac.txt") # returns list of Token objects
                tr_list_bigrams = self.build_n_gramas(2,tr_token_list) # get bigrams
                tr_list_trigrams= self.build_n_gramas(3,tr_token_list) # get trigrams

                # TEST
        #       test_token_list=self.read_tokens("test_tokens.txt") # returns list of Token objects
                #count_tokens=self.count_tokens(test_token_list)
        #       print "Count_Tokens: ",count_tokens
                #raw_input()
                #test_list_bigrams = self.build_n_gramas(2,test_token_list) # get bigrams


                # create lemma_freq_list
                lemma_freq_list=self.get_lemma_freq_list(tr_token_list)
                self.write_lemma_freq_list(lemma_freq_list,"lemma_freq_list.txt")

                # create types_freq_list
                types_freq_list=self.get_types_freq_list(tr_token_list)
                self.write_lemma_freq_list(types_freq_list,"types_freq_list.txt")


                # create pos_list
                pos_list=self.get_pos_list(or_token_list)
                sum_lexical_words=self.sum_lexical_words(pos_list)
                sum_functional_words=self.sum_functional_words(pos_list)
                #print "Lexical Words: ",sum_lexical_words
                #print "Functional Words: ",sum_functional_words
                #raw_input()
                self.write_pos_list(pos_list,"pos_list.txt")
                #raw_input()


                #for bigram in tr_list_bigrams:
                #       bigram.get_string()

                for trigram in or_list_trigrams:
                        trigram.get_string()

                high_freq_trigrams=self.get_highest_freq_bigrams(or_list_trigrams)
                #print high_freq_bigrams
                #raw_input()
                sorted_freq_trigrams=sorted(high_freq_trigrams.items(),reverse=True,key=lambda x: x[1])# nach value ordnen, absteigend
                #print sorted_freq_bigrams
                #raw_input()
                self.write_bigram_freq_to_txt(sorted_freq_trigrams,"freq_trigrams.txt")

                # TEST OF POSSIBLE FUTURE IMPLEMENTATIONS
                mle =  self.maximum_likelihood_estimation_bigrams(test_list_bigrams,"en","un")
                print "MLE: ",mle
                raw_input()
                sentence= self.sentence_generator(test_list_bigrams)
```

```
		"""
		for n_gram in list_bigrams:
			for token in n_gram:
				if token != None:
					print token.form
			print "=============="
		"""

		"""
		found_n_gramas=self.get_all_x_at_position_y("anar",1,or_list_trigrams)	# searched string, position, list of n_gramas
		print "Found n_gramas:"
		for n_gram in found_n_gramas:
			for token in n_gram:
				if token != None:
					print token.form
			print "=============="
		"""

		"""
		# for orignal corpus
		self.write_n_gram_to_txt(or_list_bigrams, "Output/original/bigrams.txt")
		self.write_n_gram_to_txt(or_list_trigrams, "Output/original/trigrams.txt")

		# for translated corpus
		self.write_n_gram_to_txt(tr_list_bigrams, "Output/translated/bigrams.txt")
		self.write_n_gram_to_txt(tr_list_trigrams, "Output/translated/trigrams.txt")
		"""

	def sum_lexical_words(self,pos_list):
		sum=0
		for (pos,freq) in pos_list:
			if pos in ["Nom","Verb","Adj","Adv"]:
				sum+=freq

		return sum
	def sum_functional_words(self,pos_list):
		sum=0
		for (pos,freq) in pos_list:
			if pos in ["Det","Prep","Pron","Conj"]:
				sum+=freq
		return sum
	def get_lemma_freq_list(self,token_list):
		lemma_freq_hash={}
		for token in token_list:
			lemma=token.lemma
			if lemma in lemma_freq_hash:
				lemma_freq_hash[lemma]+=1
			else:
				lemma_freq_hash[lemma]=1

	lemma_freq_list=sorted_freq_bigrams=sorted(lemma_freq_hash.items(),reverse=True,key=lambda x: x[1])# nach value ordnen, absteigend ; list of tuples
		return lemma_freq_list

