March 1993

Corpora and Translation: Uses and Future Prospects

Tony McEnery and Andrew Wilson

I. Introduction

Although corpora have been an object of study for some decades, the nineteen eighties saw an increased interest in their use and construction. With this increased interest and awareness has come an expansion in the application areas for which corpus based approaches have been deemed relevant. This paper will seek to define the concept of a corpus, and discuss its relevance to two application areas in particular, automatic and manual translation.

II. Corpora

A corpus, simply defined, is a large body of text. Corpora may exist in machine readable form or in their natural state as written texts or recorded speech, but increasingly the term “corpus” is used to refer to the machine readable variety.

Machine readable corpora have a number of advantages over other forms of storage. Firstly, and most importantly, machine readable corpora may be searched and manipulated in ways which are simply not possible with the other formats. Secondly, machine readable corpora can be swiftly and easily enriched with additional information. But so far corpora have been discussed as though they are an undifferentiated mass. This is not the case: corpora can be adapted in many ways. In the following sections the varying forms that a corpus may take will be discussed, begining with one basic distinction which bifurcates the concept of the corpus; should they be unannotated or annotated?

A. Enriched corpora — unannotated or annotated?

Unannotated corpora are corpora which are left in their ‘raw’ (or ‘pure’) state. These corpora may be of use for many purposes, but it is important to note that retrieval from such a corpus may take a more linguistically expert user than would otherwise be required. Also, programs manipulating and processing such data will either be:

1. More ‘intelligent’, to be able to identify some of the implicit linguistic categories and relations. Even where such an intelligent program is successful, it inevitably leads to an increase in cost and processing time.
2. Less useful — the programs may not succeed in identifying this implicit information, and consequently their functionality is limited.

Thus although untagged corpora have their uses, the range of functionality for automated retrieval and manipulation of corpora is greatly enhanced by the provision of annotation in a corpus.

Annotated corpora are corpora to which additional information (especially linguistic information) has been added (Leech 1992). This usually involves the attachment of some kind of coding to the machine readable language material itself.

Many types of linguistic information may be encoded in a corpus:

1. Part of speech

Part-of-speech (or POS) tagging, also known as grammatical tagging, entails the assignment of a part of speech to every lexical item in a corpus. One automatic POS tagging suite, CLAWS (Garside 1987), uses a set of part-of-speech codes (up to 169 codes in size) to mark such information in texts. An example of one of these codes is ‘NN1’, used to denote a singular common noun. A POS tag is attached to each word in the corpus by means of a symbol such as an underscore. So in an annotated corpus the sequence ‘dog_NN1’ may be read as being composed of two parts — the word ‘dog’ and the part of speech ‘singular common noun’. The following shows an example of POS tagging from the Lancaster-Oslo/Bergen (or LOB) corpus:

```
EXAMPLE OF PART-OF-SPEECH TAGGING FROM LOB CORPUS:

ˆ another_DT new_JJ style_NN feature_NN is_BEZ
the_ATI wine-glass_NN or_CC flared_JJ heel_NN ,_,
which_WDT was_BEDZ shown_VBN teamed_VBN
up_RP with_IN pointed_JJ ,_, squared_JJ ,_, and_CC
chisel_NN toes_NNS ._.

ˆ colour_NN is_BEZ highly_RB important_JJ in_IN
choosing_VBG autumn_NN footwear_NN ._. ˆ the_ATI
autumn_NN range_NN of_IN shades_NNS is_BEZ almost_RB
bewildering_JJ ,_, and_CC there_EX are_BER some_DTI
exciting_JJ new-comers_NNS ,_, such_IN as_IN"
conker_NN calf_NN and_CC charcoal_NN ,_, rocco_NN
and_CC Russian_JNP violet_NN ._.
```

Key:

ATI	article neutral for number

BEDZ	*was* (past sing. form of the verb BE)
BER	*are* (present plural form of the verb BE)
BEZ	*is*, *'s* (-*s* form of the verb BE)
CC	coordinating conjunction
DT	singular determiner
DTI	determiner neutral for number
EX	existential *there*
IN	preposition
JJ	general adjective
JNP	adjective with word-initial capital
NN	singular common noun
NNS	plural common noun
RB	general adverb
RP	adverb which can also be a particle
VBG	present participle of lexical verb
VBN	past participle of lexical verb
WDT	WH-determiner

[In this text, IN" indicates the second part of the two-word syntactic idiom *such as* which has the function of IN.]

