On grammatical translationese

Diana Santos

Natural Language Group
INESC, Apartado 1306, R.Alves Redol, 9
P-1000 Lisboa, Portugal
and
Department of British and American Studies
University of Oslo, Postboks 1003 Blindern
N-0315 Oslo, Norway
diana.santos@{inesc.pt, iba.uio.no}

1. Introduction

In most published work on bilingual data from aligned corpora (variously called parallel corpora, translated corpora, bi-texts or corpora of parallel texts), it is the lexicon that has been studied (cf. Klavans and Tzoukermann (1990), Church and Gale (1991), Marinai et al. (1991), Isabelle et al. (1993)). In this paper, I am concerned with corpus-based contrastive studies of grammatical features.

The term "translationese" has been invoked in connection with the distribution of lexical items. To my knowledge it was first used by Gellerstam (1986) to describe vocabulary differences between original Swedish texts and Swedish texts translated from English. In principle, there is no reason why it should not be used to describe grammatical differences between original and translated texts as well, and, in fact, it has been used loosely in this way by Gellerstam himself (see below). Johansson and Ebeling (1994) and Hasselgård (1995) have also mentioned this question with respect to comparative corpus-based studies. My intention here is not to argue that such a phenomenon exists, but to propose a first systematization of general cases, as well as discuss some problems in its delimitation. In addition, some actual cases are discussed.

Having engaged in a corpus-based contrastive semantic study of Portuguese and English, I have realized that one must look at contrasts at two levels:

- a micro-level, where individual instances count;
- a macro-level, where it is frequencies and general features that are at stake.

Translationese is no exception: it is observable both in frequency differences and in individual translations, as will be illustrated below.

2. On the notion of translationese

Gellerstam (1986) looked for statistically significant differences between texts in original Swedish and in Swedish texts originating from translation from English.

He was, however, cautious to note that not all differences were attributable to the source language.

Thus, he observed that some differences had to do with cultural differences, for example, the relative frequencies of *beer* and *coffee*. This is a difference that should not be classified as translationese, because, as far as I know, most translators would not translate *have a beer* by *have some coffee*. In fact, this is precisely a case where translationese does not occur, since the difference between the two kinds of texts is maintained. (And, I would claim, it would disappear - or actually become inverse - if the original Swedish texts were about British or American society and if the English texts were depicting a Scandinavian setting.)

Then, Gellerstam also noted that, due to differences between English and Swedish in the style of describing dialogues, phrases conveying the way (or tone) in which direct speech was uttered were much more frequent in translated than in original Swedish; e.g. (the Swedish translations of) *he shrugged his shoulders* or *he said smiling*. Incidently, direct speech presentation is also one major source of discrepancies between English and Portuguese. Gellerstam argued that direct translations of words like *shoulders* or *smiling* in these contexts should not be considered instances of translationese. I agree, but I maintain that the phrases in which they occur are very good cases of this phenomenon. In fact, it seems arbitrary of Gellerstam to consider only lexical cases.

On the other hand, it was surely a way to reliably extract possible candidates. (Below, I shall discuss the added difficulty of handling grammatical features.) Gellerstam's method was as follows: he restricted the analysis to lexical items above a certain frequency (greater than 100) and a certain distribution (appearing in more than half of the texts), whose occurrence in translated texts was greater than 70%. Then, he was able to arrive at a number of generalizations. It is important to note that they were not exhaustive, i.e., not every case could be subsumed under one explanation or the other.

The generalizations he makes after a careful study of his data are that Swedish texts translated from English share the following features:

a) decrease in Swedish colloquialism (or dialectal features);

b) lexical choice of "standard translation" (even though rare in a Swedish text);

c) increase in English loanwords;

d) use of false friends in international vocabulary;

e) restructuring semantic fields according to the source language, e.g. with evaluative adjectives, or verbs of feeling.

Cases c) and d) are in fact two faces of the same coin: they constitute uses of English values in a Swedish context, be it by inventing a new word (the easiest way being "borrowing" it), or by using an already existent one but disregarding its Swedish meaning. Case e), in turn, only imports the distribution of values in the system but uses the Swedish elements. These three cases are in decreasing order regarding overtness of English influence, from the obvious case of using foreign lexical items to the much more complex and subtle case of merely borrowing paradigmatic relationships among items.

