Corpus Linguistics and Translation Studies
Implications and Applications

Mona Baker
Cobuild, Birmingham

Abstract
The rise of corpus linguistics has serious implications for any discipline in which language plays a major role. This paper explores the impact that the availability of corpora is likely to have on the study of translation as an empirical phenomenon. It argues that the techniques and methodology developed in the field of corpus linguistics will have a direct impact on the emerging discipline of translation studies, particularly with respect to its theoretical and descriptive branches. The nature of this impact is discussed in some detail and brief reference is made to some of the applications of corpus techniques in the applied branch of the discipline.

0. Introduction

A great deal of our experience of and knowledge about other cultures is mediated through various forms of translation, including written translations, sub-titling, dubbing, and various types of interpreting activities. The most obvious case in point is perhaps literature. Most of us know writers such as Ibsen, Dostoyevsky and Borges only through translated versions of their works. But our reliance on translation does not stop here. Our understanding of political issues, of art, and of various other areas which are central to our lives is no less dependent on translation than our understanding of world literature.

Given that translated texts play such an important role in shaping our experience of life and our view of the world, it is difficult to understand why translation has traditionally been viewed as a second-rate activity, not worthy of serious academic enquiry, and why translated texts have been regarded as no more than second-hand and distorted versions of 'real' texts. If they are to be studied at all, these second-hand texts are traditionally analysed with the

sole purpose of proving that they inevitably fall short of reproducing all the glory of the original. A striking proof of the low status accorded to translated texts comes from the young but by now well-established field of corpus linguistics. A recent survey commissioned by the Network of European Reference Corpora, an EEC-funded project, shows that many corpus builders in Europe specifically exclude translated text from their corpora.[1] This is presumably done on the grounds that translated texts are not representative and that they might distort our view of the 'real' language under investigation. It is perhaps justifiable to exclude translated texts which are produced by non-native speakers of the language in question, but what justification can there be for excluding translations produced by native speakers, other than that translated texts per se are thought to be somehow inferior or contrived? Biased as it may be, this traditional view of translation implies, in itself, an acknowledgement of the fact that translational behaviour is different from other types of linguistic behaviour, quite irrespective of the translator's mastery of the target language.

The starting point of this paper is that translated texts record genuine communicative events and as such are neither inferior nor superior to other communicative events in any language. They are however different, and the nature of this difference needs to be explored and recorded. Moreover, translation should be taken seriously by related disciplines such as linguistics, literary theory and cultural and communication studies, not least because these disciplines can benefit from the results of research carried out in the field of translation. At the same time, as a phenomenon which pervades almost every aspect of our lives and shapes our understanding of the world, the study of translation can hardly be relegated to the periphery of other disciplines and sub-disciplines, those listed above being no exception. What is needed is an academic discipline which takes the phenomenon of translation as its main object of study. For many scholars, this discipline now exists. Some refer to it as the 'science of translation', others as 'translatology', but the most common term used today is 'translation studies'.

Eco (1976:7) distinguishes between a discipline and a field of study. The first has "its own method and *a precise object*" (my emphasis). The second has "a repertoire of interests that is not as yet completely unified". It could be argued that translation studies is still largely a 'field of study' in Eco's terms. The vast majority of research carried out in this, shall we say emerging discipline, is still concerned exclusively with the relationship between specific source and target texts, rather than with the nature of translated text as such. This relationship is generally investigated using notions such as equivalence,

correspondence, and shifts of translation, which betray a preoccupation with practical issues such as the training of translators. More important, the central role that these notions assume in the literature points to a general failure on the part of the theoretical branch of the discipline to define its object of study and to account for it. Instead of exploring features of translated texts as our object of study, we are still trying either to justify them or dismiss them by reference to their originals.

It is my belief that the time is now ripe for a major redefinition of the scope and aims of translation studies, and that we are about to witness a turning point in the history of the discipline. I would like to argue that this turning point will come as a direct consequence of access to large corpora of both original and translated texts, and of the development of specific methods and tools for interrogating such corpora in ways which are appropriate to the needs of translation scholars. Large corpora will provide theorists of translation with a unique opportunity to observe the object of their study and to explore what it is that makes it different from other objects of study, such as language in general or indeed any other kind of cultural interaction. It will also allow us to explore, on a larger scale than was ever possible before, the principles that govern translational behaviour and the constraints under which it operates. Therein lie the two goals of any theoretical enquiry: to define its object of study and to account for it.

Section 1 below offers an overview of the emerging discipline of translation studies and explains why translation scholars are now in a position to use the insights gained from corpus linguistics, and some of the techniques developed by it, to take translation across the threshold of 'field of study' and into the realm of fully-fledged disciplines.