	def get_types_freq_list(self,token_list):
```

```
                types_freq_hash={}
                for token in token_list:
                        type=token.form
                        if type in types_freq_hash:
                                types_freq_hash[type]+=1
                        else:
                                types_freq_hash[type]=1

        types_freq_list=sorted_freq_bigrams=sorted(types_freq_hash.items(),reverse=True,key=lambda x: x[1])# nach value ordnen, absteigend ; list of tuples
                return types_freq_list

        def get_pos_list(self,token_list):
                pos_hash={}
                for token in token_list:
                        pos=token.pos
                        if pos in pos_hash:
                                pos_hash[pos]+=1
                        else:
                                pos_hash[pos]=1
                pos_list=sorted_freq_bigrams=sorted(pos_hash.items(),reverse=True,key=lambda x: x[1])# nach value ordnen, absteigend ; list of tuples
                return pos_list

        def get_highest_freq_bigrams(self,list_bigrams):
                bigram_hash={}

                for bigram in list_bigrams:
                        if bigram.string in bigram_hash:
                                bigram_hash[bigram.string]+=1
                        else:
                                bigram_hash[bigram.string]=1

                return bigram_hash

        def count_tokens(self,list_tokens):
                return len(list_tokens)
        def read_plaintext(self):
                pass

        def read_tokens(self,file):
                # see CatCG parsed input format
                print "start read_tokens"
                regex_tokens_table = re.compile("(.*)\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)")
                try:
                        f=open(file, 'r')

                        token_list=[] # saves Token objects
                        line_number=1

                        passive_const=False# init
                        last_form=""
                        passives=0
                        passive_forms={}

                        pers_pron=0

                        for line in f.readlines(): # reads too many lines here, why?
```

```
                searching=re.search(regex_tokens_table,line)
              #  print line
              #  print line_number
                if searching==None:
                        print "Line {line_number} in {file} could not be
read".format(line_number=line_number, file=file)
                else:
                        form= searching.group(1)
                        lemma = searching.group(2)
                        pos=searching.group(3)
                        pos_full = searching.group(4)
                        x= searching.group(5)  #  ???
                        y= searching.group(6) #

                        new_token=Token()
                        new_token.form = form
                        new_token.lemma =lemma
                        new_token.pos = pos

                        if len(token_list)<228476: # cut bigger corpus here
                           token_list.append(new_token)

                           # SEARCH PERSONAL PRONOUNS
                         #  if "REO" in pos_full:
                           #    pers_pron+=1

                           # PASSIVE CONSTRUCTIONS CHECK
                          # if "VC" in pos_full and passive_const==True:
                            #   passives+=1
                              # new_passive_form=last_form+" "+form
                              # if new_passive_form in passive_forms:
                               #    passive_forms[new_passive_form]+=1
                              # else:
                                 #   passive_forms[new_passive_form]=1
                           #print "Passive"
                           #print last_form
                           #print form
                           #raw_input()

                          #   if lemma=="ser" or lemma=="estar":
                            #     passive_const=True
                             #   last_form=form
                            # else:
                              #   passive_const=False




                  line_number += 1

              f.close
              #print passives
              #print passive_forms

              #print pers_pron
              #raw_input()
```

```
                        #passive_forms=sorted(passive_forms.items(),reverse=True,key=lambda x:
x[1]) # sort for value (frequency)
                        #f=open("passives.txt",'w')
                        #f.write("total number: "+str(passives)+"\n")
                        #nr=1
                        #for entry in passive_forms:
                          #  f.write(str(nr)+"#"+entry[0]+"#"+str(entry[1])+"\n")
                            # nr+=1
                        #f.close
                        #raw_input()

                    # for token in token_list:
                    #     print token.form
                    #     print token
                    #     raw_input()
                        return token_list
                except IOError: # throw exception if file can't be read and exit
                        print "Error with file(name) {file}".format(file=file)
                        sys.exit()

        def convert_text_into_tokens(self): # for input in plain text
                pass

        def build_n_gramas(self,n,token_list):
                print "start build_n_gramas"
                list_n_gramas =[]

                token_counter =1
                for token in token_list:
                        new_n_gram = N_Gram(n,token)
                        #print token.form
                        print len(token_list)
                        print token_counter
                        # complete the incomplete n_gramas
                        for n_gram in list_n_gramas:
                                if len(n_gram) < n:  # n_gram not completed yet
                                        n_gram.append(token)