POS annotation can be important in translation for several reasons. Firstly, it may be used as a preliminary to the disambiguation of homographs. It cannot differentiate word senses, but it can disambiguate word function and sometimes this can amount to the same thing: for example, *booted* as an adjective means “wearing boots” but as a verb means “kicked”. In an empirically-based machine translation system such as that proposed by Brown et al. (1990), which employs a translation model using bilingual alignment at the word level, this provides additional information about where alignments are and are not likely to constitute “correct” translations. Secondly, a syntactic idiom such as the subordinator *so that* cannot always be translated through a simple word-to-word alignment. In German, for example, the English *so that* (2 words) may be translated as *damit* (1 word). POS tagging which uses ‘ditto tags’[1] facilitates the alignment of many-to-one and one-to-many examples such as this.

2. Syntactic structure — parsing

Parsing involves the assignment of surface structure to a text, normally using a form of phrase structure grammar. Typically the constituents are indicated using labelled brackets rather than tree-like structures, though sometimes an attempt is made to provide some graphic realization of structure (cf. Marcus and Santorini 1992). The following is an example of a type of parsing using a very small set of constituent types; for this reason, it is sometimes known as “skeleton parsing” (Leech and Garside 1991).

[1] A scheme whereby a multi-word sequence such as *so that*, which has a single syntactic function, is assigned just one part of speech and each constituent word in the idiom is differentiated only by an index (cf. Blackwell 1987).

```
EXAMPLE OF SKELETON PARSING FROM THE SPOKEN ENGLISH CORPUS

[N The_AT first_MD book_NN1 [[N he_PPHS1 N][V took_VVD
[P from_II [N the_AT library_NN1 N]P]V]]N][V was_VBDZ
[N[G Darwin_NP1 ’s_$ G][ ’_" [N Origin_NN1 [P of_IO
[N Species_NN N]P]N] ’_" [Fr[N which_DDQ N][V inspired_VVD
[N him_PPHO1 N][P with_IW [N the_AT dream_NN1
[P of_IO [Tg becoming_VVG [N a_AT1
geologist_NN1 N]Tg]P]N]P]V]Fr]]N]V] ._.
```

Syntactic constituents are bounded by labelled square brackets. POS tags are linked to their words by the underscore character, “_”.

The tagset used in this example is the “CLAWS2” tagset. This is a later version of the CLAWS1 (or LOB) tagset illustrated above. The symbols for the various parts of speech are broadly similar to the CLAWS1 tags, for example common noun tags still begin with NN and adjectives with J.

The symbols for the syntactic constituents in this example are:

Fr	relative clause
N	noun phrase
P	prepositional phrase
Tg	*-ing* clause
V	verb phrase

Where unlabelled brackets occur, this indicates a constituent-type which is not included in the reduced set of constituents employed in skeleton parsing. Such indeterminacy is allowed for in the parsing guidelines used by the human analysts.

Most corpus-based research on machine translation has relied on alignment at the word level, or alternatively at the sentence level which is easy to perform on the basis of punctuation. However, Brown et al. (1990) recognize the potential of alignment at the level of syntactic constituents, which may enable the induction of a computational phrase structure grammar and hence subsequently alignment and translation at the level of the grammatical constituent. Clause-level alignment is also important in a bilingual knowledge base approach to machine translation such as is advocated by Sadler (1989).

3. Word sense

In addition to POS tagging and parsing, it is also possible to annotate semantic features in corpora. Two types of semantic annotation may be identified: the representation of word senses, normally using some form of sense classification (rather like a thesaurus), and the marking of more structural semantic relations such as agent/patient structures. The latter type of annotation is not often encountered at present, but is likely to increase in the near future. Word sense annotation is also quite rare, but is an active area of current research. This form of tagging has not been carried out to any large extent on corpora in other languages, but deserves to be as it enables large quantities of structured lexical data to be extracted.

The following is an example of semantic word-tagging, taken from the automatic content analysis described by Wilson and Rayson (1991):

EXAMPLE OF WORD SENSE TAGGING

AT1	The	Z5
MC	one	N1
NN1	disadvantage	A5.1-
IO	of	Z5
JJ	woolen	O1.1
NN2	clothes	B5
VBZ	is	A3+
CST	that	Z5
PPHS2	they	Z8
VM	can	A7+
VV0	become	A2.1
JJ	uncomfortable	O4.2-
II	in	Z5
RG	very	A13.3
JJ	hot	O4.6+
NN1	weather	W4

In this example, the text is read vertically. The first column contains the POS tags, the second column the words of the text, and the third column the semantic tags. The semantic tags are composed of:

1. an upper case letter indicating general discourse field
2. a number indicating a first subdivision of this discourse field
3. (optionally) a decimal point followed by a further number to indicate a finer subdivision
4. (optionally) one or more pluses or minuses indicating opposites and degrees of intensity on a semantic scale.

For example, the tag O4.6+ indicates a word in the field “Physical Objects and Properties” (O), in the subcategory “Physical Attributes” (O4), in the sub-subcategory “Temperature” (O4.6) and “hot” (+) rather than “cold” (–).