1. Translation studies: the state of the art

1.1 Central issues: the status of the source text and the notion of equivalence

Until very recently, two assumptions dominated all discussions of translation and were never questioned in the literature. The first is that of the primacy of the source text, entailing a requirement for accuracy and faithfulness on the part of the translator. The second is a consequence of the first and is embodied in the notion of equivalence which has been the central concern of all discourse on translation since time immemorial. Translations should strive to be as equivalent to their originals as possible, with equivalence being understood

mainly as a semantic or formal category. The implied aim of all studies on translation was never to establish what translation itself is, as a phenomenon, but rather to determine what an ideal translation, as an instance, should strive to be in order to minimise its inevitable distortion of the message, the spirit, and the elegance of the original.

The essentialist question of how equivalence per se might be established in the course of translation has gradually been tempered by experience and by an explosion in the amount and range of texts which have come to be translated in a variety of ways on a regular basis. Hence, we now have a massive amount of literature which attempts to classify the notion of equivalence in a multitude of ways, and the question is no longer how equivalence might be achieved but, increasingly, what kind of equivalence can be achieved and in what contexts. This in itself is a noticeable improvement on the traditionally static view of equivalence, but it still assumes the primacy of the source text and it still implies that a translation is merely a text striving to meet the standards of another text.

1.2 Developments which support a move towards corpus-based research

The attempt to extend and classify the notion of equivalence has brought with it a need to explore not only the source text as the model to be adhered to but also the target language, and the specific target language text type, in order to give meaning to such categories as stylistic equivalence and functional equivalence. If the idea is not simply to reproduce the formal structures of the source text but also to give some thought, and sometimes priority, to how similar meanings and functions are typically expressed in the target language, then the need to study authentic instances of similar discourse in the two languages becomes obvious.

There have been other developments which have played a more direct role in preparing the ground for corpus work. One such development is the decline of what we might call the semantic view of the relationship between source and target texts. For a long time, discourse on translation was dominated by the idea that meaning, or messages, exist as such and can, indeed should, be transferred from source to target texts in much the same way as one might transfer wine from one glass to another. The traditional dichotomy of translating word-for-word or sense-for-sense is a product of this view of meaning. At about the same time that the notion of equivalence began to be reassessed, or perhaps a little earlier, new ideas began to develop about the nature of meaning in translation. Firth (1968:91) was among the first to sug-

gest that, difficult though as it may appear, an approach which connects structures and systems of language to structures and systems in the context of situation (as opposed to structures and systems of thought) is more manageable and "more easily related to problems of translation". Similarly, Haas (1968:104) stresses that, in practice, correspondence in meaning amounts to correspondence in use and asserts that "unless we can succeed in thus explaining translation, the mystery of bare and neutral fact will continue to haunt us". Two expressions are equivalent in meaning if and only if "there is a correspondence between their uses" (ibid). The importance of this change in orientation, from a conceptual to a situational perspective and from meaning to usage, is that it supports the push towards descriptive studies in general and corpus-based studies in particular. Conceptual and semantic studies (in the traditional sense) can be based on introspection. Studies which take the context into consideration, and even more so, studies which attempt to investigate usage, are, by definition, only feasible if access is available to real data, and, in the case of usage, to substantial amounts of it.

Apart from the decline of the semantic view of translation, another, and very exciting, development has been the emergence of approaches which undermine both the status of the source text vis-à-vis the translated text and the value of the very notion of equivalence, particularly if seen as a static relationship between the source and target texts. The move away from source texts and equivalence is instrumental in preparing the ground for corpus work because it enables the discipline to shed its longstanding obsession with the idea of studying individual instances in isolation (one translation compared to one source text at a time) and creates a requirement which can find fulfilment in corpus work, namely the study of large numbers of texts of the same type. This is precisely where corpus work comes into its own.

1.2.1 New perspectives: polysystem theory

In the late seventies, Even-Zohar, a Tel-Aviv scholar, began to develop a theory of literature as a polysystem, that is as a hierarchical and dynamic conglomerate of systems rather than a disparate and static collection of texts. A given literary polysystem is seen as part of a larger cultural polysystem, itself consisting of various polysystems besides literature, for example politics and religion. These polysystems are structured differently in different cultures.