                        # add new n_gram to list
                        list_n_gramas.append(new_n_gram)
                        print len(list_n_gramas)
                        token_counter += 1
                for n_grama in list_n_gramas:
                        while len(n_grama) < n:
                                n_grama.append(None)
                return list_n_gramas

        def compare_original_vs_translated(self):
                pass

        def get_all_x_at_position_y(self,x,y,list_n_gramas):
                                                                                #
x=string, y=int
                print "get_all_x_at_position_y started"
                form=x
                position=y
                found_n_gramas=[]
```

```
                for n_grama in list_n_gramas:
                        #print n_grama
                        #print len(n_grama)
                        if n_grama[position] != None and n_grama[position].form==form:
                                for token in n_grama:
                                  # print token.form
                                  pass
                                found_n_gramas.append(n_grama)
                         #   raw_input()

                return found_n_gramas

        def maximum_likelihood_estimation_bigrams(self,list_n_grams,x,y): # MLE

                                                                        # list of N-Grams, string, string
                # "To compute a particular bigram probability of a word y given a previous word x
                 # , we'll compute the count of the bigram C(xy) and normalize by the sum of all the bigrams
                 # that share the same first word. page 123, Jurafsky "Speech and Language Processing"
                # P(Wn|Wn-1)= C(Wn-1Wn) / C(Wn-1)  # simplified equation

                count_x =0 # init
                count_bigram_xy = 0

                for n_gram in list_n_grams:  # N_gram is a list of Tokens
                        if n_gram[0].form==x and n_gram[1].form==y:
                                count_bigram_xy += 1
                         #   print x, " ",y
                         #   raw_input()
                        if n_gram[0].form==x:
                                count_x += 1
                          #  print x , " ", n_gram[1].form
                          #   raw_input()
          #  print count_bigram_xy
           # print count_x
                raw_input()
                if count_x != 0:
                        p_bigram = (float(count_bigram_xy) / float(count_x))
                  #  print p_bigram
                        return p_bigram
                else:
                        return 0 # correct?


        def sentence_generator(self,list_n_gramas):  # PROBLEM: Endless loops
                                                                                # list of N-Gram instances
                # We choose a random value between 0 and 1 and print the word whose interval includes the real value we have chosen.
                # We continue choosing random numbers and generating words until we randomly generate the sentence final token (</s> ; here:None).
                # We can choose the same technique to generate bigrams by first generating a random bigram that starts with <s> (Our n-grams don't have that yet!)
                print "sentence generator started"
                length_n_gram= len(list_n_gramas[0]) # bigram? trigram? quadrigram?
```

```
		start_n_gram=None # init
		start_candidates=[] # init

		#get start n_gram
		  # get start candidates
		for n_gram in list_n_gramas:
		   # print n_gram[0].form
			if n_gram[0].form==".": # start symbol ; we don t have <s>
				start_candidates.append(n_gram)
			 #   start_n_gram = n_gram
			  #  break
		  # choose random start
		random_start= random.randint(0,len(start_candidates)-1)
		start_n_gram=start_candidates[random_start]
	 #  print random_start
	 #  print start_n_gram
	#   raw_input()

		if start_n_gram == None:
			print "Error in sentence_generator(): No value for start_n_gram"

		n_gram_list=[] #init ; list of n-grams that build the sentence
	  # n_gram_list.append(start_n_gram) # start symbol

		# get most probable next n_gram
		next_start=start_n_gram[-1].form

		while next_start not in [".",None]: # end symbol ; </s> in Jurafsky
			counted_n_gramas = self.count_n_gramas(list_n_gramas)
			next_n_gram = self.get_most_probable_n_gram(counted_n_gramas,next_start)
			if next_n_gram[-1] != None:
				next_start = next_n_gram[-1].form
			else:
				next_start = None # set end symbol
			n_gram_list.append(next_n_gram)

			# against endless loops # should be an adequacte algorithm
			print len(n_gram_list)
			if len(n_gram_list) > 20:
				break

