Arriving at equivalence. Making a case for comparable corpora in Translation Studies

Gill Philip – University of Bologna, Italy

Abstract
When multilingual corpora are used in translation studies, it is usually assumed that they are either parallel or comparable, or both; and that their size and text composition are analogous. As general reference corpora become more widely available, it is inevitable that these too should be used to compare and contrast SL norms, thus extending the definition of comparability to include text collections whose size and content may vary considerably, and which are nevertheless considered representative of their languages. This paper addresses the contribution of comparable reference corpora to the identification of translation equivalence. Focusing in particular on native-speaker norms, it demonstrates how the effect of creative and idiosyncratic language can be identified and reproduced by the translator.

Keywords
General reference corpora, translation equivalence, synonymy

1. Introduction

Should corpus-based translation studies necessarily study translated text? This question may appear misplaced, especially considering that the greater part of corpus-based translation studies make use of parallel corpora, which inevitably include translated texts. But while translated texts reveal much about translation as a practice and as a product, they provide rather less information about TL norms; and if the ultimate aim of a translation is to avoid sounding like a translation, and for its provenance to pass unnoticed, it must reflect TL norms and bear these in mind when reproducing any idiosyncratic usage or innovative expressions that the SL text might include.

Parallel corpora containing text in two or more languages are useful repositories of translation equivalents which can be accessed with a minimum of time and effort, this being of fundamental importance to the professional translator and trainee alike. The use of such corpora makes it possible to analyse the translation choices which have already been made in preparing a TL text; they highlight the subtleties involved in choosing one possible translation equivalent over another, and they can also serve as a 'quick-fix' translation provider. Yet in spite of the many benefits that they have brought to the field of translation, the fact that they consist of mediated texts means that they do have their limitations. By presenting translation as a *fait accompli*, and by focusing on the translation product, the norms of the TL can easily be ignored, and the relationship of the translated text to SL texts in the same language is subordinate to that holding between a SL original and its translation into the TL.[1]

It is for this reason that the comparable element of a bi-directional corpus can prove invaluable, in that it can serve as a “control for translation effects” (Johansson 1998:6); but in this context is not used as the initial basis for the study of translation. The role of the comparable corpus is principally aimed at identifying – and eliminating – the more glaring inconsistencies in terminology, textual organisation and phraseology in translated texts. Less obvious anomalies, including the rendering of conventional and formulaic language, and the normalisation of SL oddities, receive far less attention. Parallel corpora, and the comparable corpora incorporated within them, are restricted in size and scope, encapsulating single or closely-related genres, and are designed to address the specific needs of genre- or domain-specific translation. They are neither large nor wide-ranging enough to be able to give an indication of more generalised norms within the languages under study – those norms that contribute to the perception of naturalness in text, but which are not as easily identified as terminological or structural aspects. It is here that the role of general reference corpora proves its worth.

As general reference corpora become increasingly available for a wide variety of languages, the potential for contrastive bilingual studies grows. While terminology and other text- or genre-specific language remains under the dominion of parallel corpora, matters of a more wide-ranging and general nature can be usefully addressed by identifying and contrasting the language norms displayed in general reference corpora for the languages under study. This paper describes how comparable general reference corpora can be used to identify translation equivalents through the analysis and matching of cotextual patterns, and reveals how knowledge of these norms can be profitably exploited to avoid normalisation in the translation of creative and idiosyncratic language.

2. Comparability and general reference corpora

In extending the definition of comparable corpus to include text collections whose size and content may vary considerably, a number of matters must be addressed regarding the composition and size of the corpora, as well as their representativeness[2] relative to their respective languages.

In an ideal world, all general reference corpora would follow the same design criteria, making them of similar size, composed of similar text types in similar proportions. However in real life several standard models coexist, and each has its proponents and detractors. The Brown corpus (1 million words) and those modelled on it, including Frown, LOB and FLOB, comprises text samples of equal length, but as a corollary of

[1] This is clearly not the case for monolingual parallel corpora such as the Translational English Corpus (http://www2.umist.ac.uk/ctis/research/TEC/tec_home_page.htm; see also Laviosa 1998) in which the TL texts are compared directly to SL texts of the same genre.