For the sake of completeness, let me mention that Gellerstam also mentions, in passing, some syntactical matters, namely the present participle construction in adjectival function, which is weird in Swedish though common in English; e.g. *the approaching car*.

In what follows, I will try to apply the notion of translationese, as well as some of the special cases distinguished, to grammatical features.

3. Grammatical translationese

I take the general term "translationese" to denote the influence of properties of the source language in a translated text in a target language.

I am interested in studying translationese from a grammatical point of view. In grammar description, I assume that the factors at play are discrete units of a finite number, usually small. In what follows, I will denote them by capital letters, and have an arrow stand for the translation relation in terms of competence, i.e., A -> A' stands for "A can be translated by A'". ("A" represents both the formal feature and its meaning. Obviously, no claim of uniqueness is being made here: the formal feature corresponding to A may have several meanings.)

Taking the opposite path from Gellerstam, who proceeded bottom-up, I first try to sort out some general cases of grammatical translationese, prior to looking for their actual occurrence. These seem to be good candidates:

a) <u>A -> B,C,D.</u> The case where there are a number of grammatical markers in the target language with (roughly) the same "meaning" or use as A in the source language. In parallel with the standard translation of lexical items, I predict that one of the target translations would be preferred, and would thus have a significantly higher frequency than in original text.

b) <u>A + obligatory B -> A' + optional B'.</u> If there is an optional marker in the target language (standardly) corresponding to an obligatory one in the source language, it is to be expected that its frequency will increase in translationese.

c) <u>vague (A,B) -> A', B'.</u> When a source item is vague between two (or more) cases, and the target language cannot preserve such vagueness, we face a translation mismatch (see Keenan (1978) for ample illustration and discussion, and Kameyama et al. (1991) for the concept definition). Strictly speaking, this is a case where the source language is unable to influence the translation, and so, no consistent pattern can be expected. However, if in the target language one of A' and B' is felt to be marked, the translator's choice should be the less marked, and translationese could show in the relative frequencies of A' and B'.

d) <u>compact (A,B) -> A' + B', A',B'.</u> When there is a device which packages more than one meaning in a compact expression, while the target language must express them separately, more often than not only one of the features is conveyed. Still, another form of translationese occurs in that there will be more conjoined cases (A' + B', i.e., cases where both A' and B' are expressed) than in original text.

In order to test for actual instances of these cases, I rely on data from a detailed study of tense and aspect in translations between English and Portuguese. See Santos (1994a) for a first sketch. Some quantitative data are furnished below:

	Words	Sentences	Tensed clauses
Portuguese original text:	25174	1494	3257
English translation:	27972	1459	
English original text:	26060	1628	3744
Portuguese translation:	23262	1861	

While this parallel corpus is too small and completely unbalanced (one single author in each language; one single translator for the Portuguese translated text; only narrative text), it was large enough to at least pose the problems that eventually led to this study. In addition, and compared to Gellerstam's study, who only had access to Swedish texts, this set up allowed me to test the source language texts for actual influence. It allowed me, therefore, to study also micro-translationese.

I proceed to discuss some cases, noting that one reason why I first discussed the abstract general cases is that it is in general not possible to give neat examples of one case only. Grammatical categories have a number of uses and, as will be seen in the examples below, often illustrate more than one case of translationese.

3.1 The progressive

In a monolingual analysis of Portuguese, one could describe three aspectualizers as progressive: *estar a* + infinitive, *andar a* + infinitive, and *ir* + gerund. A coarse characterization is as follows: *estar* is the unmarked case, *ir* conveys graduality and *andar* habituality. Possible examples are *Ele está a trabalhar muito* (he is working hard (at this moment)) versus *Ele anda a trabalhar muito* (he is working hard (lately)), and *O sol está a aquecer a casa* (the sun is heating the house (at this moment)) versus *o sol vai aquecendo a casa* (the sun is heating the house (which is becoming warmer)).