		"""
		for n_gram in n_gram_list:
			print n_gram
			print len(n_gram)
			for token in n_gram:
				if token != None:
					print token.form
			raw_input()
		"""

		sentence = "" #init
		for n_gram in n_gram_list:
			#print n_gram
		  #  for token in n_gram:
		  #      if token != None:
		  #          print token.form
		#    raw_input()
```

```
                    # n_gram.pop() # last item is start of next n_gram # deletes instance completely in
n_gram_list

                        if len(n_gram)>1:
                                counter=0# init
                                while counter < (len(n_gram)-1) : # drop last element, beacuse it is the
start of next n-gram
                                        sentence += n_gram[counter].form + " "
                                        counter += 1
                        else: # shouldn't occur normally
                                print "n_gram has only length: ", len(n_gram)
                                sentence += n_gram[0].form

                print sentence
                return sentence # string



        def count_n_gramas(self,list_n_gramas):
                                                                                        # []
                print "count_n_gramas() started"
                counted_n_gramas = {}
         #   print len(list_n_gramas)
         #   print list_n_gramas[3][0].form
                for n_gram in list_n_gramas:
                        n_gram.get_string()
                  #  print n_gram.string
                  #   raw_input()
                        if n_gram.string in counted_n_gramas:  # Problem: n_gram consists of Tokens!
                          #  print n_gram.string
                           # print counted_n_gramas[n_gram.string]
                           # raw_input()
                                counted_n_gramas[n_gram.string][0] += 1
                                counted_n_gramas[n_gram.string][1].append(n_gram)
                        else:
                                counted_n_gramas[n_gram.string]=[]
                                counted_n_gramas[n_gram.string].append(1)
                                counted_n_gramas[n_gram.string].append([n_gram]) # list that holds
instances of n_grams

           #  print counted_n_gramas
            # raw_input()
                return counted_n_gramas # returns hash  { n-gram.string : [counter,list of n_grams] }

        def get_most_probable_n_gram(self, counted_n_gramas,x):

                  # hash   { n-gram.string : [counter,list of n_grams] }, string(1.position)
                # returns n_gram with highest frequency for x at first position
                # DOESNT COMPUTE THE PROBABILITIES SO FAR
                print "get_most_probable_n_gram() started"
                print "x = ", x
                if type(counted_n_gramas) != dict:
                        print "get_most_probable_n_gram() needs a hash as paramter, you passed: " +
type(counted_n_gramas)

                for n_gram_string in counted_n_gramas: # take the first one to init ; Good idea?
                        most_probable_n_gram = n_gram_string[1][0] # first instance of n_gram list
```

```
                        break

                # put counted_n_gramas in order (highest frequency to lowest)
                for n_gram_string in sorted(counted_n_gramas.items(),reverse=True,key=lambda x: x[1][0]): # sort for value (frequency)
                 #  print  n_gram_string[1][1][0][0]
                 #  raw_input()
                        if n_gram_string[1][1][0][0].form==x:  # [list of n_grams][first one][first token (start)].form
                         #  print "new most_probable_n_gram"
                                most_probable_n_gram=n_gram_string[1][1][0]
                         #  print most_probable_n_gram
                                for token in most_probable_n_gram:
                                        if token != None:
                                          # print token.form
                                          pass
                                        else:
                                                print "end"
                 #      raw_input()
                                break

                print most_probable_n_gram
           #  raw_input()
                return most_probable_n_gram # type n_gram (list)


        def write_bigram_freq_to_txt(self, list_freq_bigramas, output_file):
                try: # if all necessary directories exist
                        os.makedirs(os.path.dirname(output_file))  # creates complete directory structure

                except OSError: # if directory structure exists
                        print "Directory structure exists already"

                except IOError:
                        print "IOError"
                        sys.exit()

                f=open(output_file, 'w')
                line_nr=1
                for bigram_freq in list_freq_bigramas: # list of tuples
                        bigram_string=bigram_freq[0]
                        bigram_nr=bigram_freq[1]
                        f.write(str(line_nr)+"#"+bigram_string+"#"+str(bigram_nr)+"\n")
                        line_nr+=1
                f.close