Polysystem theory has far-reaching implications for the status of translated literature in general and for the status of the source text vis-à-vis the target text in particular. First, the theory assumes a high level of interdependence among the various systems and sub-systems which underlie a

given polysystem, as well as among the polysystems of literature in various cultures. This means that, for instance, "literature for children would not be considered a phenomenon sui generis, but related to literature for adults" and, similarly, "translated literature would not be disconnected from original literature" (Even-Zohar 1979:13). As a consequence, the status of translated literature is elevated to the point where it becomes worthy of investigation as a system in its own right, interacting with its co-systems and with the literary polysystems of other cultures. By recognising translated literature as a system in its own right, polysystemists shifted the attention away from individual literary translations as the object of literary studies to the study of a large body of translated literature in order to establish its systemic features.

Second, one of the main properties of the polysystem is that there is constant struggle among its various strata, with individual elements and systems either being driven from the centre to the periphery or pushing their way towards the centre and possibly occupying it for a period of time (ibid:14). This constant state of flux suggests that no literary system or sub-system is restricted to the periphery by virtue of any inherent limitations on its value. Thus, the approach stresses that translated literature may, and sometimes does, occupy a central position in the polysystem and is therefore capable of providing canonised models for the whole polysystem. Moreover, given that polysystem theory recognises that intra- and inter-relations exist within both systems and polysystems, leading to various types of interference and transfer of elements, models, canons, and so on, it becomes obvious that "semiliterary texts, translated literature, children's literature - all those strata neglected in current literary studies - are indispensable objects of study for an adequate understanding of how and why transfers occur within systems as well as among them" (ibid: 25). And finally, polysystemists reject the popular view of translation as a derivative activity and stress instead that literary translation is "a creatively controlled process of acculturation in that translators can take an original text and adapt it to a certain dominant poetics or ideology in the receiving culture" (Heylen 1993:21).

This view of literature as a conglomerate of systems, as well as the growing interest in transfer and interference across systems, has gradually undermined the status of the source text in translation studies. Since the early eighties, Toury, another Tel-Aviv scholar, has been stressing that a translation belongs to one textual system only, namely the target system, and the source text has gradually been assuming the role of a stimulus or source of information rather than the starting point for analysis. Questions regarding how a translated text came into being or what type of relationship it has with a given

source text are becoming secondary to its classification as part of the target textual system. As Toury puts it in a more recent publication (1985:19):

> It is clear that, from the standpoint of the source text and source system, translations have hardly any significance at all, even if everybody in the source culture 'knows' of their factual existence ... Not only have they left the source system behind, but they are in no position to affect its linguistic and textual rules and norms, its textual history, or the source text as such. On the other hand, they may well influence the recipient culture and language, if only because every translation is initially perceived as a target-language utterance.

It is worth noting that similar, though not quite so radical, assessments of the status of the source text have also emerged among groups of scholars not specifically concerned with literary translation. For example, Vermeer (1983: 90)[2] suggests that the function of the translated text is determined by the interests and expectations of its recipients and not by the function of the source text. The SL text is a source of information and, like other sources of information, it may be exploited in a variety of ways to meet the expectations of an envisaged audience.

1.2.2 From equivalence to norms

From the late seventies onwards, the source-oriented notion of equivalence has been gradually replaced by notions which clearly take the target system and culture as a starting point. Some of these notions have evolved within theories designed to account for translation within a commercial environment. They include, for example, Vermeer's notion of coherence, defined as the agreement of a text with its situation (Vermeer 1983), and Sager's definition of equivalence as a function of the specifications that accompany a request for translation (Sager 1993). The most important, however, has been the notion of norms, introduced by Toury (1978, 1980).

Toury has developed a tripartite model in which norms represent an intermediate level between competence and performance. If we think of competence as an inventory of all the options that are available to translators in a given context, and performance as the subset of options which are actually selected by translators from this inventory, then norms are a further subset of these options. They are options which are regularly taken up by translators at a given time and in a given socio-cultural situation. In this sense, the notion of norms is very similar to that of typicality, a notion which has emerged from recent work on corpus-based lexicography and which contrasts sharply with the standard, absolute dualisms in linguistics: competence and performance,

langue and parole.

Norms, then, are a category of descriptive analysis. They can be identified only by reference to a corpus of source and target texts, the scrutiny of which would allow us to record strategies of translation which are repeatedly opted for, in preference to other available strategies, in a given culture or textual system. The concept of norms tips the balance not only in favour of the target text (as opposed to the traditional obsession with the source text), but, more important, it assumes that the primary object of analysis in translation studies is not an individual translation but a coherent corpus of translated texts. Norms do not emerge from a source text or a body of source texts. Equally, they do not emerge from the target system nor from a general collection of target texts. They are a product of a tradition of translating in specific ways, a tradition which can only be observed and elaborated through the analysis of a representative body of translated texts in a given language or culture. They can therefore be seen not just as a descriptive category but also as providing a functional, socio-historical basis for the structure of the discipline (Lambert 1985:34).