[2] Although representativeness remains a moot point in corpus linguistics, it is beyond the scope of this paper to enter into the details of the argument: the reader is referred to Biber (1993) for a comprehensive account, and to Laviosa (1997) for considerations specific to comparable corpora.

this there are very few whole texts present in the data set, which means that organisational features may not be adequately represented. The British National Corpus (BNC) contains 100 million words of both spoken and written texts (10% and 90% respectively) produced since 1964;[3] sampling only takes place in texts which exceed 45,000 words in length, and is intended to avoid the risk of a single author's idiosyncrasies skewing the data. In common with the Brown corpus, the BNC is static, and can only be kept up-to-date through re-issue. The corpus used in this study,[4] the Bank of English, is a monitor corpus which undergoes constant updating and expansion, and now comprises 450 million words of running text.

Table 1. Proportions of text types in the Bank of English and CORIS

text type	**Bank of English**	**CORIS**
misc. journalism	65.5%	47.5%
general prose	17%	25%
academic prose	1.5%	12.5%
legal prose	--	10%
ephemera	1%	5%
spoken	15%	--

Publicly-accessible corpora for Italian are few and far between. This study draws on the Corpus di Italian Scritto (CORIS), which was the only Italian language corpus available at the time when this research was being carried out.[5] The composition of the 80 million-word CORIS is modelled on the written component of the Longman corpus of Spoken and Written English, making it qualitatively different, as well as considerably smaller than, the Bank of English. Table 1 gives an indication of the distribution of text types in the two corpora.

Given these differences, it might appear far-fetched to describe the Bank of English and CORIS as comparable, but the most important consideration to bear in mind is that they are large general reference corpora, not small, text-or genre-specific parallel corpora. Dissimilar composition is not proof of incomparability: in accepting that languages are anisomorphic, it should come as no surprise that different languages give more priority to some text types, and less to others. In fact the decision to model CORIS on LSWE was taken because its make-up was deemed more appropriate to Italian than the other contending models, most notably the BNC, LOB and Bank of English (Rossini Favretti 2000: 51). General reference corpora are expected to exemplify their languages in a balanced way, and as it is true that languages are not translations of each other, so representative samples of those languages need not mirror each other's composition. Viewed in this light, then, the comparability of the Bank of English and CORIS is not *absolute*, as two corpora constructed to the same design specification, but rather *relative*, with each corpus being independently constructed to take account of language-specific features and thus constitute a representative sample of the languages concerned.

[3] See http://www.natcorp.ox.ac.uk/corpus/creating.xml for details of the text composition of the BNC

[4] The author expresses her gratitude to the University of Birmingham/HarperCollins publishers for unlimited access to the Bank of English data for the duration of her PhD research, 1997-2003.

[5] Access to CORIS is available by request http://corpus.cilta.unibo.it:8080. It should also be mentioned that a much larger corpus for Italian, composed of texts from the Repubblica newspaper, has since been constructed. See Aston & Piccioni (2004).

3. Background to this research

Having justified the use of the term *comparable*, it is now necessary to fill in some of the background to the applicability of comparable general reference corpora to the translation examples to be examined in 4.1.1 and 5.1.

The research from which this paper is drawn, Philip (2003), is a corpus-driven study of connotation in non-literary language. It examines the meaning of colour words as found in conventional linguistic expressions such as *to see red*, *to feel blue*, and *green with envy*, and explains what factors are responsible for activating the connotative meanings of the colour words when the expressions are used in running text. By comparing colour-word expressions with a number of near-synonyms which display similar phraseological patternings (e.g. *to catch red handed*, *to catch in the act*, *to catch in flagrante delicto*), it can be observed that the selection of one expression over another is largely predetermined by the situational context which the language is describing, and is both predicted and constrained by the regularity of patterning in the co-textual environment.