Now, the English progressive is much more frequent than the (sum of the) three Portuguese progressives, because it is also used for cases where the

Portuguese past tense Imperfeito is used, most notably signalling co-temporality and a series of events. Compare these (real) examples (discussed more fully in Santos (1994b)):

Não sabia o que fazia$_{\mathrm{IMP}}$. -> *I didn't know what I was doing*

Torcia$_{\mathrm{IMP}}$ *as mãos*. -> *He was twisting hands*

What I foresee in this case is that the frequency of the progressive in translated Portuguese (coming from English) will be higher than in Portuguese original texts, and that one of the three constructions will have a more pronounced rise in frequency, with possibly a decrease in the others.

All instances of the progressive were counted. The numbers relevant to the discussion (OE - original English, OP - original Portuguese, TEP and TPE, translation from English to Portuguese and vice versa) follow (The order of the Portuguese progressives is *estar*, *ir*, *andar*):

OE	58	TEP	20 (17, 3, 0)
TPE	98	OP	30 (23, 5, 2)

A closer look reveals that the instances of progressive in English translated from Portuguese, TPE, come mainly from Imperfeito (85 cases), while 12 render Portuguese progressive with *estar* and one with *ir*.

Given the small number of cases found, there is no evidence for the first kind of translationese (a), whereby the distribution of the Portuguese progressives in translation would be significantly influenced by English, although I am convinced that in a larger corpus the absence of *andar* in TEP would become significant.

However, there certainly is a mark of translationese in English, namely, the marked increase in the frequency of the English progressive, because of translation from Imperfeito, which can be classified as a case of (b).

3.2 The present perfects

In order to describe habit up to the present time in English, one uses the present perfect with obligatory marking of repetition (or of the period up to now); cf. Leech's (1971:34) *I've always walked to work* (described as "Habit in a period leading up to the present") and Mittwoch's (1988) *John has played the piano since he was five*.

On the other hand, the Portuguese Pretérito Perfeito Composto (PPC) conveys precisely repetition in an interval up to now, without need of further specification; cf. Boléo (1936:5, my translation): *what renders this tense expressive in its conciseness and Portuguese character is exactly the ability to express duration or repetition of an action (or state...) without a single additional word*, i.e., without any supplementary device.

One would thus expect that, when the present perfect with such meaning/use is used in English, an additional (unnecessary) specification will be found in Portuguese. Conversely, translations of PPC into English will either require a *lately* or *recently* overt specification (translating the optional *ultimamente*), or fail to convey one part of the meaning.

Going through the 52 occurrences of the present perfect in the OE corpus, only two were clear instances of this use, and one had been translated by the PPC, another by Perfeito. The (obligatory) specification of the interval was obviously explicit in the translation.

On the other hand, none of the 3 instances of PPC found in the OP corpus had an interval specification. One was rendered in English by the present perfect progressive and two by the simple present perfect. These latter, I believe, failed to convey that the situation depicted continued till the narrative now, being clear instances of case (d):

E eu disse [...] que tens trabalhado$_{\mathrm{PPC}}$ *muito e até tens estudado*$_{\mathrm{PPC}}$ *com o Padre Manuel* -> *And I said [...] that you've worked much and have even studied much with Padre Manuel*

The predictions in this case were born out, even though the results are not statistically significant.

As far as the adverbs *já* and *already* are concerned, also closely tied to the perfect, *já* + Perfeito is obligatory to convey that something was performed "according to plan", while *already* is optional (if not unrelated), given that the English present perfect is appropriate to the extent that the event *was foreseen, planned, striven after, wished for, etc.; in sum, to the extent that it has a preparatory stage* (Sandström, 1993:122).

One would thus foresee that the use of *already* (with either present perfect or past simple) in translated English would outnumber it in OE. The results, displayed below, seem to confirm this prediction, illustrating thus a case of pure (b) type of translationese:

OP *já* + Perfeito	6	TPE	*already*	4
TEP *já* + Perfeito	3	OE	*already*	0

Only preverbal *já* and *already* were considered. The use of *já* with Imperfeito, or of *already* with past progressive, which is much more frequent, was not considered, since it arguably is not related to the perfect at all.

3.3 Differences in aspectual class

Here, I will confine myself on describing some complex aspectual classes in the two languages which may require partitioning in their translation, given rise to various kinds of translationese. I would claim that this is not as much a property of individual lexemes as of global lexical organization, which is crucially related to grammar. No corpus investigation has (yet) been done to check whether this partitioning ever takes place in real translations, but phenomena of this sort may be an explanation for the common increase in the number of clauses in translation, noted in Santos (1994a) typology of clause mismatches.