1.2.3 The rise of descriptive translation studies

Since the seventies, several scholars have begun to express dissatisfaction with the heavy reliance on introspective methods in translation studies. Holmes (1988:101) makes the point most clearly:

> Many of the weaknesses and naiveties of contemporary translation theories are a result of the fact that the theories were, by and large, developed deductively, without recourse to actual translated texts-in-function, or at best to a very restricted corpus introduced for illustration rather than for verification or falsification.

Newman (1980:64) similarly suggests that the way out of the dilemma posed by the notions of equivalence and translatability is to look at actual instances of translation and to determine, on the basis of those instances, "the kind of generalities that might form the basis of a theory of competence or systematic description". It is however Toury who has done more to elaborate the concept of descriptive translation studies than anyone else in the discipline.

For Toury, it is vital for translation studies to develop a descriptive branch if it is ever to become an autonomous discipline. Without this, translators will continue to rely on other disciplines such as linguistics to provide them with theoretical frameworks and the means to test their hypotheses. Descriptive Translation Studies, or DTS for short, is not reducible to a collection of case studies or comparative analyses of source and target texts. It is

that branch of the discipline which must provide a sound methodology and explicit research procedures to enable the findings of individual descriptive studies to be expressed in terms of generalisations about translational behaviour. Its agenda consists, primarily, of investigating what translation is "under any defined set of circumstances ... and WHY it is realized the way it is" (Toury 1991a:186). One of its main objectives is to render the findings of individual studies intersubjective and to make the studies themselves "repeatable, either for the same or for another corpus" (Toury 1980:81).

It is perhaps worth noting at this point that although the words *corpus* and *corpora* are beginning to figure prominently in the literature on translation, they do not refer to the same kind of corpora that we tend to talk about in linguistics. Corpora in translation studies have so far been very modest affairs. Their size is not generally expressed in terms of number of words but of number of texts, and they are searched manually. For example, Vanderauwera (1985) is a study of "50 or so novels" translated from Dutch into English in "roughly the period 1960-1980" (ibid:1-2). This is a very small corpus, and yet the experience of searching it manually leads Vanderauwera to suggest that "serious and systematic research into translated texts is a laborious and tiresome business" (ibid:6). Toury himself seems torn between the need to set an ambitious program for DTS and the recognition that "the larger and/or more heterogeneous the corpus, the greater the difficulties one is likely to encounter while performing the process of extraction and generalization" (1980:66-7). In an earlier publication, Toury (1978:96) argues for a distributional study of norms based on statistical techniques but concludes that

> ... as yet we are in no position to point to strict statistical methods for dealing with translational norms, or even to supply sampling rules for actual research (which, because of human limitations, has nearly always been applied to samples only, and will probably go on being carried out in much the same way). At this stage we must be content with our intuitions ... and use them as keys for selecting a corpus and for hitting upon ideas.

One of John Sinclair's major achievements for linguistics has been his success, through the collection of computerised corpora and the development of a relevant research methodology, in providing ways of overcoming our human limitations and minimising our reliance on intuition. His work can provide solutions for precisely the kind of problems that translation scholars are still struggling with today.

2. Corpus work in translation studies: the potential

There is no doubt that the availability of corpora and of corpus-driven methodology will soon provide valuable insights in the applied branch of translation studies, and that the impact of corpus-based research will be felt there long before it begins to trickle into the theoretical and descriptive branches of the discipline. Sinclair (1992:395) touches very briefly, and strictly from the point of view of a linguist, on one obvious application:

> The new corpus resources are expected to have a profound effect on the translations of the future. Attempts at machine translation have consistently demonstrated to linguists that they do not know enough about the languages concerned to effect an acceptable translation. In principle, the corpora can provide the information.

In the above statement, which is one of the very few Sinclair has made on translation, the concern is merely with improving the performance of translators and of machine translation systems in terms of approximating to the structures and natural patterns of a given language or languages. This same concern underlies most of the expressions of interest in corpus studies which are beginning to take shape in the literature.[3] It is of course a legitimate concern and one which will be shared widely by scholars within and outside translation studies, theorists and practitioners alike. I would, however, like to think that the 'profound effect' which Sinclair refers to will not be understood merely in terms of knowing enough about the languages concerned to approximate to their patterns. After all, once we are in a position to describe and account for our object of study, namely translation, we might find that approximating to the patterns of the target language, or any language for that matter, is not necessarily as feasible as we seem to assume, and that it is not the only factor at play in shaping translational behaviour. Several scholars, for example Toury (1991b:50) and Even-Zohar (1979:77) have already noted that the very activity of translating, the need to communicate in translated utterances, operates as a major constraint on translational behaviour and gives rise to patterns which are specific to translated texts. Thus Even-Zohar (ibid) stresses that "we can observe in translation patterns which are inexplicable in terms of any of the repertoires involved", that is patterns which are not the result of interference from the source or target language. Examples of these patterns are discussed as *universal features of translation* in section 2.1 below. The profound effect that corpora will have on translation studies, in my view,

