Monolingual comparable corpora and parallel corpora in the search for features of translated language

Silvia Bernardini

University of Bologna

Summary

For almost two decades now, mainstream corpus-based research in descriptive translation studies has focused on the computerized analysis of translated and comparable non-translated texts (so-called monolingual comparable corpora). These have been suggested to better serve a target-oriented, descriptive approach to translation research than parallel corpora, which in turn have come to be perceived as somewhat backward-looking, favouring source-oriented, equivalence-focused, at worst anecdotal, observations. In this article I argue that comparable and parallel corpora in fact offer complementary perspectives on translation norms/universals, such that neither would suffice in isolation to shed full light on this complex research topic. The point is illustrated with reference to two case studies, the first dealing with borrowing in technical translation, the second with phraseological regularities in fiction translation.

1 Introduction

The approach to research on translation using electronic text collections, known as Corpus-Based Translation Studies (henceforth CBTS) has, since its very start, programmatically focused on the monolingual comparison of originals and translations in the same language, or monolingual comparable corpora (MCCs). In the words of Baker (1993: 237),

> [t]he move away from source texts and equivalence is instrumental in preparing the ground for corpus work because it enables the discipline to shed its long-standing obsession with the idea of studying […] one translation compared to one source text at a time […].

This followed from the target-oriented, descriptive approach to translation developed by Toury (1980, 1995), which proposed that the search for norms and laws of translation be based on the analysis of bodies of translated texts within their contexts of fruition (i.e., the target culture). Within CBTS, comparable non-translated texts were taken to represent the target culture, and to form the baseline against which the peculiarities of translated language (hypothesized to be universal) would emerge. MCCs were hailed as “the major methodological advance associated with corpus studies” (Pym 2008: 321-322). By comparison, parallel corpora, i.e. collections of translated texts aligned to their source texts, appeared to perpetuate a source-oriented, prescriptive view of translated texts as almost by definition inferior to their source texts. The search for translation universals thus took the form of quantitative comparisons − e.g. ratio of content to function words (Laviosa 1998), presence of target-specific lexical items (Tirkkonen-Condit 2004), number of contractions (Olohan 2003) − opposing translated and non translated texts in the same language. Differences found were interpreted in terms of e.g. a tendency for translated texts to be more/less explicit, unambiguous, repetitive, plain etc., than texts similar along all dimensions, except for the fact that they originated within the target culture instead of being imported from a different one by means of translation. While several parallel corpora as well as corpora combining the two corpus typologies exist and have been used in influential work in translation studies and contrastive linguistics (see e.g. Kenny 2001, Johansson 2007), within

CBTS there has always been an observable tendency to favour MCCs, and to relegate parallel corpora to at best an ancillary role.

The aim of this article is to question this state of affairs, to reflect on the appropriateness of MCCs for identifying typical features of translated language, and to suggest that a methodology combining different types of corpora is not only advisable, but critical if one is to avoid methodological pitfalls. To this effect, two case studies dealing with translation from English into Italian will be presented, one focusing on anglicisms in technical translation (Section 2), the other on collocations in fiction translation (Section 3). I shall conclude in Section 4 by discussing the results of these studies in terms of the methodological implications they have for CBTS at large.

2 Case study 1: Anglicisms in technical translation

The aim of the first case study presented is to investigate use of calques and borrowings in translation of computing texts from English into Italian. This technical domain is one in which the influence of English is especially strong, in Italian as well as other languages (see e.g. Piqué-Angordans et al. 2006). Both texts originally written in Italian by Italian authors and texts translated from English into Italian are likely to contain many words borrowed from English, this tolerance of foreign-looking lexis being one of the typical features of this domain. By comparing originals and translations, however, one could ascertain whether translations contain more/less or a similar number of such words. If translators were found to use more English words than Italian writers this could be taken as evidence of interference from the source text. If, on the other hand, translators used fewer English words than writers, one could hypothesize a normalizing tendency, i.e. a preference for the more normal or typical options afforded by the target language system. Investigation of anglicisms in technical translation from English thus appears to be an ideal test bed for shedding light on Toury's (1995) hypothesized laws of translation, i.e. the *law of interference*, stating that ST linguistic features are transferred onto the TT, and its converse, the *law of growing standardization*, stating that more frequent target language options are preferred in translation.