Word sense annotation is a crude basis for translation *per se*. A translation based on the fact that w_1 in L1 belongs to the same thesaurus class as w_2 in L2 does not necessarily entail that w_2 is a good translation of w_1. However, word sense annotation can be of use in creating term banks and multilingual thesauri from large quantities of text, to function as translation tools in other approaches to machine and machine-aided translation. The use of measures such as mutual information further enables the extraction of collocations at the level of semantic rather than lexical information.

4. Anaphoric relations

Anaphoric annotation indicates the co-reference of noun phrases and pronominals in text. This is an important, but not frequently encountered, type of annotation. It has been carried out successfully over some 100,000 words of English by Lancaster University in collaboration with IBM T.J. Watson Research Center, New York. Again, it has not to our knowledge been carried out on corpora in languages other than English, but its potential in modelling pronominal reference suggests that it ought to be.

An example of this form of annotation is as follows:

```
ANAPHORIC ANNOTATION OF AP NEWSWIRE

{21 (4 Civic Center 4) Director 21}} (21 Frank E. Russo Jr. 21)
said <21 he was confident (4 the $31.5 million coliseum 4)
would be ready to open as scheduled.

"There’s no turning back now", <21 he said.

Tickets for (167 (7 the Whalers 7)’ first game in (25 <7 their
home city 25) in two years 167) have been selling briskly.
```

The use of the same index number in the above indicates the co-reference of constituents. In the following, the letter *n* is used to represent an index number in the actual notation:

(n n) OR [n...] enclose a constituent (normally a noun phrase) entering into an equivalence chain

<n indicates a pronoun with a preceding antecedent

{n n}} enclose a noun phrase entering into a copular relationship with a following noun phrase

Anaphoric annotation is of particular value in research on automated pronoun translation. In order to translate a pronominal which does not enter into a one-to-one relationship with a pronominal in the target language, one requires two sets of information:

1. The antecedent of the pronominal, where one exists.
2. The number and gender of the antecedent.

Pronominals typically inherit number and gender from their antecedents[2]. Sometimes it is easy to translate from one language to another: for example, English *she* aligns unproblematically with French *elle* as feminine singular (nominative). However, the French translation of English *they* depends on whether the plural group is all female, or includes one or more male person. This may not always be clear from the text, but if the English pronoun refers to a phrase *the girls* then it is obvious that it should be translated as *elles* and not *ils*. Anaphoric annotation enables empirical research to be carried out into automatic pronoun resolution, including the examination of exceptions to general rules, and thus it will then be possible to attempt to overcome this particular translation problem.

III. Parallel Corpora

Parallel corpora are, in a very real sense, best characterized as the 'Rosetta Stone' of modern corpus linguistics. These are corpora which hold the same text in more than one language. Typically, at present, these parallel corpora are bilingual rather than multilingual.

There is a general paucity of annotated parallel corpora. A very few do exist, such as the Canadian Hansard (a parallel corpus in French and English of the proceedings of the Canadian Parliament) and a corpus of IBM Technical Manuals (English and French), but they tend to be of limited value because of restrictions of domain and availability.

Research is limited to these corpora and language pairs alone, e.g. Brown et al. (1990, forthcoming), which is hardly satisfactory. Further, their potential for yielding large automated lexicons (Garside and McEnery, forthcoming), as will be discussed later, remains largely unexploited.

But before any further discussion of the use of the corpora can gainfully take place, one necessary refinement of the form taken by a parallel corpus needs to be considered — alignment.

IV. Parallel Aligned Corpora

It is clear that simply having a corpus composed of two parallel subcorpora poses as many problems as it solves. Which sentences are translations of which? Below that level, which word (or words) are translations of which word (or words)? An aligned corpus tackles this problem, by aligning sentences which are mutual translations of one another. It may also, below the sentence level, align word units that are translations. So within a corpus we may see the sentences '*C'est magnifique, mais ce n'est pas la guerre*' and 'It is magnificent but it is not war' aligned together. Below that level we may see further alignments. '*C*' may be aligned with 'It', '*la guerre*' may be aligned with 'war' and '*mais*' may be aligned with 'but'.

[2] This is not, however, always the case. The most important exceptions are *anaphoric islands* (cf. Oakhill and Garnham 1992) and *conceptual anaphors* (cf. Oakhill et al. 1992).

This form of alignment may be achieved with a high degree of accuracy automatically, using such statistical techniques as mutual information. These techniques are currently being refined within project ET 10-63 (section 6). It is intended to further develop these techniques in the future. To give two examples of techniques that may be used to improve alignment not exploited by ET 10-63:

1. Using part-of-speech tagging to align at the level of grammatical function rather than at word string level.
2. Smoothing any skewed probabilities by using statistics not only from the current corpus, but other successfully aligned corpora.