        def write_lemma_freq_list(self, lemma_freq_list,output_file):
                                                # list of tuples (lemma, freq)
                try: # if all necessary directories exist
                        os.makedirs(os.path.dirname(output_file))  # creates complete directory structure

                except OSError: # if directory structure exists
                        print "Directory structure exists already"

                except IOError:
                        print "IOError"
```

```
                f=open(output_file, 'w')
                line_nr=1
                f.write("Total number: "+str(len(lemma_freq_list))+"\n")
                for lemma_freq in lemma_freq_list:
                        #print lemma_freq
                        f.write(str(line_nr)+"#"+lemma_freq[0]+"#"+str(lemma_freq[1])+"\n")
                        line_nr+=1
                f.close
                return 1
        def write_pos_list(self, pos_list, output_file):
                                                # list of tuples (lemma, freq)
                try: # if all necessary directories exist
                        os.makedirs(os.path.dirname(output_file))  # creates complete directory
structure

                except OSError: # if directory structure exists
                        print "Directory structure exists already"

                except IOError:
                        print "IOError"

                f=open(output_file, 'w')
                line_nr=1
                for pos in pos_list:
                        f.write(str(line_nr)+"#"+pos[0]+"#"+str(pos[1])+"\n")
                        line_nr+=1
                f.close
                return 1
        def write_n_gram_to_txt(self,list_n_gramas, output_file):
                try: # if all necessary directories exist
                        os.makedirs(os.path.dirname(output_file))  # creates complete directory
structure

                except OSError: # if directory structure exists
                        print "Directory structure exists already"
                                except IOError:
                        print "IOError"

                f=open(output_file, 'w')
                for n_grama in list_n_gramas:
                        f.write("[")
                        for token in n_grama:
                                if token != None:
                                        f.write(token.form+" ") # 1 space distance to next entry
                                else:
                                        f.write("None")
                        f.write("]")
                        f.write("\n")
                f.close