will be a consequence of their enabling us to identify features of translated text which will help us understand what translation is and how it works. The practical question of how to improve our translations will find more reliable and realistic answers once the phenomenon of translation itself is explained in its own terms.

Practical applications aside, what kind of queries can access to computerised corpora help us resolve in our effort to explicate the phenomenon of translation? Given that this question, to my knowledge, has not been addressed before, what follows has to be seen as a very tentative list of suggestions which can provide a starting-point for corpus-based investigations in the discipline but which do not, by any means, address the full potential of corpora in translation studies.

2.1 Universal features of translation

The most important task that awaits the application of corpus techniques in translation studies, it seems to me, is the elucidation of the nature of translated text as a mediated communicative event. In order to do this, it will be necessary to develop tools that will enable us to identify universal features of translation, that is features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems.

It might be useful at this point to give a few examples of the type of translation universals I have in mind. Based on small-scale studies and casual observation, a number of scholars have noted features which seem, intuitively, to be linked to the nature of the translation process itself rather than to the confrontation of specific linguistic systems. These include:

(i) A marked rise in the level of explicitness compared to specific source texts and to original texts in general (see for instance Blum-Kulka 1986:21). In Baker (1992), I discussed several examples of translations which build extensive background information into the target text. In one case (*Autumn of Fury: the Assassination of Sadat* by Mohamed Heikal; 1983:3), a simple clause - *The example of Truman was always present in my mind* - is rendered into Arabic as follows:

> In my mind there was always the example of the American President Harry Truman, who succeeded Franklin Roosevelt towards the end of World War II. At that time - and after Roosevelt - Truman seemed a rather nondescript and unknown character who could not lead the great human struggle in World War II to its desired and inevitable end. But Truman - faced with the

> challenge of practical experience - grew and matured and became one of the most prominent American presidents in modern times. I imagined that the same thing could happen to Sadat.

Toury (1991b:51) accepts that 'explicitation'[4] is a feature of all kinds of mediated events, including interaction in a foreign language, but wonders whether there are any differences in the level and nature of explicitation by, for instance, language learners vs. translators, professional vs. non-professional translators, or in oral vs. written translation. The techniques currently available in corpus linguistics can in principle cope with such tasks as the measurement of expansion in two or more corpora. Moreover, specific conventions and software tools are now available for recording and retrieving automatically such things as information about a writer or speaker, for example his/her first language, nationality, and gender.

(ii) A tendency towards disambiguation and simplification. For example, Vanderauwera (1985:97-8) notes that, in her corpus of English translations of Dutch novels, potentially ambiguous pronouns are replaced by forms which allow more precise identification, and difficult syntax is made easier. Similarly, she reports that "where quotation marks fail to distinguish a person's speech or thought in the source text, they are almost invariably restored in the target text" (ibid:94). Again, corpus linguistics now has the necessary tools to measure such things as simple vs. complex syntax, and the automatic retrieval of pronominal as opposed to precise forms of identification, as well as elements of punctuation such as quotation marks, is a fairly simple operation in terms of current corpus techniques.

(iii) A strong preference for conventional 'grammaticality'. In interpreting, that is oral translation, this manifests itself in an overriding tendency to round off unfinished sentences, 'grammaticise' ungrammatical utterances and omit such things as false starts and self-corrections, even those which are clearly intentional in a courtroom context (Shlesinger 1991:150). Vanderauwera (1985:93) records a similar tendency towards general textual conventionality in a corpus of English translations of Dutch novels.

(iv) A tendency to avoid repetitions which occur in source texts, either by omitting them or rewording them (Shlesinger 1991, Toury 1991a). Toury (ibid:188) reports this feature as "one of the most persistent, unbending norms in translation in all languages studied so far".

(v) A general tendency to exaggerate features of the target language. For example, binomials composed of synonyms or near-synonyms, which are a common feature of Hebrew writing, tend to occur more frequently in translated than in original Hebrew texts and to replace non-binomials in source

texts (Toury 1980:130). Vanderauwera (1985:11) suggests that translations overrepresent features of their host environment in order to make up for the fact that they were not originally meant to function in that environment.

