Collocations in Translated Language: Combining Parallel, Comparable and Reference Corpora

Silvia Bernardini[1]

1. Introduction

This paper describes an attempt at investigating some collocational properties of translated language. The notion of collocation is one of the cornerstones of corpus linguistics, and has been the subject of substantial speculation and empirical research (section 2.1). Within translation studies, few works have tackled this issue, partly because of methodological conundrums (section 2.2). Yet collocations – and idiomaticity in general – would seem to be relevant to the "elucidation of the nature of translated text as a mediated communicative event" (Baker, 1993: 243), and thus central to corpus-based translation studies.

The paper claims that the research questions regarding collocations in translated language posed so far need to be reframed in order to avoid the methodological problems faced by previous studies. The method devised to answer these questions is described in some detail (section 3), and a case study is presented (section 4) of a single phraseological pattern in translated and original Italian fiction texts (*Noun preposition|conjunction Noun*). Section 5 concludes the paper and makes suggestions for further research.

2. Background

2.1 Collocations

The search for collocations is one of the driving forces behind corpus linguistics. The notion is traditionally associated with the work of J.R. Firth, who promoted "the study of key-words, pivotal words, leading words, by presenting them in the company they usually keep" (Firth 1956:106-107). Scholars inspired by his work have attempted to make the notion less obscure and more operational, and to pursue its study through the use of text corpora. Jones and Sinclair (1974) describe significant collocation as the "regular collocation between items, such that they occur more often than their respective frequencies and the length of the text in which they occur would predict".

In recent years, several definitions of collocation have been proposed, usually falling within one of two general approaches (Nesselhauf 2005). *Phraseological* approaches attempt to tell collocations apart from free combinations on the one hand, and from other lexical restriction phenomena on the other. A typical phraseological definition is Howarth's (1996: 37), who describes collocations as "fully institutionalised phrases, memorized as wholes and used as conventional form-meaning pairings". Clearly, collocations here are viewed as abstract entities with instantiations in texts: the main focus is on the language user's competence. *Frequency* approaches focus less on classifying collocations, and more on identifying them in texts. Compare Kjellmer's (1987: 133) definition: "A sequence of words that occurs more than once in identical form and is grammatically well-structured

[1] School for Translators and Interpreters, University of Bologna, Italy
e-mail: silvia.bernardini@unibo.it

(Kjellmer, 1987: 133). This definition (like Jones and Sinclair's above) steers clear of criteria for collocativeness such as commutability of elements, semantic opacity and figurativeness, which are often called upon by phraseology scholars to delimit the notion from a theoretical point of view, and focuses instead on the parameters needed to automatise the extraction of collocations from corpora.

The present work falls within the frequency approach. It makes no attempt at distinguishing e.g. semantically-motivated combinations from (arbitrary) collocations, or restricted collocations from idioms. This follows from the view that any "lexicalised expression" - i.e. resulting from the operation of the idiom principle (Sinclair 1991) - is potentially relevant to our analysis (see sections 3-4 below). We shall therefore adopt Manning and Schütze's (1999: 151) rather general definition of collocation as "an expression consisting of two or more words that corresponds to some conventional way of saying things", and focus on 2-word collocations only.

2.2 Corpus-based Translation Studies

Following Baker's seminal paper (1993), a large body of research in translation has adopted a corpus-based methodology to try and shed light on the "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1993: 243). This way of envisaging translation, not as an individual act of transfer of a source text into a target language, but rather as a socio-culturally regulated communicative event in the target language community, has its roots in the work of translation studies theorists, particularly Toury (1995). Toury's notion of "translation norms", i.e., of socio-cultural constraints regulating the behaviour of professional translators and leaving traces in translated texts, has been the object of substantial research, along with the more controversial notion of "translation universals". Several features (whether universal or not) have been isolated, that would seem to characterise translated language with respect to other kinds of text production, and to point at norms of translational behaviour. These are (the list is not exhaustive) explicitation/explicitness, simplification, disambiguation, homogeneity, conventionality, normalisation/sanitisation and so forth (see Laviosa (2002) and Olohan (2004) for more exhaustive discussions).

These studies have been conducted using two kinds of resources: the more traditional parallel corpora, made of originals in language A and their translations into language B, and the innovative monolingual comparable corpora, made of originals in language A and comparable translations into language A. Sometimes these resources are combined, i.e., to form bidirectional corpora, made of originals in languages A and B, and the respective translations in languages B and A. Reference corpora of the languages under analysis are also sometimes employed as benchmarks. By limiting their scope to the target language, studies of monolingual comparable corpora are unaffected by language system-specific differences, an important variable in the parallel approach. Therefore, they have been used extensively to compare overall textual features such as sentence length, lexical variety, ratio of content words to function words (Laviosa, 2002), or more specific patterns of use of (semi-)grammatical (Olohan, 2001; Olohan and Baker, 2000) and lexical features (Tirkkonen-Condit, 2004; Mauranen, 2000).

Parallel corpus approaches are more appropriate for the analysis of local shifts and strategies. Studies following this approach have focused, e.g., on explicitating shifts (Øverås, 1998), on normalising/sanitising shifts (i.e., the tendency to select habitual target language expressions to render creative turns of phrases in the source text; Kenny, 2001) and on translator choices with implications for a description of translator's style (Malmkjær, 2004; Marco, 2004).