Although colour words are widely considered to be highly salient, Philip demonstrates that when they occur as part of conventional, non-compositional expressions, their meaning is subjected to the process of delexicalisation in the same way as any other component of a non-compositional chunk, in accordance with Sinclair's (1991) idiom principle and Louw's (2000) theory of progressive delexicalisation. As a result, the metaphorical-connotative meaning potential of these words remains latent, only rearing its head when the expressions undergo creative variation.

When the canonical forms of conventional expressions are altered, the way in which the phrase is interpreted changes radically, because the novel element has to be integrated into the whole. In order to do this, the non-compositional phrase is broken down into its component parts, and regains a degree of compositionality. The meaning is then reprocessed to make the relationship of the novel element to the underlying canonical form contingent; in doing so, meanings which are normally delexicalised regain a degree of saliency and metaphorical life. This can be observed by comparing the canonical forms in examples 1 and 2 with the creative variants in examples 3 and 4.

1. The gang was finally *caught red-handed* in an armed police ambush in September 1992

2. At the time it was claimed Kerr had been *caught red-handed* trying to smuggle arms to the Irish Republic.

3. A car *hi-fi* thief was *caught Simply Red-handed* when he took a *CD player* into a store owned by his victim.

4. Mr. *Green* apparently had been *caught scarlet-handed* at his own *blackmail* game. Pictures of him with Miss *Scarlet* were found hidden in Scarlet's bedroom.

A similar phenomenon takes place when the cotext includes an element that favours a salient interpretation within the non-compositional phrase, as is the case with example 4 which includes a colour-word in the proper name, in addition to the colour word component of *blackmail*. The proximity of colour words in the phraseological core and in the co-text causes the delexical colour word to be relexicalised, thus re-activating the salient meaning.

In both these types of variation – phrase-internal, and phrase-external – the chunk is read as a phraseological palimpsest, the sum of the underlying conventional, delexicalised meaning and the novel, salient one that is superimposed on top of it.

4. Identifying translation equivalents

Conducting such a study with reference to two languages has translation as its ultimate aim. Monolingual reference corpora make it possible to identify the mechanisms which drive creativity in both languages concerned, and the results obtained demonstrate that the changes in meaning are governed by the same general principles – the combination of delexical meaning and a contextually-relevant salient add-on. This knowledge provides the basis for the informed translation of unconventional language, especially that found in literature, journalism and advertising, where word-play and anomalous language often falls victim to the normalisation process in translation (Kenny 2001: 65-69).

Translation involves a great deal of choice, whether explicit or implicit. Choice implies the selection of one interpretation, expressed by a particular sequence of words, over other possible contenders, with the aim to achieve as close an effect as possible to that obtained in the SL text. As Halliday puts it:

> "The translator is aware that a given item in the source has a set of possible equivalents in the target language. [S/he is] aware that these are not free variants but they are contextually conditioned. By 'contextually conditioned' I do not mean that in a given context you must choose A and cannot choose B or C, but that if you choose A or B or C then the meaning of that choice will differ according to what the context is." (Halliday 1992: 16)

So what is the translator's choice based on? Expert knowledge of the languages provides a substantial degree of intuition regarding equivalence; and language reference books and media fill in the gaps to a certain extent; but when the translator is faced with a range of apparently synonymous possibilities, how should he or she proceed? No two expressions are identical in meaning and function, but the fine details of the distinctions all too often escape our conscious knowledge.

The concept of the translation network as a reference point is very attractive to translation scholars, and corpus data can be usefully exploited in order to identify series of translations equivalents for given words or expressions. Reference to corpus data makes it possible to identify where differences and similarities lie across languages, thus fine-tuning the translator's knowledge; but while academically interesting, the procedures followed in order to identify equivalences are cumbersome, involving umpteen translation and back-translation phases, and they are predominantly based on existing translation practice as represented in the parallel corpus data rather than on native speaker norms as represented in the untranslated language of monolingual reference corpora.