(vi) Point (v) above not withstanding, it has been shown that the process of mediation often results in a specific type of distribution of certain features in translated texts vis-à-vis source texts and original texts in the target language. For example, Shamaa (1978:168-71) reports that common words such as *say* and *day* occur with a significantly higher frequency in English texts translated from Arabic than they do in original English texts. At the same time, their frequency of occurrence in the translated English texts is still considerably lower than the frequency of the equivalent Arabic items in the source texts. Shamaa suggests that, although subtle and elusive, this unusual distribution of features contributes to the identification of a text as a translation and "leave[s] a vague impression of being culturally exotic" (ibid:172). It is a symptom of what is sometimes referred to as "the third code" (Frawley 1984:168), which is a result of the confrontation of the source and target codes and which distinguishes a translation from both source texts and original target texts at the same time.[5] Another example of the operation of the 'third code' is provided by Blum-Kulka (1986:33) who suspects that research is likely to reveal "that cohesive patterns in TL texts are neither TL nor SL norms oriented, but form a system of their own". Research into the nature of the third code in translation might also provide answers to queries about the nature of 'pseudotranslations', that is texts which are regarded as translations but for which no genuine source texts exist.

It will clearly take some time and ingenuity to develop a corpus methodology for capturing universal features of the type discussed above. But, assuming that we have a corpus of texts translated into, say, English from a variety of languages, we might attempt to isolate patterns which

(i) occur across the corpus, irrespective of whether the source texts are French, Hebrew or Chinese,

(ii) do not occur, or do not occur to the same degree/with the same frequency, in original English texts .

We then need to compare our results with the results of similar research carried out on corpora of translated texts in, let us say, French and German. As a starting point, this will help us identify certain patterns or tendencies which occur in the three corpora and which, therefore, are good candidates for universal features of translation. Further research on similar corpora in other languages will either confirm or disprove our hypotheses.

2.2 Translational norms operating in a given socio-cultural context

Universal features such as those discussed in 2.1 above can be seen as a product of constraints which are inherent in the translation process itself, and this accounts for the fact that they are universal (or at least we assume they are, pending further research). They do not vary across cultures. Other features have been observed to occur consistently in certain types of translation within a particular socio-cultural and historical context. These are a product of norms of translation which represent another type of constraint on translational behaviour (see section 1.2.2 above). Like universal features of translation, textual exponents of translational norms can be efficiently and reliably investigated if access to computerised corpora is made available to scholars in the field.

Toury (1978) distinguishes between various types of norms which operate at different stages of the translation process. The techniques of corpus linguistics are particularly suited to the identification of what Toury calls 'operational norms': norms which "affect the matrix of the text, that is, the modes of distributing linguistic material [matricial norms] ..., and the actual verbal formulation of the text [textual norms]" (ibid:87). Matricial norms cover the occurrence of omissions, additions, substitutions and transpositions in translated texts. For instance, Toury (1980:127) reports that in translations of prose fiction from English and German into Hebrew during the years 1930-45, there was a tendency to omit substantial subtitles (that is subsidiary rather than main titles of literary works), and that this contrasts with a tendency to add subtitles of a certain type in literary translation into French in the early eighteenth century.

Textual norms have been discussed more widely in the literature, and examples include the following:

(i) Toury (1980:129) reports, in the same corpus of Hebrew translations of prose fiction, a high level of dependence on a repertory of fixed collocations derived from canonised religious texts. The same corpus also reveals that special importance was attached to direct speech: pieces of dialogue were regularly turned into independent paragraphs, indirect speech was replaced by direct speech, and phrases which indicated a move from narration to dialogue were omitted.

(ii) In a corpus of 425 mystery books translated into Hebrew since the early sixties, Toury (ibid:104) identifies a strong norm which he expresses as "The title should never be too complex, witty or sophisticated". This norm manifests itself in two ways. First, sophisticated titles (in terms of lexis) are re-

placed by simple titles which contain one of a stock of items that include the Hebrew equivalents of 'mystery', 'murder', 'blood', 'death', and so on. Thus, *The Case of the Ice-Cold Hands* becomes "The Mystery of the Murder in the Motel". Second, the Hebrew titles follow a limited range of simple syntactic patterns, notably 'the x of y', as in the above example, and 'the x's y' as in "The Black Candle's Mystery".

(iii) In a corpus of Dutch novels translated into English, Vanderauwera (1985:93) notes that foreign words and dialogue which occur in the source texts are either replaced in the translation by target language items (for example Malay *klamboe* is replaced by *mosquito curtain*) or glossed in the target language.