program=Corpus_Admin()
```

Bibliography

Aarts, J. & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on Computer* (pp. 132-141). London: Longman.

Alsina, À., Badia, T. *et al*. (2002). CATCG : un sistema de análisis morfosintáctico para el catalán in *Procesamiento del Lenguaje Natural*, n. 29 (2002), pp. 309-310.

Baker, M. (1993) Corpus linguistics and translation studies: Implications and applications, in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and technology* (Amsterdam: John Benjamins), 233-250.

Baker, M. (1995) Corpora in translation studies: An overview and some suggestions for future research. *Target* 7 (2): 223-243.

Baker, Mona. (1996). Corpus-based translation studies: The challenges that lie ahead. Harold Somers, ed. *Terminology, LSP and translation*: *Studies in language engineering in honour of Juan C. Sager*. Amsterdam—Philadelphia: John Benjamins, 1996. 175–186.

Baroni, M. and Bernardini, S. (2003). A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora. *Proceedings of Corpus Linguistics 2003*, Lancaster, UK, March.

Baroni, M. and Bernardini, S. (2006) A new approach to the study of tranlsationese: Machine-learning the difference between original and translated text, accepted for publication in Literary and Linguistic Computing.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. and Mazzoleni, M. (2004) Introducing the La Repubblica corpus: a large, annotated,

TEI(XML)-compliant corpus of newspaper Italian, in *Proceedings of LREC 2004* (Lisbon 26-28 May 2004), 1771-1774.

Bernardini, S. and Zanettin, F. (2004). When is a Universal not a Universal? In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they exist?* Amsterdam: Benjamins, pp. 51–62.

Borin, L. and Prütz, K. (2001) Through a glass darkly: part of speech distribution in original and translated text, in W. Daelemans, K. Sima'an, J. Veenstra and J. Zavrel (eds) *Computational linguistics in the Netherlands 2000* (Rodopi, Amsterdam), 30-44.

Blum-Kulka, S. (1986) Shifts of Cohesion and Coherence in Translation, J. House & S. Blum-Kulka (Eds), *Interlingual and Intercultural Communication*, Tübingen.

Clement, R & Sharp, D. (2003). Ngram and Bayesian Classification of Documents for Topic and Authorship. Literary and Linguistic Computing. vol. 18, issue 4 pp. 423-447. Oxford.

Corpas, G. (2008). Investigar con corpus en traducción: los retos de un nuevo paradigma. *Frankfurt, Peter Lang,*

Coseriu, E. (1987a) Gramática, semántica, universales. *Estudios de lingüística funcional*, Madrid; 2., n.132.

Coseriu, E. (1987b) Palabras, cosas y términos*, In Memoriam Inmaculada Corrales, I, Estudios lingüísticos*, Universidad de La Laguna, Sta. Cruz de Tenerife, S. 175-185.
Even-Zohar, I. (1979/1990): Polysystem Theory, *Poetics Today*, Special Issue on Polysystems Studies, Vol. 11, No. 1, pp. 9-26.

Eskola, S. (2002). Untypical frequencies in translated language: a corpus-based study on a literary corpus of translated and non-translated Finnish. In A. Mauranen & P. Kujamäki (Eds.), *Translations Universals - Do They Exist*? Amsterdam/Phildelphia: John Benjamins.

Espunya, Anna. 'Informativeness and explicit linking in the translation of the English V-ing free adjuncts into Catalan'. *Languages in Contrast* 7 (2): 143-166.

Espunya, Anna. 'Is Explicitation in Translation Cognitively Related to Linguistic Explicitness? A Study on interclausal relationships'. *Belgian Journal of Linguistics,* 21(1): 67-86.

Finn, A. and Kushmerick, N. (2003) Learning to classify documents according to genre in *Proceedings of IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*.

Frawley, W. (1984) Prolegomenon to a theory of translation, in W. Frawley (ed.) *Translation: Literary, Linguistic and Philosophical Perspectives* (London and Toronto: Associated University Presses), 159-175.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English, in L. Wollin and H. Lindquist (eds) *Translation Studies in Scandinavia* (Lund: CWK Gleerup), 88-95.

Gellerstam,M. (1996). Translations as a source for cross-linguistic studies. In K. Aijmer, B. Altenberg and M. Johansson (eds) *Languages in Contrast* (Lund: Lund University Press), 53-62.

Granger, S. (2010). Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. In : *Journal of Shanghai Jiaotong University.*

Hansen, S. (2003). The Nature of Translated Text . (Saarbrücken: Saarland University)..

Hermans, Theo (1985). The Manipulation of literature: studies in literary translation / edited by Theo Hermans. Croom Helm. London :

Jantunen, J.H. (2001). Synomymity and lexical simplification in translations: A corpus-based approach. *Across Languages and Cultures* 2 (1), 97-112.

Joachims, T. (1997). Text Categorization with Support Vector Machines: Learning with Many Relevant Features (University of Dortmund: Department of Computer Science).

Joachims, T. (1999) Making large-scale SVM learning practical, in B. Schölkopf, C. Burges, and A. Smola (eds) *Advances in Kernel Methods - Support Vector Learning* (Cambridge, MA: MIT Press)

Kenny, D. (1998). Creatures of Habit? What Translators Usually Do with Words. *Meta*, XLIII(4):515–524.