The corpus used in this study contains documentation about the *Perl* programming language. While most communication about this popular programming language takes place in English, as is the case for programming in general, Italian Perl enthusiasts are trying to make it more familiar to newbies by producing documentation in Italian and/or translating it from English. Hence the interest for our purposes: the Perl corpus contains texts belonging to a single genre, dealing with a specific topic, produced and translated by members of the same community (area experts, not linguists). The structure of the corpus is tripartite: originals in Italian (**PERL-OR-IT**), translations from English (**PERL-TR-IT**), and the latter's English source texts (**PERL-OR-EN**). In other words, a combination of a MCC and a parallel corpus. Table 1 provides some data about the texts in the corpus. The actual titles and authors/translators are not included since they are unlikely to be familiar to the audience (interested readers who would like more details or indeed the corpus itself, can contact the author). This corpus, like the one used in the case study described in Section 3 below, was Part-of-Speech tagged and lemmatized using the Tree tagger (Schmid 1994), indexed with the Corpus WorkBench (CWB, Christ 1994), and queried using the companion Corpus Query Processor (CQP).

Table 1. PERL: basic corpus data

	PERL-OR-EN	PERL-TR-IT	PERL-OR-IT
Tokens	298,346	321,405	305,537
Types	18,639	22,768	22,495
Texts	43	43	89
Writers	19	---	33
Translators/revisors	---	16	---

For the purposes of this study, only overt, unadapted borrowings are focused upon, i.e. words imported from English recently and preserving their alien aspect. Clearly, other ways of borrowing lexis from a foreign language exist (see e.g. the typology in Gottlieb 2004), which cannot be dealt with here for reasons of space (for a more complete treatment of anglicisms in translated technical Italian see Bernardini and Ferraresi forthcoming). The search for borrowings in corpora cannot easily be automated (Furiassi and Hofland 2007), and the procedure used here is largely manual.

Focusing on the monolingual comparable dimension of the analysis, *keyword lists* were produced to identify words (potentially including borrowings) that are significantly more frequent in translated than in original texts, and vice versa. The rationale for adopting this procedure as a starting point is that if a borrowing were used by translators and authors alike, it would not be relevant for this study. Frequency lists were obtained for all word forms in the two Italian subcorpora, and they were then compared using the Log-Likelihood statistic (following Rayson and Garside 2000, henceforth LL), using each subcorpus as a reference for the other.

The top 100 words (in LL order) with frequency equal to or greater than 5 were examined and all English-looking words were retained for further analysis. While arbitrary, this cut-off point was necessary to keep the extent of manual perusal manageable. Several English words were identified among the top keywords in the two lists: these constitute the potential "key" borrowings from the two subcorpora (Table 2).

Table 2. Candidate English key-borrowings in translations and originals

English borrowings in PERL-TR-IT			**English borrowings in PERL-OR-IT**		
Word	**Fq**	**LL**	**Word**	**Fq**	**LL**
package	357	178.4	script	472	131.0
match	174	148.2	expression	104	130.7
char	70	94.6	regular	103	123.2
filehandle	234	87.7	array	882	118.7
locale	115	83.7	overloading	59	75.0
require	112	83.3	print	920	54.1
unpack	60	72.3	reference	88	50.7
socket	102	66.9	matching	69	37.5
shift	208	66.9	hello	24	34.1
local	140	65.6			
buffer	62	63.7			
point	55	54.9			
record	70	54.4			

long	50	53.4
pack	59	51.7
thread	282	50.5
encode	36	48.6
pipe	99	46.5

As can be observed in Table 2, more candidates are found in the translated corpus list than in the original one. One could infer from this that Italian translators use borrowings more liberally than do Italian authors, yet these results are just a starting point for the analysis. Several checks have to be carried out to confirm them. First, one has to ascertain that a certain expression is found in at least two texts written by different authors, to rule out the possibility that it is idiosyncratic to a single individual; in the case of translations this check is not deemed necessary since all translated texts are the product of at least a translator and a reviser, working independently of each other. Then one has to check that the words in the lists are indeed used within "normal" Italian text (and not e.g., as part of a quotation, of untranslatable code text, etc.), and finally that the high frequency of a given English word in one subcorpus is indeed likely to be due to a preference by writers/translators for an anglicism and not to some other reason: in other words, there must be a contextually appropriate alternative Italian expression that could have been used in place of the English one, and was not.