As far as I know, the question of compactness is the best documented one: Compactness in one language will tend to be expressed by only one of the two conjuncts in its translation, except if the two (or more) meaning pieces are equally crucial for understanding. Compactness is called "event conflation" in Talmy (1991). One case that has been especially discussed is that of verbs of motion also expressing manner (typical of languages like English) translated into verbs expressing only direction (typical of languages like Portuguese); cf. Talmy (1985) for the relevant typology.

Slobin (1994) presents the following real example (in published translation from English into Spanish): *He strolled across the room to the door* -> *Se dirigió a la puerta* ('He went to the door'). Still, he observes that locative detail is more common in Spanish translations than in Spanish original text, which constitutes evidence for the (d) kind of translationese.

This sort of example is worth discussing in more detail because it involves a special sort of (English) predicate that denotes the conjunction of (i) the action towards a goal, (ii) the attainment of the goal <u>and</u> (iii) the resulting state: Compare *He ran out of the house,* which could be rendered in Portuguese by *Ele saiu da casa a correr* ('He exited running') or *Ele correu para fora da casa* ('He ran towards the outside'). However, neither sentence conveys exactly the three pieces of information, since the first only states that when he went out he was running - not that he started running in order to exit, and the second sentence does not actually state that he went out, as can be seen by the possible continuation *mas não conseguiu sair* ('but did not manage to go out'). One would probably need something like *ele correu para fora da casa e saiu* ('He ran outwards and exited') in order to express explicitly both the action and the resulting state, which would constitute another case of translationese of the (d) kind.

Another aspectual class that may bring a problem for translation are acquisitions in Portuguese, i.e., verbs which denote both the inception and the resultant state (distinguishable by tense or context). So, the verb *conhecer* used in two different past tenses will be rendered by the simple past of two different English verbs, *meet* and *know*: *Ele conheceu-o* -> *He met him, Ele conhecia-o* -> *He knew him.*

In cases where the Portuguese text does not distinguish between inception and state, the English translator is forced to select one arbitrary interpretation, as is the case in *Conhecer uma pessoa como ele é sempre uma aventura*, that could be rendered as *Meeting* (or *knowing*) *a person like him is always exciting*. This is a clear example of (c). Only comparing the distribution of *meet* and *know* in non-tensed clauses in translated and original texts can one identify translationese.

Finally, it is interesting to note the fairly frequent existence of translation pairs involving the English verb *be* where the Portuguese past perfect of a change of state verb is employed as translation, as illustrated by *He was quiet now* -> *Ele acalmara* ('he had calmed down'). This can be described by noting that *be* is vague, while Portuguese makes explicit both the inception (through the lexical item) and the resulting state (through the use of the past perfect). We have thus a case where the target language is able to translate vague(A,B) by compact(A',B'), thus being able to avoid translationese of the (c) kind. However, this difference may result in a conspicuously higher frequency of compact constructions in translated texts.

4. Conclusion

Summing up, by considering some differences in the tense and aspect systems of English and Portuguese, I was able to illustrate several kinds of translationese, while at the same time demonstrating that a relatively fine analysis was required. On the other hand, by comparing with the source text, I could identify cases of influence of the source language in the translated text with a very small corpus. An analysis of considerably larger amounts of text would be required if only target language texts were used.

The most important message of this paper, however, should be that, if one wants to rely on aligned corpora of actual translations for natural language processing, and/or for general contrastive studies, which seems clearly fashionable these days, it is fundamental to pay attention both to questions of translation quality of the individual translations and to general properties of the translated texts.

One cannot take it for granted a priori that the translated text is a good representative of the target language. On the contrary, one has to acknowledge, with Baker (1993), the specificity of texts which originated from a different language by translation. However, an apparently paradoxical property should be mentioned: If translationese stems from the fact that different languages have different systems, it is also related to language closeness: the closer the languages the larger the quantity of false friends and cognates, both in lexicon and in grammar. The closer the languages the easier to translate the surface and not the content, and therefore the more possible to "level" the two languages, i.e., even out their differences. (Translationese was also described by Gellerstam as a levelling of the two languages involved.)

With this paper, I hope to have contributed to the identification of translationese in the realm of grammatical mechanisms, whose considerably greater complexity is due to the widely known fact that any grammatical device has a variety of meanings and performs a varied number of functions in the language.