The building up of translation networks from parallel corpora generally starts with a word (only rarely is a phrasal expression the primary focus) and a single translation in the TL language(s), this translation usually being the most salient of the possible translations. This single translation is extended as the different patternings of the SL term are matched up to equivalents – these equivalents being defined functionally (Tognini Bonelli 2000). Once translations have been identified for the main patterns of the SL term, the first back-translation phase takes place, as the TL terms' patternings are matched up in the SL, with the inevitable result that new translation equivalents are identified. The procedure is resumed, taking these new SL terms as the departure point, and so on. The network, or "translation web" (Tognini Bonelli 2001: 150-154), becomes more complex with every process of translation and back-translation as more and more terms are added and connected up with their equivalents.[6] It is also time-consuming as new items are added both to the SL and the TL sides at each stage in the process, making it potentially never-ending.

4.1 Translation equivalence and the paradigmatic axis

One way to place a control on the network is to work on the basis that the SL word is one of several members of a larger semantic set, and as such it is distinguished and distinguishable from a range of near-synonyms. In this way, the analysis of the patternings of the SL term and its near-synonyms is carried out before embarking on the translation process. If the same procedure is applied to the posited TL equivalent term, i.e. that it is considered as a member of an analogous paradigm, for which the various patternings have to be identified, then the location of translation equivalents becomes a matter of matching up patternings, rather than searching for new expressions every time a new pattern appears. The correspondences are more detailed and accurate, and the tangled web of translations can be replaced by a more robust and linear schema of one-to-one correspondences that are arrived at independently of which language is to be considered the source or the target.

4.1.1 Case study: go red

The approach outlined above is best illustrated with a practical example. By taking the expression *to go red* as a point of departure, it is necessary to decide which other expressions belong to the paradigm in English, to posit an equivalent term in the TL (here, Italian), and to identify its near-synonyms. This stage does not require use of a corpus, as the necessary information can be found in standard language reference works (mono-and bilingual dictionaries and thesauri). Thus, from the single item *to go red*, the English paradigm can be identified as *to go red*, *to become red*, *to blush*, *to flush*, *to redden*, and *to turn red*. The Italian equivalent selected is *diventare rosso*; the other members of its paradigm are *arrossare, arrossarsi, arrossire, arrossirsi, farsi rosso* (*in viso/faccia*), and *far salire il sangue*.[7]

[6] As an illustration of the complexity involved, the reader is referred to the schematic representation of *sorrow* and its translation into German in Váradi and Kiss (2001: 169).

[7] This represents the full paradigm of translations found in Ragazzini 1995. It should be noted that this is

The expressions are initially analysed without any reference being made to their translatability. The English terms are studied via corpus data from an English general reference corpus (in this case, the Bank of English), and the Italian terms are examined though Italian general reference corpus data (CORIS). Each term is broken down into its sense divisions (for example, separating out reflexive and non-reflexive forms of *arrossar(si*) and *arrossir(si),* and dividing the transitive and intransitive forms of *flush*) and extended units of meaning (Sinclair 1996). The expressions are profiled in terms of their collocational patterns, their colligational and semantic preferences, any extra-linguistic function or context of use that is indicated in the data, and, once the more detailed sub-senses and phraseologies have been identified, any apparent semantic prosody that these suggest is also noted.[8] Only once this detailed monolingual examination is complete is it possible to match up terms on the basis of the linguistic (and extra-linguistic) features that they have in common. This makes it possible to identify translation equivalence in a much more detailed and consistent way than any approach which takes the word alone as its starting point. By subdividing all the terms into their smaller units, it is possible to recognise, for example, that the presence of reflexivity can be a determining feature in arriving at translation equivalence (*arrossare* corresponds to *redden*, yet its corresponding reflexive form *arrossarsi* shares the same patterning as *become red*); or that the terms give rise to similar (and equivalent) phraseological or terminological constructions, as is the case with *have the grace to blush* and *degnare di arrossire*; and that the same subdivisions of meaning may be expressed in similar ways across both languages, for example *go red as a beetroot* and *diventare rosso come un peperone* which refer to embarrassment, while their related forms *go red as a lobster* and *diventare rosso come un gambero* describe sunburn.