Turning specifically to collocation, few studies have tackled the issue. Identifying collocations in (monolingual) corpora is far from straightforward, and researchers adopting a parallel paradigm may have felt that adding a bilingual dimension would render the task a daunting one. Advocates of the monolingual comparable corpus approach, for their part, have typically tended to focus on aspects that could be identifiable via automatic procedures; this requires some ingenuity in the case of collocations, as we shall see. Yet the issue is central to an understanding of strategies and norms for dealing with lexicalisation and creativity in translation.

Kenny (2001) and Øverås (1998) provide some evidence of *normalising* shifts affecting collocations, i.e., in Toury's (1995) terms, of a tendency for translators to produce *repertoremes* (lexicalised target language collocations) in place of *textemes* (creative source text coinages). Neither method can be applied to a systematic analysis of collocations in translation, though: in the case of Øverås, because the finding is incidental, and in the case of Kenny because the starting point is the (manual) identification of the creative combinations formed around a single node word in the source text.

Attempts at analysing collocations in translated language systematically have been made by Baroni and Bernardini (2003) and Danielsson (2001). The former approaches the issue of collocations in translation from the target perspective, using a monolingual comparable corpus of Italian original and translated articles from a single geopolitics journal. All bigrams from the translated sub-corpus and from the original sub-corpus were ranked according to their log-likelihood ratio value. The bigrams most representative of the translated subcorpus (i.e., infrequent in the original subcorpus) and those most representative of the original subcorpus (i.e., infrequent in the translated subcorpus) were extracted for manual comparison. The authors report that translations in the corpus show a tendency to repeat structural patterns and strongly topic-dependent sequences, whereas originals show a higher incidence of topic-independent sequences, i.e., the more usual lexicalised collocations in the language. This work has the merit of proposing an original method for identifying collocations in translated language, that relies on a mono-source monolingual comparable corpus and goes beyond local observations of single cases selecting candidates on statistical grounds. However, the results are rather difficult to interpret. This is a common problem with quantitative studies of monolingual comparable corpora, since the general tendencies observed are difficult to pin down and interpretation is often not straightforward.

Danielsson's (2001) is an attempt at identifying “units of meaning” (~ collocations) in two monolingual corpora (one English, one Swedish), with the ultimate aim of finding “units of translation” (i.e., bilingual collocation pairings) in parallel corpora. Based on a frequency list of all the words in the corpus, word-forms occurring 200 times of more are extracted for further analysis. Upward and downward collocates (cut-off point: ≥5) are searched for and the evidence is combined to produce citation forms. As she moves on to search for units of translation, Danielsson’s work is plagued with data-sparseness problems. In the source text component of her parallel corpus of fiction texts translated from Swedish into English (~400,000 words per component), she finds that only 2 units of meaning (of the 12,099 previously identified) occur five times or more. Similar results are obtained from the English target text corpus. Danielsson is well aware of the limits of her method when it gets to the translational perspective, and acknowledges the need for much larger corpora. Unfortunately, parallel corpora are costly to assemble and tend to be small (unless one contents oneself with some widely-available text types, such as EU parliament proceedings). Therefore a method such as Danielsson's, which starts with units of meaning in reference corpora and then proceeds to look for units of translation in parallel corpora (rather than the other way round), despite its obvious value from a monolingual perspective, is bound to result in a substantial amount of precious evidence being wasted in the parallel phase.

To obviate this problem, a change of perspective is needed. Rather than identifying collocations based on the frequency and/or relatedness of word combination *tokens* in a monolingual comparable or parallel corpus, we extract word combination *types* from the corpus under study (however small), and obtain their frequency and relatedness in a (large) reference corpus. In other words, we are using reference corpora to approximate "the collective linguistic experience of a language community" (Howarth, 1996: 72), and thus bypass the data sparseness bottleneck inherent in the corpora currently available to translation scholars. Section 3 below describes the method in more detail.

3. Studying collocations in translated language

3.1 Research questions

This study addresses the following research questions:

1. Are translated texts more/less *collocational* than original texts in the same language?
    - i.e., the collocation *types* they contain are more/less frequently attested and/or significant than the collocation types found in originals?

2. If any difference can be identified, is it likely to be a consequence of the translation process?
    - i.e., can we isolate shifts (less-to-more collocational or more-to-less collocational) that can point us towards possible reasons for the observed differences?

Question 1 requires a monolingual comparable corpus and a reference corpus of the source and target languages, while question 2 requires a parallel corpus.

3.2 Corpus resources

Two tiny parallel corpora are used for this study, one containing extracts from novels and short stories in original and translated English (source language: Italian), the other containing similar extracts in original/translated Italian (source language: English). Details of these corpora, referred to below as the *LIT* corpora, are provided in tables 1-2.[2]

[2] A corpus containing open source software manuals was also analysed, to check whether the same tendencies would be observable in literary as well as technical translation. For reasons of space, results regarding this second corpus are not discussed here.