Close inspection of both monolingual and parallel concordances makes it possible to exclude from the lists words occurring mainly within (untranslatable) examples of code text (*char*, *filehandle*) and words that are more frequent in one list for topic-related issues (e.g., *locale* or *encode* from the translated list, which refer to specific topics not covered to the same extent in original texts, and for which no equivalent Italian word exists). For the remaining candidates, an Italian equivalent is searched for in the Italian Perl subcorpora (original and translated) and on the web, to confirm that writers and translators using a borrowing *did* have a choice to use a native equivalent. In most cases it is found that the English word is in fact the only option available to express the concept in question, even in texts addressing a lay audience: this is the case with *socket*, *buffer*, *record*, *thread* and *pipe*. All these words have entries in the Italian *Wikipedia*, where they are defined and used as if they were standard Italian words. In some instances a literal translation is provided, but this is not always the case, cf. the case of *pipe*:

> Nei sistemi operativi una **pipe** è uno degli strumenti disponibili per far comunicare tra loro dei processi. (from Wikipedia: Pipe)
>
> In operating systems, a pipe is one of the tools that make processes communicate with each other. (my translation)

If there is no obvious alternative to the use of the term *pipe*, its presence in the corpus is indicative of topic rather than translator/writer preferences, therefore these cases are disregarded. Finally, parallel concordances are browsed to find out whether alternative translations exist for the words borrowed. Notice that a dictionary is no use in this case, since we are dealing with contextually restricted technical terms. This detailed analysis leads to the exclusion of several borrowings, and to the trimming of the list in Table 2. Table 3 lists actual borrowings, i.e. words for which an alternative attested in this specialized field exists in Italian.

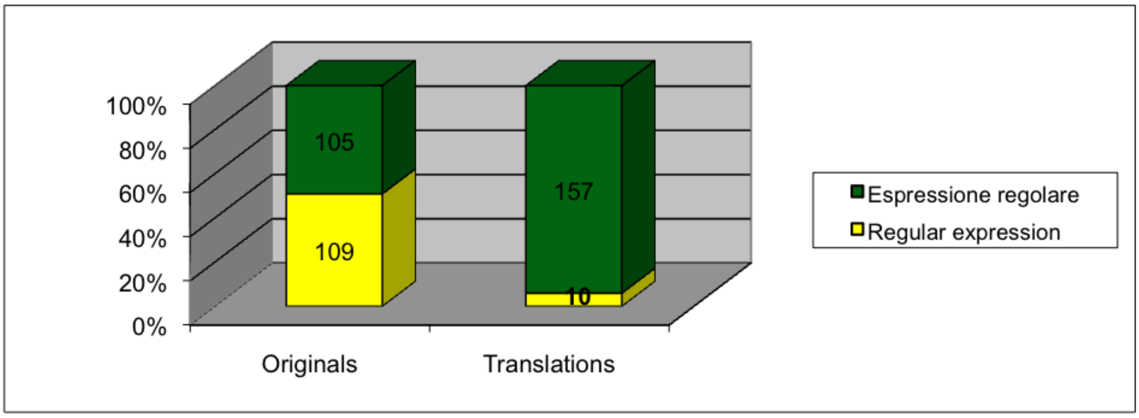
Table 3. Relevant borrowings in translations and originals

English borrowings in PERL-TR-IT		**English borrowings in PERL-OR-IT**	
Word	**Attested Italian alternative**	**Word**	**Attested Italian alternative**
package	pacchetto	expression	espressione
		regular	regolare
		reference	riferimento
		hello	ciao

The situation is reversed with respect to the initial results (18 potential borrowings in translations vs. 9 in originals). The only borrowing seemingly favoured by translators over the Italian alternative term is *package*, which occurs 357 times in translations and 81 times in originals. However, the Italian equivalent pacchetto is also much more frequent in translations than in originals (453 vs. 96 solutions), suggesting once again that we are observing a topic-related difference (translated texts deal with *packages/*pacchetti more than native Italian texts).

Moving to the non-translated subcorpus, the first two key-borrowings are in fact part of the same phrase, namely *regular expression/*espressione regolare. If we add the frequencies of the English borrowing and of its Italian alternative, we get the number of times that the notion is referred to explicitly in the two subcorpora: 167 times in the translated subcorpus and 214 times in the non-translated subcorpus. In the overwhelming majority of cases (94%) translators opt for the Italian term, while writers use the English borrowing or the Italian equivalent to the same extent (51% vs. 49%; see Figure 1).