4.2 Manual and automatic profiling

With its requirement for detailed analysis of members of a semantic set rather than of a single term, the paradigmatic model may give the impression of being perhaps unnecessarily time-consuming, but it should be remembered that it is proposed as an alternative to the existing – and considerably more onerous – method involving successive stages of translation and back-translation. If the intention is to compile some sort of translation database or to improve translators' reference works, then the corpus approach gives the most comprehensive account of how cotextual features contribute to the building up of meaning. It can provide extremely detailed information about how the words in question combine, the units of meaning that they generate, their textual positioning and their extra-linguistic function; all these aspects are potentially necessary to the translator of text.

Word profiling of the sort discussed here can be done manually, automatically or through a combination of both. The precise approach taken depends on time available, the potential of the analysis tools, and indeed the corpus itself, as some can only be interrogated through their built-in query software, which may limit the degree to which

not an exhaustive list of every possible comparable expression, and it avoids paraphrase.

[8] The analysis of this data set made it evident that semantic prosodies cannot be identified for each term as a whole, but that it is specific to the particular patterns formed around the node. For this reason it is identified last of all, as is tied to units of meaning which include the node, but not the node alone.

the analysis can be automated. The profiling discussed in this paper was carried out mainly by hand, the choice being determined by the corpora used: both the Bank of English and CORIS are only available by remote access, and can only be interrogated by their built-in query software.

Manual profiling is time-consuming, but generally highly accurate, as the human analyst is able to recognise semantic relations between collocates more easily than a computer can, and this is especially important when it is a semantic preference rather than an individual word-collocates which forms the impression of a recurring pattern. Humans also find it easier to spot long-distance collocates (Seipmann, 2005), and incomplete fragments or fractured phraseological patterns (Moon 1998). Automatic profiling software such as the *Sketch Engine*[9] running on BNC data (Kilgarriff and Tugwell 2002, Kilgarriff et al 2004) can be very sophisticated and detailed, though it is less able to spot semantic sets or phraseological patterns than the human analyst. Less sophisticated applications are quicker, but less time often results in less detailed information. The “picture” option in the Bank of English’s suite of tools (see Krishnamurthy 2000: 36-39) gives an overview of collocational frequency at the various positions around the search term; most PC concordance packages come ready-equipped with options for calculating collocations, patterns, n-grams and so on by frequency, and Fletcher’s *Phrases in English*[10] (again operating with BNC data) also allows n-grams, phraseological frames, POS-grams and chargrams patterns to be identified.
As time goes on, corpus query software offers more and more sophisticated facilities which make profiling quick and reliable, making manual analysis necessary only to verify, fine-tune and trouble-shoot. This is clearly a boon to the translator and language researcher analyst alike, as the most frequent patterns are flagged up automatically, thus saving time on the initial phases of analysis.

5. Native norms and creativity in translation

In all attempts at pattern matching there will inevitably be some forms that appear not to have an equivalent, at least insofar as the paradigms studied are concerned. This is the case with the sense of *turn red* that collocates with leaves and berries. Although *turn red* can nearly always be translated as *diventare rosso*, this form in Italian never occurs with plant collocates to give the meaning “ripening”, nor do any of its near-synonyms. In such a case as this, a new, related paradigm can be opened up for exploration (*ripen* and *maturare*, with their synonyms). On the other hand, should no translation be found to be appropriate, then, as Baker reminds us, “[a] certain amount of loss, addition, or skewing of meaning is often unavoidable” (1992: 57). The recurring phrase *arrossire fino ai radici dei capelli* (literally, “to blush to the roots of one’s hair”) is one such case in point. The translator is likely to resort to traditional remedies such as paraphrase (*to go bright red*), or an untranslated borrowing, which would however be novel and marked in English. A literal translation with gloss would not be likely with this example, but may prove the best solution elsewhere.