Author	Title (IT)	Sample size	Translator	Title (EN)	Sample size	
F. Camon	*La malattia chiamata uomo*	16,230	J. Shepley	*Sickness called man*	18,074	
G. Celati	*I narratori delle pianure*	19,144	R. Lumley	*Voices from the plains*	20,903	
C. Comencini	*Le pagine strappate*	23,219	G. Dowling	*The missing pages*	27,199	
Luther Blissett	*Q*	16,295	S. Whiteside	*Q*	18,247	
D. Maraini	*Donna in guerra*	17,669	D. Kitto, E. Spottiswood	*Woman at war*	19,531	
G. Pontiggia	*Il giocatore invisibile*	12,408	A. Cancogni	*The invisible player*	14,962	
G. Tomasi di Lampedusa	*Il Gattopardo*	22,275	A. Colquhoun	*The Leopard*	23,816	
Total		**127,240**			**142,732**	**269,972**

Table 1: Composition of the *LIT* corpora: the Italian=> English sub corpus

Author	Title (EN)	Sample size	Translator	Title (IT)	Sample size	
M. Atwood	*The handmaid's tail*	15,647	C. Penati	*Il racconto dell'ancella*	15,184	
M. Atwood	*Cat's eye*	15,146	M. Papi	*Occhio di gatto*	15,134	
M. Cruz Smith	*Gorky Park*	10,863	P. F. Paolini	*Gorky Park*	10,181	
C. Fowler	*Red bride*	12,350	S. Bini	*Nozze di sangue*	12,566	
N. Gordimer	*My son's story*	13,999	F. Cavagnoli	*Storia di mio figlio*	14,897	
G. Greene	*The tenth man*	11,916	B. Oddera	*Il decimo uomo*	12,284	
D. Leavitt	*A place I've never been*	15,010	A. Cossiga	*Un luogo dove non sono mai stato*	15,476	
R. Rendell	*Kissing the gunner's daughter*	14,329	H. Brinis	*Oltre il cancello*	14,284	
Total		**109,260**			**110,01**	**219,266**

Table 2: Composition of the *LIT* corpora: the English=> Italian sub corpus

We might describe this resource as a very small and opportunistically built *bidirectional corpus* (Johansson, 2000), i.e., a combination of parallel and monolingual comparable corpus resources. Yet the texts included in each parallel sub-component differ considerably from each other, such that doubts about their comparability are not unwarranted. Italian novels in translation tend to be more highbrow and to have been published by niche publishers, while English originals, with some exceptions, typically belong to more low-brow, mass-market fiction. These characteristics reflect real-world tendencies in the translation market, and cannot be swept under the carpet. They should be kept in mind when attempting to relate the results of the comparable corpus analysis and of the parallel investigation of translation shifts to the wider socio-cultural norms regulating translation – an aspect relevant to theoretical (more than descriptive) translations studies, and beyond the immediate concerns of this paper.

The reference corpora used in this study are:[3]

1. The *British National Corpus* (*BNC*) for English (100 million words from various sources)
2. The *Repubblica* Corpus for Italian (340 million words from a single newspaper)

These corpora are a) not comparable with the study corpora, i.e. they are not made of fiction texts and b) not comparable with each other (one being a “balanced” corpus, the other a single-source corpus). These should not constitute major problems for the present purposes. With regards to a), the point is to use the reference corpus as a repository of collocations that language users would recognise as well-established, filtering out sequences produced by the operation of the open-choice principle (Sinclair 1991); a fiction corpus, being potentially rich in creative combinations, might actually be detrimental. The non comparability of the two reference corpora (b) could constitute a potential problem at the stage where we try to draw conclusions about the universality of the claims, i.e. whether the shifts we observe apply to English and Italian to the same extent, and therefore could be candidates for “universal” status. Yet at the stage where we compare original and translated texts in the same language, no bias is inserted due to the choice of the reference corpus.

3.3 Corpus preparation

The reference corpora were already available (tagged, lemmatised and indexed with the *Corpus Work Bench* (CWB, Christ 1994)). The *LIT* corpora were:

1. scanned in from the paper sources
2. tokenised
3. tagged
4. lemmatised
5. indexed
6. sentence aligned

Steps 2-4 were carried out by the *tree-tagger*, a freely-available language-independent tagger pre-trained on English and Italian, as well as a few other languages (Schmid, 1994). Steps 5-6 are taken care of by CWB. At this point the different sub-corpora could be searched using the C*orpus Query Processor* (CQP), the interrogation companion to CWB.

3.4 Extraction of candidate collocations

For the present purposes, and in order to make the data set manageable, the object of study was arbitrarily restricted to collocations made of two lexical words that are either contiguous or separated by at most two function words. POS patterns matching these criteria that are likely to yield lexical collocations were then obtained from the

[3] Data were also collected from the Web through automatic queries to *Google*. These will be used in follow up studies that attempt to evaluate the effects of a massive scaling up of the reference corpus size on the results obtained.

available literature (Benson *et al.*, 1997; Dzierżanowska and Kozłowska 1999; *Oxford collocations dictionary for students of English* 2002; Jezek 2005; Voghera 2004). Examples of the patterns selected for the study are listed in table 3.