Kenny, D. (1999). Norms and creativity: Lexis in translated text. PhD Thesis. Mimeo. Manchester: Centre for Translation and Intercultural Studies, UMIST.

Kenny, D. (2001). Lexis and Creativity in Translation: a corpus based study. Manchester: St. Jerome.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics 6(1)*, 97-133.

Kindermann, J., Diederich, J., Leopold, E. and Paass, G. (2003) Authorship attribution with support vector machines. *Applied Intelligence 19*, 109-123.

Kolz, Benjamin, and Giménez, Pau. (2012) Translationese Analizer in Python. To appear.

Koppel, M., Argamon, S. and Shimoni, A. (2002) Automatically categorizing written texts by author gender. *Literary and Linguistic Computing 17(4)*, 401-412.Laviosa, S. (1997): How Comparable can 'Comparable Corpora' be?, *Target*, Vol. 9, No. 2, pp. 289-319.

Laviosa-Braithwaite, S. (1996). Comparable corpora: Towards a Corpus Linguistic Methodology for the Empirical Study of Translation. In M. Thelen & B. Lewandoska-

Tomaszczyk (Eds.), *Translation and Meaning* (Part 3) (pp. 153-163). Maastricht: Hogeschool Maastricht.

Laviosa, S. (1997). The English comparable corpus (ECC): A resource and a methodology for the empirical study of translation. Manchester: UMIST.

Laviosa, S. (1998a). The corpus-based approach: a new paradigm in translation studies. *Meta*, Volume 43, number 4,. 474-479.

Laviosa, S. (1998b) Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta 43(4)*, 557-570.Laviosa, S. (2003). Corpora and the translator. In H. Somers (ed.) *Computers and translation: a translator's guide*. John Benjamins, Amsterdam, pp. 105-117.

Laviosa, S. (2003). Corpora and the translator. In H. Somers (ed.) *Computers and translation: a translator's guide*. John Benjamins, Amsterdam, 105-117.

Malblanc, A. (1968) Stylistique comparée du français et de l'allemand: essai de représentation linguistique comparée et étude de traduction. Paris: Didier,

Malmkjaer, K. (Ed.). (1998). Translation and Language Teaching : Language Teaching and Translation. Manchester: St. Jerome.

Mauranen, A. (2000). Strange strings in translated language: A study on corpora. Maeve Olohan, ed. *Intercultural faultlines: Research models in Translation Studies I: Textual and cognitive aspects*. Manchester: St. Jerome. 119–141.

Mauranen, A. (2005). Contrasting languages and varieties with translational corpora. *Languages in Contrast 5(1)*.

Olohan, M. & Baker, M. (2000). Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation. *Across Languages &* Cultures, 1 (2), 141-158.

Olohan, M. (2001) Spelling out the optionals in translation: a corpus study, in *Proceedings of CL 2001*.

Olohan, M. (2004). Introducing corpora in translation studies. Routledge, London.

Øverås, L. (1998). In search of the third code: an investigation of norms in literary translation. *Meta, 43*(4), 571-588.

Puurtinen, T. (1995). Linguistic Acceptability in Translated Children's Literature. *Joensuu*: University of Joensuu Publications in the Humanities.

Puurtinen, T. (1997). Syntactic Norms in Finnish Children's Literature. *Target*, 9 (2) pp. 321-334. John Benjamins.

Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, 18(4): 389–406.

Puurtinen, T. (2004). Explicitation of clausal relations. In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they Exist?* Amsterdam: Benjamins. 165-176.

Rayson, P. *et al*. (2008). Quantitative analysis of translation revision: contrastive corpus research on native English and Chinese translationese in *Proceedings of XVIII FIT World Congress*, Shanghai, China, August 4-7, 2008.

Santini, M. (2004). State-of-the-art on automatic genre identification. Brighton: ITRI, University of Brighton.

Santos, D. (1995). On grammatical translationese. In Koskenniemi, Kimmo (comp.), paper presented at the *Tenth Scandinavian Conference on Computational Linguistics* (Helsinki).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34(1)*, 1-47.

Tirkkonen-Condit, S. (2002). Translationese – a myth or an empirical fact? *Target*, 14(2): 207–20.

Tirkkonen-Condit, S. (2004). Unique items – over or under-represented in translated language? In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they Exist?* Amsterdam: Benjamins. 177–84.

Toury, G. (1979). Interlanguage and its Manifestations in Translation, *Meta, Translator's Journal,* Vol. 24 No. 2.

Toury, G. (1980). In Search of a Theory of Translation. Tel Aviv.

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

Toury, G. (2004). Probabilistic explanations in translation studies: welcome as they are, would they qualify as universals? In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they Exist*? Amsterdam: Benjamins, 15–32.

Vanderauwera, R. (1985). Dutch Novels Translated into English: The Transformation of a "Minority"Literature, Amsterdam, Rodopi.

Venuti, L. (2004) .The Translation studies reader. Routledge. New York.

Vinay, J.P. and Darbelnet, J. (1958). Stylistique comparée du français et de l'anglais. Paris: Didier.

Vinay, J.P. and Darbelnet J. (1995). Comparative stylistics of French and English: a methodology for translation / Jean-Paul Vinay, Jean Darbelnet ; translated and edited by Juan C. Sager, M.-J. Hamel. John Benjamins. Amsterdam.