[9] http://www.sketchengine.co.uk/
[10] http://pie.usna.edu/

The adoption of a paradigm in translation adds a further degree of consciousness to the translation process. The translator is able to enter into an awareness of the language choices made by the author, and thus not only find the most accurate translation, but also note the differences between this term and the others which could have been used, but were not. This notion takes on particular importance when the language being translated differs from the norm – either in extreme cases such as the translation of poetry, or in the day-to-day inventiveness that characterises normal language use. Peculiarities and deviations from the SL norm can be assessed in relation to that norm and replicated in the TL, in full consciousness rather than by mere instinct. This means that the translation can match the effect of the original, because the mechanisms governing the effect can be identified and reproduced.

Using general reference corpora as an aid to the translation process means using data which makes it possible to assess and compare norms across the languages involved. Translated text does not fail utterly in this role, but it bears the sign of translation choices already made – for good or ill. With normalisation prevalent in translation of (apparently) atypical language, it is useful to be able to compare, as Kenny does (2001: 125ff.), parallel corpus data with comparable corpus data, and to do so both for the SL and the TL.

5.1 Expressing emotion through colour

Colour words are typically used in European languages to express emotional states, mainly because there is a fairly transparent metonymical connection between, for instance, adrenaline speeding up the flow of blood through the body, and the face becoming flushed or red. So it is justifiable to expect that colour-word expressions should be used to refer to the manifestation of emotion in several languages. What may come as something of a surprise, however, is that the colour words typically used are not necessarily the same. Within Europe, English is odd in that it associates the colour green with anger, when other languages prefer yellow, the colour of bile. But the non-equivalences do not end here.

Casting aside any cultural reasons why colours and emotional states should not correspond exactly (see Nieimeier 1998; Philip 2003: 151-164), the fact remains that there is a degree of language variation in this area, and that a translator should be in a position to address it appropriately. Consider the following corpus extracts (Examples 5-7), in which *rabbia* (rage) is assigned different colours – *nero* (black), *viola* (purple), and *verde* (green).

5. Chi sta vicino al Castel dè Britti, dice che è *nero di rabbia*, che sogna la rivincita.
6. È *viola di rabbia*, una furia scatenata
7. Quando mia nonna le ha risposto: “Speriamo di no, altrimenti verrà fuori una puttana come te!”, ho visto mia madre diventare *verde di rabbia*.

How normal is it to use these colours to describe anger? An English-speaker wishing to translate these examples might (erroneously) consider all three to be innovative, variations to the canonical form *rosso di rabbia* (red with rage). This preconception derives from the translator’s L1 in which *red with anger* is the canonical expression;

and while *viola* may not seem unusual because *purple with rage* is quite common (see Table 2 for frequencies), *nero* would appear odd as *black* is very uncommon, and in fact does not appear at all with either *anger* or *rage* in the Bank of English data.

Table 2. Colours of rage and anger in Italian and English.

Italian	**English**
nero di rabbia (8)	*black* (0)
rosso di rabbia (5)	red with anger (18)/ rage (28)
verde di rabbia (5)	green with anger (1)/ rage (2)
bianco di rabbia (2)	white with anger (7)/ rage (9)
blu dalla rabbia (1)	*blue* (0)
viola (0)	purple with anger (1)/ rage (21)
rosa (0)	pink with anger (3)/ rage (2)

If these preconceptions are checked against Italian norms, as presented in the general reference corpus, a different picture emerges. The fact of the matter is that *nero* (8 of the 22 occurrences) is the most commonly used colour word in this context, followed by *rosso* and *verde* (5 each), and *bianco* (white). It therefore becomes apparent that the only unusual colour of the three that appear in the examples is *viola*, with both *nero* and *verde* being at least as frequent as the expected *rosso*.[11]