Figure 1. Regular expression vs. espressione regolare in originals and translations



Figures 2 and 3 present data for *reference/*riferimento and *hello/*ciao (used in the example sentence *Hello world*) that confirm the pattern observed in the case of *regular expression*. Where an Italian alternative is available, translators show a very clear preference for it (over 90% of the total). Writers also use the Italian term in a majority of cases, but in over 30% of the total they opt for the English word, again showing a less clear-cut preference for the Italian word over the English one.

Figure 2. Reference vs. riferimento in originals and translations

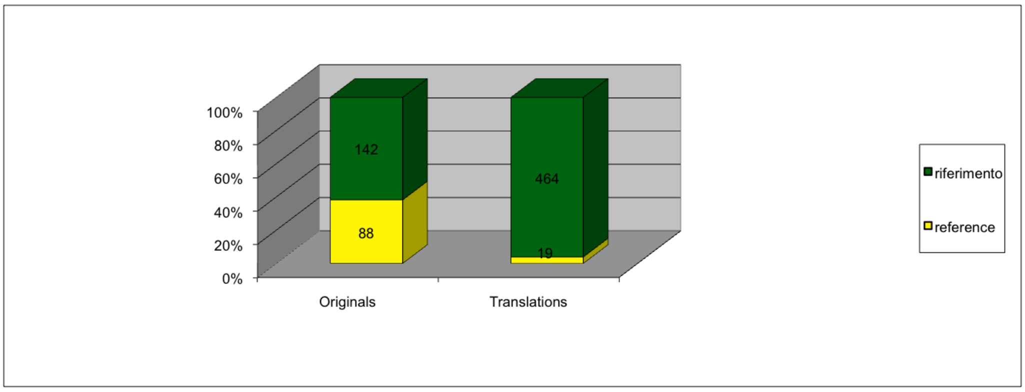
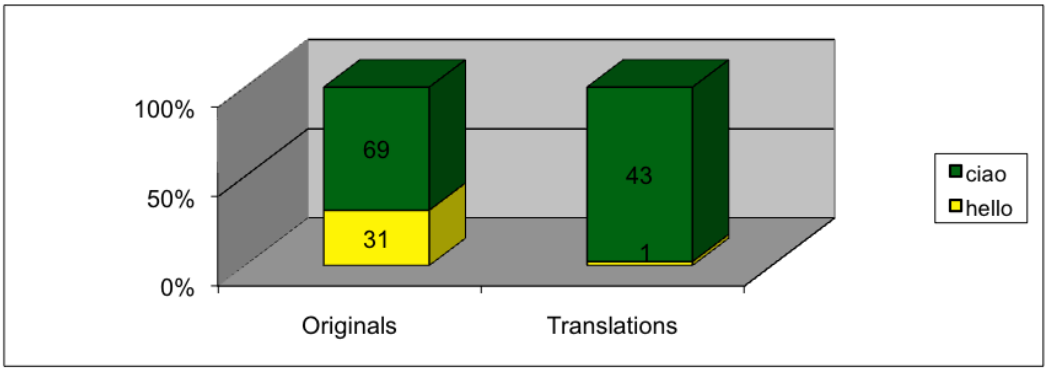


Figure 3. Hello vs. ciao in originals and translations



Taken together, the results just presented suggest that Italian translators of programming documentation seem less comfortable with anglicisms than technical writers of comparable texts: compared to the former, the latter use more unadapted borrowings when alternative Italian renderings are available. In turn, this seems to indicate a trend toward normalization in this translation setting, or that the "law of growing standardization" predominates over the "law of interference", to use Toury's (1995) terms, a trend that is also confirmed by results obtained for other types of anglicisms (adapted borrowings, semantic loans and morphosyntactic calques, see Bernardini and Ferraresi forthcoming).

This case study suggests that a parallel comparison is essential to disentangle variables even in a register-controlled study with near-experimental conditions. If monolingual and parallel concordances, as well as other sources (e.g. the web), had not been checked manually, preliminary results from the MCC (cf. Table 2) would have led us to conclude that technical translators borrow more liberally than authors, while the opposite is in fact the case.

3 Case study 2: Collocations in fiction translation

While in case study 1 I looked at (voluntary) technical translation, in case study 2 the setting is professional fiction translation. The object of analysis changes as well: since anglicisms are likely to be rare in this setting, here we focus on phraseology, and more specifically on collocations, i.e.