Acknowledgments

I am grateful to Stig Johansson, Hilde Hasselgård and Jan Engh for reading and commenting on an earlier version of this paper. The work reported here was supported by a PhD scholarship granted by JNICT (Junta Nacional de Investigação Científica e Tecnológica).

References

Baker, Mona. "Corpus Linguistics and Translation Studies: Implications and Applications", in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*, Benjamins, 1993, pp. 233-50.

Boléo, Manuel de Paiva. *O Perfeito e o Pretérito em português em confronto com as outras línguas românicas,* separata de *Cursos e Conferências* da Biblioteca da Universidade de Coimbra, vol. VI, 1936.

Church, Kenneth W. & William A. Gale. "Concordances for Parallel Text", *Using Corpora: Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research* (Oxford, September 29 - October 1, 1991), pp. 40-62.

Gellerstam, Martin. "translationese in Swedish novels translated from English", in Lars Wollin & Hans Lindquist, *Translation Studies in Scandinavia*, CWK Gleerup, Lund, 1986, pp. 88-95.

Hasselgård, Hilde. "Some methodological issues in a contrastive study of word order in English and Norwegian", in Bengt Altenberg & Karin Aijmer (eds.), *Languages in Contrast: Text-based cross-linguistic studies*, Lund University Press, in print.

Isabelle, Pierre, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliot Macklovitch, François Perrault, Xiaobo Ren & Michel Simard. "Translation Analysis and Translation Automation", *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, *TMI'93* (Kyoto, July 14-16, 1993), pp. 201-17.

Johansson, Stig & Jarle Ebeling. "The English-Norwegian Parallel Corpus: Introduction and Applications", paper presented at The XXVIII International Conference on Cross-Language Studies and Contrastive Linguistics (Rydzyna, Poland, 15-17 December 1994).

Kameyama, Megumi, Ryo Ochitani & Stanley Peters. "Resolving Translation Mismatches With Information Flow", *Proceedings of the 29th Annual Meeting of the ACL* (Berkeley, 18-21 June 1991), pp. 193-200.

Keenan, Edward L. "Some Logical Problems in Translation", in F. Guenthner & M. Guenthner-Reutter (eds.), *Meaning and Translation: Philosophical and Linguistic Approaches*, Duckworth, 1978, pp. 157-89.

Klavans, Judith & Evelyne Tzoukermann. "The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries", in Hans Karlgren (ed.), *Proceedings of COLING'90*, Vol. 3, pp. 174-9.

Leech, Geoffrey N. *Meaning and the English Verb*, Longman, 1971.

Marinai, Elisabetta, Carol Peters & Eugenio Picchi. "Bilingual Reference Corpora: A System for Parallel Text Retrieval", *Using Corpora: Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research* (Oxford, September 29 - October 1, 1991), pp. 63-70.

Mittwoch, Anita. "Aspects of English Aspect: On the interaction of perfect, progressive and durational phrases", *Linguistics and Philosophy 11*, 1988, pp. 203-54.

Sandström, Görel. "*When*-clauses and the temporal interpretation of narrative discourse", PhD dissertation, Department of General Linguistics, University of Umeå, Report nr. 34, DGL-UUM-R-34, May 1993.

Santos, Diana. "Bilingual alignment and tense", *Proceedings of the Second Annual Workshop on Very Large Corpora* (Kyoto, August 4th, 1994), extended version: INESC Report AR/10-94.

Santos, Diana. "Imperfeito: a broad-coverage study", *Actas do X Encontro da Associação Portuguesa de Linguistica* (Évora, 6-8 de Outubro de 1994).

Slobin, Dan I. "Two Ways to Travel: Verbs of Motion in English and Spanish", in M. Shibatani & S. A. Thompson (eds.), *Essays in Semantics*, Oxford University Press, 1994.

Talmy, Leonard. "Lexicalization patterns: semantic structure in lexical forms", in T. Shopen (ed.), *Language typology and semantic description, vol.3: Grammatical categories and the lexicon*, Cambridge University Press, 1985, pp. 57-149.

Talmy, Leonard. "Path to Realization: a Typology of Event Conflation", *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society*, 1991, pp. 480-519.