What are the implications of this for translation? If the corpus data shows that *nero* is commonly used in this pattern but *black* is not, how should the translator proceed? At this stage the principles of delexicalisation in conventionalised phraseology come back to centre stage. *Nero di rabbia* is unmarked, and the term which is correspondingly unmarked in English is *red with anger/rage*. By matching these expressions, the salient meaning of the colour word has to be ignored in favour of the unmarked phraseological meaning, which in this case is equivalent. The same is true for *verde di rabbia*, again an unmarked form. Should there be text-internal reasons for considering the colour to be relevant, the translator could use the alternative, *livid*; but if there are no special circumstances to take into consideration, then again *red* would serve to translate *verde*. The anomalous *viola di rabbia* would be inaccurately rendered by an unmarked form such as *purple with rage*, so some alternative rendering would be desirable; the most likely course of action would be to move away from the basic colour terms (Berlin & Kay 1969) and select a particular shade such as *plum*, *puce* or even *regal purple*. In doing so, the colour is perceived accurately, but the effect of the SL original is preserved because the phrase is not normalised.

Creative use of language may well make up only a small proportion of the language that is translated every day, but it is important both culturally and linguistically for a translator to render it in an appropriate manner. If the innovative and marked can be compared to related, unmarked forms in the SL, then the translator's job is facilitated greatly. By considering conventional language as largely delexicalised, and innovative language as being a combination of a delexical support and a contextually relevant addition, it is possible to go about achieving the same effect in the TL by adhering to the

[11] The example illustrating *viola di rabbia* was located on the Internet; there were no occurrences in the CORIS data.

same principles, essentially re-creating the TL text in the same way as the SL text was constructed. In order to do this, however, the translator must have access to a large quantity of data from which to identify language norms, and that data comes in the form of general reference corpora. Smaller corpora are simply inadequate when it comes to dealing with stretches of text, fixed and semi-fixed phrases, and less-frequently used words and expressions, though they serve a fundamental role in the identification of genre-related phenomena.

6. Discussion

The use of comparable general reference corpora as an aid to the translation process is often one-sided. TL corpora are often used as a control to ensure that the translation produced sounds natural, but less use is made of corpora in assessing the naturalness of the SL original. While must be acknowledged that absolute equivalence remains an elusive and rare phenomenon, adopting a data-assisted approach facilitates the identification of patterns and preferences across languages, making it possible to match up SL and TL expressions that are functionally as well as formally similar.

Choice in translation is related to choice in the SL, and this can be identified by comparing chosen expression against its possible alternatives along the paradigmatic axis. In this way the translator obtains a more instant and detailed impression of the meaning being conveyed, and if an equivalent paradigm of choice is set up for the TL, the most suitable correspondences can be identified and used in the translated text.

Delexicalisation is fundamentally important as a concept when translating both conventional and unconventional language. There is a distinct difference between the meaning values of words in conventionalized utterances and their values in non-standard uses of the language, and this should be acknowledged and acted upon when translating text. Corpus data highlights the conventional and recurrent, and the rarity of unusual structures stands out. An awareness of the norms underlying non-standard and unconventional language makes it possible for the translator to recreate its effect in a structured and systematic way, rather than rely on the “intuition and hunch, inspiration and even flashes of genius” (Firth 1968: 85) that seem integral to the translation process.

Corpus tools are becoming increasingly sophisticated, and word profiling is therefore a much more straightforward matter than it was a few years ago. By teaching trainee translators how and when to use these tools their sensitivity to language patternings will be heightened, and their translations will improve in accuracy and fluency. A combination of automatic processing, manual analysis and greater awareness of how languages make meaning, will give translators the chance to have equivalence at their fingertips.

References

Aston, G. and Piccioni, L. 2004. “Un grande corpus di italiano giornalistico.” In *Atti del convegno nzionale AitLA*, G. Bernini, G. Ferrari and M. Pavesi (eds.). Perugia: Guerra. Available from http://www.sslmit.unibo.it/~guy/aitla_repubblica.htm (accessed 03 April 2006).