> lexicalized, partly lexicalized and other "habitual" word combinations [including] conventionalized word combinations [that] do not show the typical linguistic hallmarks of lexicalization, i.e. non-compositionality, non-substitutability and non-modifiability. (Evert 2008:3)

The corpus setup is also partly different. As in the previous study, the core corpus is formed of translations into Italian, their source texts, and comparable originals − some data about the corpus, including the titles and authors/translators of the works sampled, are provided in Table 4. In this case however we work with text samples rather than whole texts, and we also use reference corpora of Italian and English, both hand-made (the BNC for English and the *Repubblica* corpus for Italian) and obtained by crawling the web (see Aston and Burnard 1998, Baroni et al 2004 and Baroni et al. 2009 for a detailed description of these corpora).

Table 4. FICTION: basic corpus data

	FICTION-OR-EN	**FICTION-TR-IT**	**FICTION-OR-IT**
Tokens	109,974	110,409	128,036
Types	13,481	17,476	20,203
Text samples	8	7	8
N. of writers/ translators	7 ---	--- 8	8 ---
Works sampled	Atwood, The handmaid's tale Atwood, Cat's eye Cruz Smith, *Gorky Park* Fowler, Red bride Gordimer, My son's story Greene, The tenth man Leavitt, A place I've never been Rendell, Kissing the gunner's daughter	Il racconto dell'ancella (Penati) Occhio di gatto (Papi) *Gorky Park* (Paolini) Nozze di sangue (Bini) Storia di mio figlio (Cavagnoli) *Il decimo uo*mo (Oddera) Un luogo dove non sono mai stato (Cossiga) Oltre il cancello (Brinis)	Camon, La malattia chiamata uomo Celati, I narratori delle pianure Comencini, Le pagine strappate Luther Blissett, *Q* Maraini, Donna in guerra Pontiggia, Il giocatore invisibile Tomasi di Lampedusa, *Il Gattopardo*

Finally, the devising of an appropriate research methodology in this case required more ingenuity. Since definitions of collocations tend to be fuzzy, the process of systematically comparing collocations across texts cannot be carried out manually without running the risk of inserting subjective biases in the data. On the other hand, the relatively small size of the corpora makes it difficult to extract collocations automatically, since only very few sequences are likely to be repeated frequently enough for statistical methods to reliably identify them. Previous work in CBTS has shown that this is indeed a thorny issue, since most corpora used in translation studies tend to be small (Dayrell 2007, Danielsson 2001, Kenny 2001).

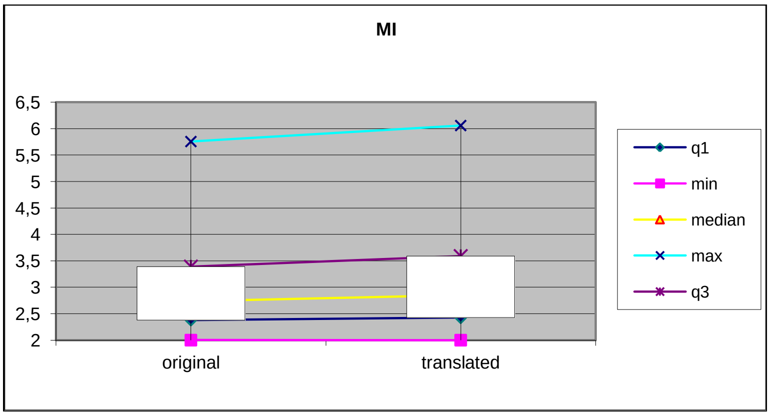
The method used in this study consists in first identifying structural patterns that were likely to form lexical collocations in Italian − e.g. Adjective-Noun (JN), Verb-Noun (VN), Noun-Preposition/Conjunction-Noun sequences (NprepconjN). This is done through research in the literature followed by random checks in the corpus. All matching sequences, regardless of their frequencies, are extracted from the Italian MCC. The frequencies of the lexical words within these structures are then collected from the *Repubblica* reference corpus. In other words, intervening function words are used for filtering out noise, but only the lexical words are kept for analysis. The collocation candidates from the (small) corpus of translated and non-translated Italian are then ranked according to their Mutual Information (MI, Church and Hanks 1990) scores in the many times larger reference corpus of general Italian. After setting a conservative threshold (frequency > 1 and MI ≥ 2), the significance of the difference between each pair of rankings is calculated using the Mann-Whitney-Wilcoxon ranks test (henceforth Mann-Whitney test). The test answers the question whether two rankings − e.g. collocation candidates belonging to the VN structural pattern extracted from the translated subcorpus vs. the non-translated subcorpus − are likely (not) to belong to the same population. Out of the many comparisons thus made, three return significant results, as shown in Table 5.