	Italian	English
contiguous	Adj-Noun Noun-Adj Verb-Noun	Adj-Noun Noun-Noun Verb-Noun
1 intervening function word	Noun-prep|conj-Noun Verb-prep-Verb	Noun-prep|conj-Noun Adj-conj-Adj
2 intervening function words	Noun-prep-pron-Verb	Verb-pron-pron-Noun

Table 3: Example patterns retrieved from the *LIT* corpora

3.5 Evaluation

All the sequences matching a given pattern are retrieved from the LIT corpora, and frequency information about the combination and about its constituents is obtained from the relevant reference corpus (see table 4).

Original (BNC)					Translated (BNC)				
W1	W2	Fq1	Fq2	Fq1-2	W1	W2	Fq1	Fq2	Fq1-2
absence	game	5625	13978	1	act	deception	11021	642	1
absence	pain	5625	6690	2	act	foundation	11021	2118	1
absence	people	5625	113684	6	activities	rules	11091	9624	1
aches	pains	162	1007	69	admission	guilt	1998	1547	17
act	adultery	11021	248	5	admission	order	1998	31665	1

Table 4: Sequence types matching the *N prep|conj N* pattern and their *BNC* frequency data

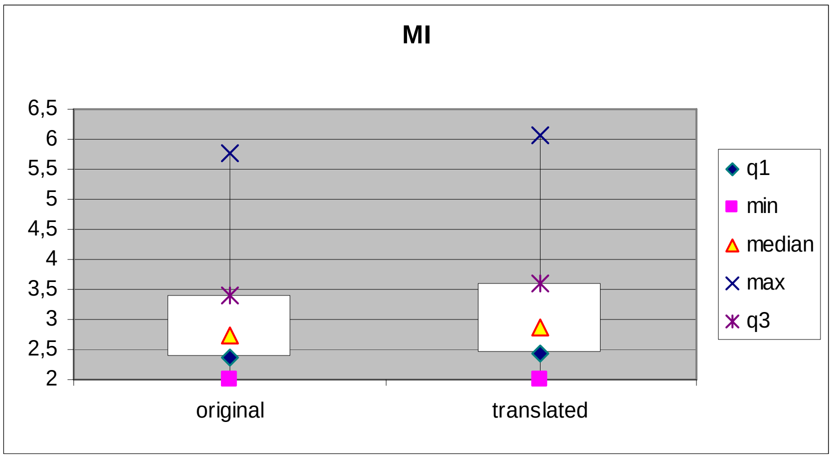
The next step consists in calculating Mutual Information (MI; Church and Hanks, 1990) values for each sequence using the UCS toolkit (Evert, 2004-2006).

Original (BNC)						Translated (BNC)					
MI	W1	W2	Fq1	Fq2	Fq 1-2	MI	W1	W2	Fq1	Fq2	Fq1-2
6.1765	Uncles	aunts	3	222	1	7.0731	Snakes	Ladders	52	13	8
5.9863	narcissi	hyacinths	24	43	1	5.9385	constancy	inconstancy	72	16	1
5.8184	aunts	uncles	222	219	32	5.7706	knives	forks	609	181	65
5.7604	uncles	aunts	219	222	28	5.7590	scribes	Pharisees	95	110	6
5.7180	frogs	toads	393	151	31	5.5936	forks	spoons	181	169	12

Table 5: Sequence types matching the *N prep|conj N* pattern, ranked according to their MI values

Once the results are ordered as in table 5, all sequences with an MI>2 and a fq>1 are selected for further analysis. These are arbitrary cut-off points that take into account the relatively small size of the reference corpora. Further work is required to find empirically the most appropriate cut-off point relative to size/specialisation of the reference corpus, as well as to evaluate the effects of other statistical measures of collocativeness. As will be seen, however, the combinations selected for further analysis would in general appear to match intuitions about institutionalisation/lexicalisation. For our purposes, as long as the same cut-off point is used for both translated and original language, no bias is inserted.

The final step of the monolingual analysis consists in comparing the distribution of MI and frequency values in each matching original/translated ranking. The Mann-Whitney significance test is used for this purpose. If the test returns a significant (p ≤ .05) or near-significant difference, descriptive statistics are obtained for the relevant ranking pair to determine which of the two rankings (original/translated) displays higher/lower values for either MI or (LOG)FQ. Table 6 summarises these data for the *Noun prep|conj Noun* pattern in Italian. The difference between the two MI distributions is statistically significant (p-value = 0.007834), while that between the LOG(fq) distributions is not (p-value = 0.8724).



MI		
statistic	original	translated
min	2.001	2.000
q1	2.381	2.425
median	2.736	2.853
q3	3.392	3.590
max	5.757	6.059
mean	2.954	3.069

Table 6: Descriptive statistics for the *N prep|conj N* pattern in Italian (MI)

This result may be partly due to the different number of candidates in each set (691 original vs. 855 translated). Notice though that this difference in size between the two sets is not an artifice of the research setup, i.e., it is not a consequence of a difference in size between the two sub-corpora. Prior to the comparisons, the number of token bigrams for each pattern in the larger file was artificially reduced to the number of occurrences in the smaller matching file (through random sampling), thus ensuring that any subsequent difference was not due to this initial bias. Furthermore, significant or near-significant differences were reported for rankings with the same or very similar numbers of observations (e.g. *N prep|conj N* in English).