(iv) Based on a study of a limited corpus of translations of modern, non-literary English texts in a variety of languages, I have suggested (Baker 1992:36) that Japanese seems far more tolerant of the use of loan words in translation than, for instance, Arabic and French. The norms that govern translational behaviour in the three languages are noticeably different in this regard.

2.3 Other issues for corpus research

Apart from universal and norm-oriented features of translational behaviour, access to computerised corpora should enable us to explore a number of other theoretical issues which are difficult to deal with on the basis of small-scale studies.

First, there is the question of the intermediate stages of translation, or how the final product evolves over a period of time. Recent developments in the working practices of professional translators will soon make this type of research feasible. For instance, the Institute of Translation and Interpreting of Great Britain is currently considering a proposal to adapt a variant of the British Standard BS5750[6] for professional translators working in a variety of contexts: as staff translators, on their own or through agencies or co-operatives. If adopted, this standard will require translators to follow certain routines in documenting the work they receive and despatch, whether in manuscript or disk form, and to maintain clearly labelled and dated versions of each translation. Access to this type of text in electronic form can be used to explore the process of translation through a retrospective analysis of successive versions of the product. New software tools, including tools for fuzzy matching (that is identifying stretches of text which are similar but not identical), are currently being developed by corpus linguists. Svanholm (1992) reports that IBM already offer their translators fuzzy matching software as a basic tool for

revising and updating existing translations of large documents.

Second, central questions which have been on the agenda for decades can be resolved more efficiently and reliably through the investigation of large computerised corpora. These questions include the size and nature of the unit of translation, the type of equivalence which is achieved in practice and the level at which it is achieved. Again, new software tools are now available for the investigation of parallel corpora, that is corpora of source texts and their translations (see for example Marinai et al 1991, Brown & Cocke 1988, Church & Gale 1991). New and sophisticated methodologies are also evolving for investigating the nature and limits of equivalence on the basis of comparable corpora, that is corpora of comparable original texts in several languages. Sinclair has pioneered the investigation of comparable corpora through a project on multilingual lexicography (Sinclair 1991).

3. Conclusion

I have argued in this paper that translation studies has reached a stage in its development as a discipline when it is both ready for and needs the techniques and methodology of corpus linguistics in order to make a major leap from prescriptive to descriptive statements, from methodologising to proper theorising, and from individual and fragmented pieces of research to powerful generalisations. Once this is achieved, the distinction between the theoretical and applied branches of the discipline will become clearer and more convincing.

There is now an urgent need to explore the potential for using large computerised corpora in translation studies. It seems to me that most of the components for realising this potential are in place. The emphasis has shifted from meaning to usage, and the notion of equivalence is gradually giving way to that of norms. The status of the source text has been undermined and we have managed to make the leap from source-text-bound rules and imperatives to descriptive categories. There is increasing interest in features of translated texts per se and we are beginning to develop a descriptive branch of the discipline with well-defined objectives and an explicit program. What we need is a research methodology and a set of tools that can help us put this program into action. A suitable methodology and a set of very powerful and adaptable tools are now available from corpus linguistics.

Author's Address:
Cobuild, Westmere, 50 Edgbaston Park Road, Birmingham. B15 2RX. U.K.

Notes

1. INL Working Paper 92-11. J. G. Kruyt & E. Putter, Corpus Design Criteria: report submitted to the European Commission by the Instituut voor Nederlandse Lexicologie, Leiden, as a contribution to a European enquiry into corpus design criteria.
2. I do not have access to the original German text quoted in the references, but have relied on an unpublished translation by Professor J. C. Sager, UMIST, of short extracts from various German sources.
3. For example, Lindquist (1984) is a corpus-based study of adverbials which aims not at explicating translational behaviour but at providing a 'translation grammar' which can guide translators' choices.
4. This is the term used in the literature.
5. In some cases, when an unusual distribution of features is clearly a result of the translator's inexperience or lack of competence in the target language, this phenomenon is referred to as 'translationese'.
6. BS5750 is a British standard for quality assurance. It specifies a set of controls which have to be implemented in order for an organisation to claim that it is 'quality assured'. It is therefore a form of company certification, guaranteeing that the final product marketed by a given company meets the national standard.

References

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

Blum-Kulka, S. 1986. "Shifts of Cohesion and Coherence in Translation". *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* ed. by J. House & S. Blum-Kulka, 17-35. Tübingen: Gunter Narr.

Brown, P. & J. Cocke. 1988. "A Statistical Approach to Language Translation". *Proceedings of the 12th International Conference on Computational Linguistics: COLING '88* ed. by D. Vargha, 71-6. Budapest.