Table 5. Results of FICTION-IT significance testing (MI)

Pattern	**W**	**P value (confidence)**	**Higher in**
NV	16974.5	0.008979	Tr
VN	9936.5	0.01088	Tr
NprepconjN	272195	0.007834	Tr

In all cases, significant differences were due to translated texts displaying overall higher values. This seems to suggest that the translated texts in the corpus contain more collocations than their comparable non-translated texts. However it is well known that comparability is a problematic notion, both in corpus design in general (Kilgarriff 2001) and in CBTS in particular (Laviosa 1997, Bernardini and Zanettin 2004). Therefore, before we draw any conclusions based on these results, it is necessary to turn to a parallel comparison of the translated texts and their English source texts, to highlight any shifts that might account for them. Manual browsing of the parallel concordances is thus in order. This is a lengthy process, not free from risks of subjectivity. Yet it is necessary, if one is to avoid hasty judgments based on thorough and objective, but also rather superficial, evidence. For space reasons, here I will focus only on one construction, namely NprepconjN. A more extensive report on these results will be published in the near future (Bernardini in preparation). The boxplot graph in Figure 4 provides a visual representation of results for this pattern.

Figure 4. Boxplot graph for NprepconjN



The analysis of parallel concordance lines (1061 altogether) highlighted 67 shifts where translators made marked choices from among the options available from the target language (Italian) to render the corresponding ST expressions. These were then classified bottom up in terms of the effect they had on the target text and, indirectly, of the hypothesized motivation behind them. Apart from 7 shifts that escaped categorization, the rest were classified as making the text 'More explicit', 'More formal/precise', or 'More lexicalized/less creative'. I will only present few representative examples for each category − the number of shifts in brackets next to each category gives an idea of their relative weight.

- More explicit (32)

[1] The sitting-room door is closed. **Sun comes** through the fanlight, falling in colours across the floor: red and blue, purple. (*The handmaid's tale*)

La porta del salotto è chiusa, i **raggi del sole** filtrano dalla lunetta sulla porta, frangendosi sul pavimento: rosso e blu, violaceo. (lit. <u>rays of the sun</u>)

[2] Giggles and laughter moved the children now, like one of the **gusts** that kicked dust spiralling away in the trampled yard. (*My son's story*)

Ora i ragazzi erano scossi da fremiti di risa come le **folate di vento** che sollevavano vortici di polvere nel cortile percorso da mille passi. (lit. <u>gusts of wind</u>)

[3] Helen – She raised her palm to him. (*Red bride*)

Helen... Alzò verso di lui il palmo della mano. (lit. palm of the hand)

In examples [1]-[3] the translator opted for a more explicit solution, even though translation equivalents exist in Italian, that have exactly the same denotation as the words in bold in the examples, and that would intuitively be equally appropriate in these contexts: <u>sole</u>, <u>folate</u>, <u>palmo</u>. Of course, by using a recognizable set phrase formed of words that normally go together, the translator is able to make the text smoother, more "proper", as well as more explicit.

At times the tendency to explicitate through use of a set phrase (the data provide no clues as to the exact motivation behind these shifts) is even more marked, as in examples [4] and [5].

[4] Ram Gopal had begun passing **packs of notes** under the glass barrier. (*Kissing the gunner's daughter*)

Ram Gopal aveva cominciato a passare **mazzette di banconote** sotto la barriera di vetro. (lit. <u>wads of money of banknotes</u>)

[5] **Salesladies** intimidate me, I don't like to be caught shopping. (*Cat's eye*)

Le **commesse dei negozi** mi intimidiscono, non mi piace essere guardata quando faccio compere. (lit. <u>salesladies of shops</u>)

The Italian renderings in this case are utterly redundant, since the head nouns by themselves (<u>commesse</u> and <u>mazzette</u>) translate the English *salesladies* and *packs of notes*. The impression here is that translators have reproduced the structural layout of the ST, displaying a level of interference. The resulting phrases however are collocations in Italian, since the two lexical words that form the phrases are strongly tied to each other. The somewhat paradoxical result is a case of normalization caused by interference (see Toury 1995 and Pym 2008 for theoretical discussions about this possibility).

- More formal/precise (12)

The tendency toward greater precision and propriety seems to be the leading motivation for the shifts occurring in examples [6] and [7].