Based on this monolingual comparison, we can hypothesise that Italian translators tend to make use of *N prep|conj N* established sequences (potential collocations) more than Italian authors do. To confirm this hypothesis, we need first of all to ascertain that the difference observed is indeed relatable to the process of translation, and not to other variables, e.g. to the differences in corpus set-up described in 3.2 above. To investigate this issue, we need to look at parallel concordance data.

4. Case study: *N prep|conj N* sequences

1,061 parallel concordance lines were browsed and clear cases of shifts that might affect the collocativeness of the pattern observed in the TT were extracted. Even though this procedure has the methodological limit of not accounting for all the data, it does have the advantage of providing a conservative estimate: in case of doubt, the concordance line was assigned to the "irrelevant" set. Altogether, 127 shifts (11.9 percent of concordance lines) were found and grouped into sets according to the strategy that might have motivated the shift. Tentative labels were then provided for these sets, following a bottom-up approach. Clearly, the data available only allow us to make informed guesses about translator motivation; other analytical tools would be needed to proceed in this direction (i.e., think-aloud protocols, corpora including different drafts of the same translation and so forth).

Table 7 summarises the shift types observed and the number of times they occur in the data. For reasons of space, only a single example per shift will be discussed (4.1-4.5).

Shift type	**n.**
creative → institutionalised	7
institutionalised → institutionalised (different meaning)	7
free → institutionalised	11
more explicit	86
more formal/precise	16
total shifts observed	127
total concordance lines analysed	1,061

Table 7: Shifts leading to increased institutionalisation: A quantitative estimate

4.1 Creative → institutionalised

In a few cases the institutionalised noun phrase in Italian would appear to result from a process of normalisation or "sanitisation" (Kenny, 2001), in which an unusual or stylistically marked expression in the source text is rendered with a more common, recognisable phrase in the target text. *Cf.*:

[1] TT: l'odore della terra smossa , il <senso di pienezza> che davano le forme tonde dei bulbi chiusi nella mano
(the smell of the turned earth, the sense of fullness that gave the round shapes of the bulbs held in the hand)
ST: the smell of the turned earth, the plump shapes of bulbs held in the hands, fullness
The handmaid's tale

In example [1], the noun *fullness* in the source text, isolated form the surrounding text by two commas, follows and comments on the scene just described (*the plump shapes of bulbs held in the hands*). In the target text, the syntax is normalised and made more explicit (*lit.*: "*the sense of fullness that gave the round shapes of the bulbs held in the hand*"), and coherently the rather vague noun "pienezza" is resolved as "il senso di pienezza", using a familiar Italian phrase (14 occurrences in *Repubblica*).

4.2 Institutionalised → institutionalised (different meaning)

This set groups those cases in which the tendency is not towards greater explicitness or normalisation in translation, but rather towards preserving a level of "collocativeness" similar to that of the source text. When an expression preserving both the denotative meaning and the collocative value of the source text expression is not immediately available to the translator, s/he may decide to favour the latter at the expense of the former.

[2] TT: Fa collezione di <cartine di sigarette> con disegni di aeroplani, e ne conosce tutti i nomi.
ST: He collects <cigarette cards> with pictures of airplanes on them, and knows the names of all the planes.
Cat's eye

Cigarette cards "were issued by tobacco manufacturers both to protect the cigarettes by stiffening the pack, and also to gain customer loyalty to their particular brand of cigarettes. The cards […] are considered […] one of the first collectibles".[4] *Cigarette* cards did not exist in Italy, though cards did: they accompanied products like *Lavazza* coffee and *Liebig* stock cubes. The latter are commonly known as "figurine". By analogy with these products, a possible translation for *cigarette cards* could have been "figurine delle sigarette". This option would select a free combination as a translation for an institutionalised phrase (see table 8 below). The translator instead has chosen to employ a more institutionalised phrase in Italian, i.e. "cartine di sigarette", which however has a very different meaning, namely *rolling papers*. While this choice may create a misunderstanding if one stops to think about it, it is very likely that readers do not stop at all. One of the effects of using run-of-the-mill expressions is exactly that they do not draw attention to themselves.

Expression	**Attestedness (occurrences)**	**Meaning**
Cigarette cards	*BNC*: 16 *Google*: 491,000	Collectible cards found in cigarette packs
Cartine da/per/di/delle sigarette	*Repubblica*: 3 *Google*: 726	Rolling papers, i.e. small sheets of paper which are sold for rolling one's own cigarettes
Figurine da/per/di/delle sigarette	*Repubblica*: 0 *Google*: 2	(by analogy with other products) collectible cards found in cigarette packs

Table 8: *cards* and "cartine" compared (*Google* data obtained on 29 June 2007)

4.3 Free → institutionalised

Examples in this category are clear instances of normalisation, i.e. cases where the translator has used an institutionalised Italian collocation to render an English free

[4] http://en.wikipedia.org/wiki/Cigarette_card (accessed: 29 June 2007)

combination. A rather straightforward case is [3]:

[3] TT: decorazioni di <spicchi d'aglio>, si rende conto che
ST: handpainted by Alex with purple garlic bulbs, she sees that
A place I've never been

Here the English *garlic bulbs* was rendered as “spicchi d’aglio” (*garlic cloves*) in Italian. There is no question that the translator may have got the term wrong, as *bulb* is “bulbo” in Italian, or more appropriately in the case of garlic, “testa” (*head*). Besides, serving dishes are typically decorated with *bulbs*, not *cloves* of garlic in Italy (and indeed, the scene is set in Tuscany). A tentative explanation for this choice could be that the translator (rightly) felt that “bulbi d’aglio” and “teste d’aglio”, while perfectly acceptable Italian phrases, would not be as familiar to the reader as “spicchi d’aglio”. It is interesting at this point to consider the relative frequencies of the different phrases in English and in Italian (table 9, frequency data from the Internet are given as some of the phrases have no occurrences in the *BNC*/*Repubblica*).

Expression	**Attestedness**	**Attestedness**	**Expression**
Garlic	34,300,000 (100%)	2,580,000 (100%)	aglio
garlic bulbs + bulbs of garlic	109,600 (0.31%)	1305 (0.05%)	bulbi d’aglio + bulbi di aglio
garlic heads + heads of garlic	59,300 (0.17%)	612 (0.02%)	teste d’aglio + teste di aglio
garlic cloves + cloves of garlic	2,207,000 (6.43%)	229,100 (8.87%)	spicchi d’aglio + spicchi di aglio

Table 9: *bulbs* and “spicchi” compared (*Google*, 29 June 07)

Taking these data to somehow approximate the perception of speakers of the two languages, we might conclude that the Italian translator did not have at her/his disposal an expression at exactly the same level of familiarity/collocativeness as the one used in the ST. Faced with the choice to either employ a slightly less familiar phrase (“teste d’aglio” o “bulbi d’aglio”) with the same denotational meaning, or a much more familiar phrase with a different though related meaning, s/he may have opted for the second.

4.4 Explicitation

4.4.1 Straightforward explicitation

[4] TT: l’avrei schiacciato sotto il <tacco della scarpa>, seppellito.
ST: ground away under my heel, buried along with it.
My son’s story

In example [4], an *of*-construction in Italian (“N di N”) is used to make the meaning of the English ST more explicit: while English has *heel,* Italian specifies exactly *what heel*, (i.e., *the heel of the/one’s shoe*). Notice that the Italian head noun has exactly the same denotation as its English counterpart: with no further specification, they would be taken to refer to the same objects.

4.4.2 Partitives

Very often the text is made more specific (and more familiar) through the insertion of (quantity) partitive nouns. In example [5], a partitive construction is formed with a noncount noun and a "typical" partitive that is restricted in its combinability:

[5] TT: Non riuscivo a prendere sonno , così sono sceso a bere un <sorso d'acqua> . [*gulp of water*]
ST: I couldn't sleep , so I came down for <u>water</u>.
The tenth man

4.4.3 Specificity of head nouns

A somewhat more subtle form of explicitation is also observable. In several cases, Italian and English use a similar phrase, but the head noun in Italian (being more specific or explicit than its English equivalent) could alone express the meaning of the whole English expression. Since exactly the same level of explicitness could not be obtained, the translator has opted for the more explicit solution, which is also an institutionalised phrase in the language. *Cf.* example [6]:

[6] TT: anello giallastro e parecchi <mozziconi di sigarette> che galleggiano
ST: yellowish-brown ring around the inside of the bowl and several floating <u>cigarette butts</u>
Cat's eye

In the BNC, *butt* collocates with *cigarette*, *cigar* and especially *water* and *rifle*. In *Repubblica*, "mozzicone" in its non figurative use collocates with "sigaretta" (*cigarette*), "sigaro" (*cigar*) and "candela" (*candle*). Both *cigarette butt* and "mozzicone di sigaretta" are institutionalised sequences in the two languages, and dictionary equivalents of each other. However, if we look up the corpora for occurrences of *butt* <u>not</u> accompanied by *cigarette* (or any other noun), and "mozzicone" <u>not</u> accompanied by "sigaretta" (or any other noun), we notice some differences (concordances 1-2).

```
1      new building. Stubbing out <his butt,> Stan shudders at the
2     to " get Central Office off <its butt,> and put the Governme
3     w status as the comedians ' <favourite butt.> There were so
4     oys taking snapshots of his <own butt,> and loses his dad’s
5     h close simultaneously by a <direct butting,> as distinguis
6     er, were eager to be out at <the butts.> Elizabeth herself h
7     ow by late June with a lush <green butt,> with heads formed b
8     still public land, known as <The Butts.> The library itself
9     greenheart. With a spike in <the butt.> For sticking into th
10     starting at 1.30 p.m. from <The Butts.> The race, organised by
11     werful as it bends towards <the butt.> It was designed for li
12       corner of my eye. I grab <the butt,> with my fingers curled a
13      rifle furniture -- stocks <and butts.> There was only one ma
14     ered, and the hole through <the butt,> which was counterbored
15     bolt, the head to the foot <of butt,> and held the oil bottle
16     ptain’s " Let's go back to <the butts,> gentlemen " is the sign
77     n the joystick and go kick <some butt.> Chris Nicholls from P
18     rom inexperienced dealers. <No butts...> I have a five year ol
19      factory they 're actually <half butts,> for no other reason
20     the shape and condition of <the butt,> Klondyke would expect t
21      Verlaine with a rocket up <his butt --> hardly the sound of a
22      s all the time. " You see <her butt?> What a bitch of a butt,b
23      her butt? What a bitch of <a butt,> boy! " Sooner or later on
24      Gambling is a big pain in <the butt,> " he says. " A lot of c
25     inished it and tossed away <the butt,> will add to the rubbish
```