Church, K. & W. Gale. 1991. "Concordances for Parallel Text". *Using Corpora: Proceedings of the Seventh Annual Conference of the UW Centre for the New OED & Text Research.* St. Catherine's College: Oxford.

Eco, U. 1976. *A Theory of Semiotics.* Bloomington & London: Indiana University Press.

Even-Zohar, I. 1978. "The Position of Translated Literature within the Literary Polysystem." *Literature and Translation* ed. by J. S. Holmes, J. Lambert & R. van den Broeck 117-27. Leuven:ACCO.

_____. 1979, 1990. "Polysystem Theory." *Poetics Today (Special Issue on Polysystem Studies by Itamar Even-Zohar)* 11, 1. 9-26.

Firth, J. R. 1956, 1968. "Linguistics and Translation" *Selected Papers of J. R. Firth 1952-59* ed. by F. R. Palmer, 84-95. London: Longman. (first read at Birbeck College, London in 1956).

Frawley, W. 1984. "Prolegomenon to a Theory of Translation". *Translation: Literary, Linguistic and Philosophical Perspectives* ed. by W. Frawley, 159-75. London & Toronto: Associated University Presses.

Haas, W. 1968. "The Theory of Translation" *The Theory of Meaning* ed. by G. H. R. Parkinson , 86-108. Oxford: Oxford University Press.

Heylen, R. 1993. *Translation, Poetics, and the Stage: Six French Hamlets.* London & New York: Routledge.

Holmes, J. S. 1988. *Translated! Papers on Literary Translation and Translation Studies.* Amsterdam: Rodopi.

Lambert, J. 1991. "Shifts, Oppositions and Goals in Translation Studies: Towards a Genealogy of Concepts." *Translation Studies: The State of the Art* ed. by K. M. van Leuven-Zwart & T. Naaijkens, 25-37. Amsterdam: Rodopi.

Marinai, E., C. Peters & E. Picchi. 1991. "Bilingual Reference Corpora: A System for Parallel Text Retrieval". Unpublished manuscript. Istituto di Linguistica Computazionale: Pisa, Italy.

Lindquist, H. 1984. "The Use of Corpus-based Studies in the Preparation of Handbooks for Translators". *Translation Theory and its Implementation in the Teaching of Translating and Interpreting* ed. by W. Wilss & G. Thome, 260-70. Tübingen: Gunter Narr.

Newman, A. 1980. *Mapping Translation Equivalence.* Leuven: ACCO.

Sager, J. C. 1993. *An Extended Communicative Theory of Translation.* Unpublished Manuscript, University of Manchester Institute of Science and Technology.

Shamaa, N. 1978. *A Linguistic Analysis of Some Problems of Arabic to English Translation.* D. Phil. Thesis, Oxford University.

Shlesinger, M. 1991. "Interpreter Latitude vs. Due Process. Simultaneous and Consecutive Interpretation in Multilingual Trials". *Empirical Research in Translation and Intercultural Studies* ed. by Sonja Tirkkonen-Condit, 147-55. Tübingen: Gunter Narr.

Sinclair, J. M. 1991. *Council of Europe Multilingual Lexicography Project.* Unpublished report submitted to the Council of Europe under contract no. 57/89.

_____.1992. "The Automatic Analysis of Corpora". *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991* ed. by J. Svartvik 379-97. Berlin & New York: Mouton de Gruyter.

Toury, G. 1978. "The Nature and Role of Norms in Literary Translation." *Literature and Translation* ed. by J. S. Holmes, J. Lambert & R. van den Broeck 83-100. Leuven: ACCO.

_____. 1980. *In Search of a Theory of Translation.* Tel Aviv: The Porter Institute for Poetics and Semiotics.

_____. 1985. "A Rationale for Descriptive Translation Studies". *The Manipulation of Literature: Studies in Literary Translation* ed. by T. Hermans, 16-41. London & Sydney: Croom Helm.

_____. 1991a. "What are Descriptive Studies into Translation Likely to Yield apart from Isolated Descriptions." *Translation Studies: The State of the Art* ed. by K. M. van Leuven-Zwart & T. Naaijkens 179-92. Amsterdam: Rodopi.

_____. 1991b. "Experimentation in Translation Studies: Achievements, Prospects and Some Pitfalls". *Empirical Research in Translation and Intercultural Studies* ed. by Sonja Tirkkonen-Condit, 45-66. Tübingen: Gunter Narr.

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

Vermeer, H. J. 1983. *Aufsätze zur Translationstheorie.* Heidelberg: Groos.