[6] I walk around to the back **door**, open it, go in, set my basket down on the kitchen table. (*The handmaid's tale*)

Proseguo fino alla **porta d'ingresso** sul retro, l'apro, entro, depongo il mio canestro sul tavolo di cucina. (lit. <u>entrance door</u>)

[7] Kicking around among the **clothes** on the floor, he found no trace of the incriminating article. (*Red bride*)

Spostando col piede i **capi di vestiario** sul pavimento, non trovò traccia della prova incriminante. (lit. items of clothing)

In these cases more informal alternatives exist in Italian, that would be (more) acceptable in context, i.e. vestiti for *clothes* and porta for *door*. Notice in the latter case how a back door, una porta sul retro in Italian, is by definition an entrance door, therefore there was no need to be more precise.

- More lexicalized/less creative (16)

Shifts in this category seem solely motivated by the desire to make the TT sound more idiomatic. At times a creative expression is rendered by means of a lexicalized one, as in examples [8] and [9], leading to a more sanitized TT (Kenny 2001). At other times the very sense expressed in the ST changes, possibly through a domesticating effort, as in examples [10] and [11].

[8] The air was reverberant, gray from the light that seeped from the faraway platform. (*Gorky Park*)

Nel tunnel soffiavano correnti d'aria e la scarsa luce, grigia, proveniva dalla lontana banchina. (lit. in the tunnel blew gusts of air)

[9] I can remember the smell of the turned earth, the plump shapes of bulbs held in the hands, **fullness**, the dry rustle of seeds through the fingers. (*The handmaid's tale*)

Ricordo l'odore della terra smossa, il **senso di pienezza** che davano le forme tonde dei bulbi chiusi nella mano, il fruscio secco dei semi tra le dita. (lit. sense of fullness)

[10] Celia has always loved it, and now, as she spoons penne into a bowl handpainted by Alex with purple **garlic bulbs**, she sees that Sylvie is an expert. (*A place I've never been)*

A Celia è sempre piaciuta moltissimo, e adesso, mentre rovescia le penne in un piatto che Alex ha dipinto a mano con decorazioni di **spicchi d'aglio**, si rende conto che Sylvie è un' esperta. (lit. cloves of garlic)

[11] On a shelf of the urinal there was a **screw of paper** - three hundred francs. (*The tenth man*)

Su una sporgenza dell'orinatoio trovò un **rotolo di banconote** - trecento franchi. (lit. roll of banknotes)

The parallel concordance analysis thus suggests that the differences observed at the MCC level are likely to follow from the translation process instead of being due to other unrelated variables (e.g. less-than-perfect comparability of the translated vs. non-translated texts). We observed a tendency for translators to opt for established collocations in the target language regardless of the presence of a corresponding collocation in the source text, and that this tendency for growing standardization (Toury 1995) goes hand in hand with a tendency to explicitate and/or make texts more formal and more proper. While one has to be careful when drawing conclusions about the translation process based on corpus work, these decisions on the part of translators might explain the MCC results observed.

4 Discussion and conclusions

The two case studies have shown that a quantitative analysis carried out on a MCC has a number of advantages for corpus-based research in translation studies: apart from the well-known focus on the target language/text, it overcomes the tendency to look at isolated, hand-picked cases, favouring a more systematic, thorough approach (all anglicisms/collocations corresponding to explicitly stated parameters); it can be carried out with limited manual intervention, and reduced risk of inserting a personal bias; and it is suitable for statistical testing of significance, since two varieties of the same language are being compared. Yet I believe that they have also shown the limits of the methodology.

MCC analyses highlight differences, and as such are ideal tools for making strong hypothesis about norms operating in a given target culture. But they have little explanatory power and rely too much on the problematic notion of textual comparability. The parallel corpus approach is still indispensable to confirm that observed differences are indeed due to the translation process (vs. unrelated variables), to add explanatory power to "black box" observations when trying to infer decision-making processes (translator strategies and procedures) from corpus data and also, crucially, to link theory/description to teaching/practice. On the basis of data such as those presented here it is possible to develop teaching materials for classroom activities in which students interpret professional translator's choices, criticize them if necessary, suggest alternatives, relate them to prevailing norms at a given point in time, and so forth. MCC results are not as easily imported into the translation classroom, because the insights they provide are at a remove from the actual translation process.

The lesson to be learnt is simple. Relying on superficial quantitative data in the search for translation norms/universals can be misleading. Insights and hypotheses should emerge from the accumulation of results of analyses conducted on closely **comparable** corpora, checked through painstaking manual scrutiny against their **parallel** text component(s) and/or against **reference** corpora of the source and target language, and any other source of contextual information available. Operatively, I would suggest that a tripartite corpus structure (source texts in language A, target texts in language B and comparable/reference originals in language B) is the minimal setup to start to shed light on the distinctive features of translated language.