Concordance 1: *BUTT* followed by punctuation and not preceded by a noun (*BNC*)

```
1        del sesto è rimasto " un <mozzicone ">. Le rimesse dei liba
2       si a gettare per terra un <mozzicone,> però, e un pezzo di
3       cipale responsabile, il " <mozzicone ">, che ha un alto cont
4       erà 50.000 lire chi getta <mozziconi,> lattine e cartacce s
5       tacenere conserva un solo <mozzicone,> nessun cuscino non è
6         traccia, scatole vuote, <mozzìconi,> ogni genere di rifiuti
7       ltsin -- mi ha risposto a <mozzìconi,> ha detto che Gorbaci
8       a manipolato in seguito i <mozzìconi )>, ma scientificament
9        ri, portaceneri pieni di <mozzìconi,> e libri gialli; un co
10      di cui si raccolgono solo <mozzìconi "> Basta... Anch' io sono
11       il tabaccaio ritirava il <mozzicone,> lo spegneva e lo rivende
12      volontariamente ". Questo <mozzicone,> è possibile -- continua
13      mmacolate -- non c’era un <mozzicone -->, immacolati la scuola
14       te, il piattino pieno di <mozzìconi;> e impepate di cozze, me
15       , terreno mitragliato di <mozzìconi,> si pregano gli spettator
16      on mano una candela; o un <mozzicone,> un moccolo della medesim
17      n la sabbia senza trovare <mozzìconi ">, ha detto l’assessore
18      che il giallo parte da un <mozzicone,> ma in questo episodio i
19      ri che la colpa era di un <mozzicone,> di una stufetta, la cla
20       sua casa non restano che <mozzìconi,> per la strada non c’è n
21      ndo per terra bottiglie e <mozzìconi,> come in una finale di Co
22       izzato in michette... al <mozzicone (> multa di 850 mila lire )
23      che una campagna contro i <mozzìconi,> le cicche, quei pezzett
24       rigano, sono in fondo un <mozzicone --> un plastico, un ricordo
25       a Lagorio: " Miliardi di <mozzìconi,> graffiti sui muri. Non
26      to la Fenice a meno di un <mozzicone:> imputati di strage e di
27       per un incendio basta un <mozzicone,> i bianconeri si sentono
```

Concordance 2: "MOZZICONE" followed by punctuation (*Repubblica)*

Out of 58 occurrences of the lemma *BUTT* followed by punctuation and not preceded by a noun in the BNC, only 2 lines refer to the smoking sense (Concordance 1 shows a selection of the hits found). If we compare a concordance from *La Repubblica*, we find instead that two thirds of the concordance lines refer to cigarette butts. In other words, while *butt* does not imply *cigarette* (i.e., a *butt* is not necessarily understood as a *cigarette butt*, but more likely as a body part or part of a weapon), “mozzicone” does imply “sigaretta” (i.e., “un mozzicone” is almost certainly to be understood as “un mozzicone di sigaretta”). Therefore, while *cigarette butt* and “mozzicone di sigaretta” may appear to be obvious decontextualised equivalents of each other, the decision to translate *cigarette butt* with “mozzicone di sigaretta” in fact results in a target text that is more explicit than its source text, and one that adopts an established restricted combination in Italian where there is no strict need for it.

4.5 More formal/precise

This set groups cases where the expression used in Italian a) has an alternative more informal or more vague rendering that would be equally adequate and (possibly) more acceptable in context, and/or b) displays a concern with exactness of expression that does not follow from the ST.

[7] TT: Spostando col piede i <capi di vestiario> sul pavimento , non trovò traccia della prova incriminante. [*items of clothing*; instead of “vestiti” (*clothes*)]
ST: Kicking around among the clothes on the floor, he found no trace of the incriminating article.
Red bride

[8] TT: Da un bel pezzo, un’eternità, non provava più un vero e proprio <senso di nausea> davanti a spettacoli del genere. [*a sense of nausea*]
ST: It was a long time, an age, since he had felt actual physical nausea at such sights.
Kissing the gunner’s daughter

5. Concluding remarks

This paper has described a method for comparing use of collocations in original and translated language that relies on monolingual comparable corpora and reference corpora. Combining this monolingual perspective with a parallel focus, a case study has been subsequently presented of the *Noun preposition|conjunction Noun* pattern in original and translated Italian. Going back to our research questions, and relying only on the analysis of a single collocation pattern, we can tentatively suggest that translated texts would seem to be more *collocational* than original texts in the same language, and that there is some evidence that this is a consequence of the translation process.[5]

Overall, collocations following this pattern have higher MI values in translated than in original fiction texts. Over a thousand parallel concordance lines were analysed, and almost 12 percent of these showed shifts towards greater collocativeness. In many cases, the increase in collocativeness was accompanied by other phenomena, i.e. normalisation and explicitation (the quantitatively most prominent set). It is difficult to tell what was the principal driving force behind these shifts. Yet the very relationship we observed between (some of) the regularities of translation behaviour observed in the literature (*cf.* 2.2 above) and the (higher) level of institutionalisation of the language used in translated texts is in itself an intriguing finding, that deserves further investigation.