Baker, M. 1992. *In Other Words: A Coursebook on Translation*. London/New York: Routledge.

Berlin, B. and Kay, P. 1969. *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press.

Biber, D. 1993.“Representativeness in Corpus Design.” *Literary and Linguistic Computing* 8 (4): 243-257.

Firth, J.R. 1968 “A Synopsis of Linguistic Theory, 1930-5.” In *Selected papers of J.R. Firth 1952-1957*, F.R. Palmer (ed.), 168-205. London/Harlow: Longmans.

Halliday, M.A.K. 1992. “Language Theory and Translation Practice.” *Rivista internazionale di tecnica della traduzione* 0 (pilot issue): 15-25.

Johansson, S. 1998. “On the role of corpora in cross-linguistic research”. In *Corpora and cross-linguistic research,* S. Johansson and S. Oksefjell (eds), 3-24. Amsterdam: Rodopi.

Kenny, D. 2001. *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester: St. Jerome.

Kilgarriff, A. and Tugwell, D. 2002. “Sketching words.” In *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, Marie-Hélène Corréard (ed), 125-137. Göteborg: EURALEX.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. “The Sketch Engine.” *Proceedings of the Eleventh EURALEX International Congress*, 105-116. Lorient: Université de Bretagne-Sud.

Krishnamurthy, R. 2000. “Collocation: from silly ass to lexical sets.” In *Words in Context: A Tribute to John Sinclair on his Retirement*, C. Heffer and H. Sauntson (eds), 31-47. Birmingham: The University of Birmingham.

Laviosa, S. 1998. “The English Comparable Corpus (ECC): A Resource and a Methodology.” In *Unity in Diversity? Current Trends in Translation Studies*, L. Bowker, M. Cronin, D. Kenny and J. Pearson (eds), 101-112. Manchester: St. Jerome.

Laviosa, Sara. 1997. “How Comparable Can ‘Comparable Corpora’ Be?” *Target* 9 (2): 289-319.

Louw, W.E. 2000. “Some implications of progressive delexicalisation and semantic prosodies for Hallidayan metaphorical modes of expression and Lakoffian ‘Metaphors we Live By’.” privately-distributed version of “Progressive delexicalization and semantic prosodies as early empirical indicators of the death of metaphors” paper read at the 11th Euro-International Systemic Functional Workshop: Metaphor in systemic functional perspectives, University of Gent (Belgium), 14-17 July 1999.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon.

Niemeier, S. (1998) “Colourless green ideas metonymise furiously” *Rockstocker Beträge zur Sprachwissenschaft* 5, 119-146.

Philip. G. 2003. *Collocation and Connotation: a corpus-based investigation of colour words in English and Italian.* PhD thesis. The University of Birmingham, UK.

Ragazzini, G. (ed.) 1995. *Il Ragazzini: dizionario inglese italiano - italiano inglese* (3rd edition). Bologna: Zanichelli.

Rossini Favretti, R. (2000) "Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS." In *Linguistica e informatica: corpora, multimedialità e percorsi di apprendimento*, R. Rossini Favretti (ed.), 39-56. Rome: Bulzoni.

Siepmann, D. 2005. "Collocation, colligation and encoding dictionaries. Part 1: Lexicological aspects." *International Journal of Lexicography* 18 (4): 409-443.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J.M. 1996. "The Search for Units of Meaning." *TEXTUS* 9 (1), 75-106.

Tognini Bonelli, E. 2000. "Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach." In *Lexis in Contrast*, B. Altenberg and S. Granger (eds.), 73-95. Amsterdam and Philadelphia: John Benjamins.

Tognini Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

Váradi, T. and Kiss, G. 2001. "Equivalence and Non-equivalence in Parallel Corpora." *International Journal of Corpus Linguistics* 6 (special issue): 167-177.