References

Aston, Guy /Burnard, Lou (1998) *The BNC handbook: Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

Baker, Mona (1993) Corpus linguistics and Translation Studies. Implications and Applications. In: Baker, Mona et al. (eds) *Text and Technology*. Amsterdam: John Benjamins. 233-250.

Baroni, Marco /Bernardini, Silvia /Comastri, Federica /Piccioni, Lorenzo /Volpi, Alessandra /Aston, Guy /Mazzoleni, Marco (2004) Introducing the *La Repubblica* Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *Proceedings of LREC 2004*. Lisbon: ELDA. 1771-1774.

Baroni, Marco /Bernardini, Silvia /Ferraresi, Adriano /Zanchetta, Eros (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. *Journal of Language Resources and Evaluation* 43 (3). 209-226.

Bernardini, Silvia (in preparation). *Collocations in translation: A corpus-based study* (provisional title). Amsterdam: John Benjamins.

Bernardini, Silvia /Ferraresi, Adriano (forthcoming) Practice, Description and Theory Come Together − Normalization or Interference in Italian Technical Translation? *Meta. Translator's Journal.*

Bernardini, Silvia /Zanettin, Federico (2004): When is a Universal not a Universal? Some Limits of Current Corpus-based Methodologies for the Investigation of Translation Universals. In: Mauranen, Anna /Kujamäki, Pekka (eds) *Translation Universals. Do They Exist?* Amsterdam: John Benjamins. 51-62.

Christ, Oli (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System. *COMPLEX'94*, Budapest, 1994. URL http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench <09.08.2011>

Church, Kenneth Ward /Hanks, Patrick (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16 (1). 22-29.

Danielsson, Pernilla (2001) *The Automatic Identification of Meaningful Units in Language*. Unpublished doctoral dissertation. Göteborg: Göteborg University.

Dayrell, Carmen (2007) A Quantitative Approach to Compare Collocational Patterns in Translated and Non-translated Texts. *International Journal of Corpus Linguistics* 12 (3). 375-414.

Evert, Stefan (2008) A Lexicographic Evaluation of German Adjective-Noun Collocations. In: *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco.

Furiassi, Cristiano /Hofland, Knut (2007) The Retrieval of False Anglicisms in Newspaper Texts. In: Facchinetti, Roberta (ed.) *Corpus Linguistics 25 Years On*. Amsterdam: Rodopi. 347-363.

Kenny, Dorothy (2001) *Lexis and Creativity in Translation*. Manchester: St. Jerome.

Kilgarriff, Adam (2001) Comparing Corpora. *International Journal of Corpus Linguistics* 6 (1). 97-113.

Laviosa, Sara (1997) How Comparable can 'Comparable Corpora' Be? *Target. International Journal of Translation Studies* 9 (2). 289-319.

Laviosa, Sara (1998) Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta. Translator's Journal* 43 (4). 557-570.

Olohan, Maeve (2003) How Frequent are the Contractions? A Study of Contracted Forms in the Translational English Corpus. *Target*. *International Journal of Translation Studies* 15 (1). 59-89.

Piqué-Angordans, Jordi /Posteguillo, Santiago /Melcion, Lourdes (2006) The Development of a Computer Science Dictionary, or How to Help Translate the Untranslatable. In*:* Arnó Macià, Elisabet et al. (eds) *Information Technology in Languages for Specific Purposes. Issues and Prospects.* New York: Springer. 213-229.

Pym, Anthony (2008) On Toury's Laws of how Translators Translate. In: Pym, Anthony et al. (eds) *Beyond Descriptive Translation Studies. Investigations in Homage to Gideon Toury.* Amsterdam: John Benjamins. 311-328.

Rayson, Paul and Garside, Roger (2000) Comparing Corpora using Frequency Profiling. *Proceedings of the Workshop on Comparing Corpora at ACL 2000*. 1-6.

Schmid, Helmut (1994) "Probabilistic part-of-speech tagging using decision trees". *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, 14-16 September 1994.

Tirkkonen-Condit, Sonja (2004) Unique Items − Over- or Under-represented in Translated Language? In: Mauranen, Anna /Kujamäki, Pekka (eds) *Translation Universals: Do they Exist?* Amsterdam: John Benjamins. 177-184.

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

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