Several directions for further study can be taken at this point. To shed light on the “universality” of the shifts described above, this method could be applied to other mediating contexts (e.g. technical translation, interpreting) as well as to other

[5] Analyses of all the significantly different ranking pairs (not reported on here) are consistent with this claim.

language pairs. To confirm the hunches about translator motivation, corpus-based (product-oriented) studies could be combined with process-oriented approaches to translation research. Lastly, from a methodological point of view it would be important to experiment with different statistical measures, frequency cut-off points, and reference corpora.

References

Baroni, M. and S. Bernardini (2003). A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora. In D. Archer, P. Rayson, A. Wilson and A. McEnery (eds.) Proceedings of Corpus Linguistics 2003, pp. 82–91. Lancaster: UCREL.

Benson, M., E. Benson and R. Ilson (1997). The BBI dictionary of English word combinations. Amsterdam: Benjamins.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. *COMPLEX '94, Budapest 1994*.

Church, K.W. and P. Hanks (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics 16,1*: 22–29.

Danielsson, P. (2001). The Automatic Identification of Meaningful Units in Language. Unpublished PhD Thesis. Göteborg University.

Dzierżanowska, H. and C. Kozłowska (1999). LTP Dictionary of Selected Collocations (edited by J. Hill and M. Lewis). Hove: LTP.

Evert, S. (2002–2004). The UCS Toolkit. Available online from http://www.collocations.de/software.html (accessed: 29 June 2007).

Firth, J.R. 1956. Descriptive linguistics and the study of English. in F.R. Palmer (ed.) (1968). Selected papers of J.R. Firth 1952-1959. London and Harlow: Longman. 96–113.

Howarth, P. (1996). Phraseology in English Academic Writing. Tübingen: Niemeyer.

Jezek, E. (2005). Lessico. Classi di Parole, Strutture, Combinazioni. Bologna: Il Mulino.

Johansson, S. (2000). "Reflections on Corpora and their Uses in Cross-linguistic Research", in F. Zanettin, S. Bernardini and D. Stewart (eds.) Corpora in Translator Education, pp. 135–44. Manchester: St Jerome.

Jones, S. and J. McH. Sinclair (1974). English Lexical Collocations. *Cahiers de Lexicologie 24*: 15–61.

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

Kjellmer, G. (1987). Aspects of English Collocations. In W. Meijs (ed.) Corpus Linguistics and Beyond. Proceedings of the seventh international conference on English language research on computerised corpora, pp. 133–40. Amsterdam: Rodopi.

Laviosa, S. (2002). Corpus-based Translation Studies: Theory, Findings, Applications. Amsterdam: Rodopi.

Malmkjær, K. (2004). Translational Stylistics: Dulcken's Translations of Hans Christian Andersen. *Language and Literature 13,1*: 13–24.

Manning, C. and H. Schütze. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Marco, J. (2004). Translating Style and Styles of Translating: Henry James and Edgar Allan Poe in Catalan. *Language and Literature 13,1*: 73–90.

Mauranen, A. (2000). Strange Strings in Translated Language: A Study on Corpora. In M. Olohan (ed.) Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects, pp. 119–41. Manchester: St.

Jerome.

Nesselhauf, N. (2005). Collocations in a Learner Corpus. Amsterdam: John Benjamins.

Olohan, M. (2001). Spelling out the Optionals in Translation: A Corpus Study. In P. Rayson, A. Wilson, A. McEnery, A. Hardie and S. Khoja (eds.) Proceedings of Corpus Linguistics 2001, pp. 423–32. Lancaster: UCREL.

Olohan, M. (2004). Introducing Corpora in Translation Studies. London: Routledge.

Olohan, M. and M. Baker (2000). Reporting that in Translated English: Evidence for Subconscious Processes of Explicitation? *Across Languages and Cultures 1,2*: 141–58.

Øverås, L. (1998). In Search of the Third Code: An Investigation of Norms in Literary Translation. *Meta 43, 4*: 571–88.

Oxford Collocations Dictionary for Students of English (2002). Oxford: Oxford University Press.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. September 1994.

Sinclair, J. McH. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Tirkkonen-Condit, S. (2004). Unique items – Over- or Under-represented in Translated Language?, in A. Mauranen and P. Kujamäki (eds.) Translation Universals. Do they Exist?, pp. 175–84. Amsterdam: John Benjamins.

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

Voghera, M. (2004). Polirematiche, in M. Grossman and F. Rainer (eds.) La formazione delle parole in italiano. Tübingen: Niemeyer.