V. Uses of Parallel Aligned Corpora

The uses of parallel aligned corpora are potentially many, but two obvious areas would be machine translation and lexicon construction.

A. Machine translation

Brown et al. (1990, forthcoming) attempt to build upon the success of probabilistic methods in other areas of language processing and apply them to the problem of machine translation. They have produced a probabilistic machine translation system trained on an aligned French-English corpus. This system chooses the most probable translation sentence in the target language given a sentence in the source language using two probability models: a trigram language model based on three-word sequences, originally developed for a speech recognition system, and a translation model derived from the word-level alignment of their English and French parallel subcorpora and information about word positions within the corpus sentences.

Sadler (1989) proposes an alternative approach in which a very large bilingual database is constructed, with each language parsed using a form of dependency grammar. The resulting units are then aligned between the languages involved. Translation is carried out by isolating possible units in the source text, retrieving these units and their translations in the database, and combining the retrieved translation units. The work carried out by Tsujii (1992) on machine translation by example also requires parallel aligned corpora for its operation.

B. Lexicon construction

It is possible to extract correspondences between languages not only at the word level, but also above the level of the word. Multiword units would also be retrieved from a parallel aligned corpus, making multilingual dictionary building an easier task.

If the corpus is machine readable, it can also be scanned for frequent collocations. With a specialized corpus it is also possible to construct terminology databases.

Multilingual parallel aligned annotated corpora open up many possibilities for future development. Yet their full potential may only become known when an end user has the opportunity to actively exploit such a corpus. That presupposes the existence of one.

VI. The ET 10-63 project

ET 10-63 is a project under the EC's EUROTRA programme currently running at IBM Paris, C2V Paris, Essex University and Lancaster University. Its aim is to develop a large, part-of-speech tagged, bilingual parallel aligned corpus of EC telecommunications texts and carry out work on lexicon building techniques, including term extraction and argument frame extraction.

This project has developed a data model to store parallel aligned bilingual corpora as databases (e.g. McEnery and Daille 1993). This has the advantages that:

1. A standard for corpus storage may be created.
2. The powerful query languages used with databases, such as SQL and other relational query languages, can be used to retrieve linguistic information from annotated corpora.
3. Data may be shared more readily between sites, as the data model remains constant.

A new proposal has recently been formulated, called the MACE project. MACE seeks to extend and widen the scope of this research, taking it beyond the bilingual domain, and producing the first multilingual parallel aligned POS-annotated corpus. Greek, French, Italian and English are the languages that the corpus hopes to cover. The corpus will be based upon the Official Journal of the EC, as this is an excellent source of parallel texts. Also, the Official Journal is available in all of the official languages of the EC, so the corpus would be readily extensible beyond its original four core languages.

The ET 10-63 model will be extended in the near future to cover the form of corpora proposed within the MACE project so that a standard for the storage of multilingual parallel aligned corpora can be achieved.

VII. Conclusion

This paper has outlined the nature of corpora, and has concentrated specifically upon the applications of corpora within the field of translation studies. This influence is growing as the availability of relevant corpora increases. It would seem that this trend is set to continue, especially when it is considered that corpora are seen to be an increasingly significant part of the CEC's strategy for the development of linguistic engineering in the next decade as is stated in the Technical Background Document for the LRE Call for Proposals 1992:

> "The availability of large, duly classified and annotated text corpora is a sine qua non for any linguistic R&D work."
>
> CEC (1992:15)

Hence the presence of corpora in translation studies, as well as other areas of linguistic study, seems destined to become ever greater.

References

Blackwell, S. (1987). Syntax versus Orthography: Problems in the Automatic Parsing of Idioms. In: R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. & Roosin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics* 16:2, 79-85.

Brown, P., Della Pietra, V., Della Pietra, S., Lafferty, J., & Mercer, R. (forthcoming). Analysis, Statistical Transfer and Synthesis in Machine Translation. To appear 1993.

CEC. (1992). *LRE Programme Call for Proposals 1992: Technical Background Document*. Luxembourg: CEC.

Garside, R. (1987). The CLAWS Tagging System. In: R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R. and McEnery, A. (forthcoming). Treebanking: The Compilation of a Corpus of Skeleton-Parsed Sentences. To appear in: E. Black, G. Leech and R. Garside (eds), *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach*. Amsterdam: Rodopi.

Leech, G. (1992). Corpus Annotation Schemes. Paper presented at Pisa Workshop on European Corpus Resources, January 1992

Leech, G. and Garside, R. (1991). Running a Grammar Factory. In: S. Johansson and A.-B. Stenström, *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter.

Marcus, M. and Santorini, B. (1992). Building Very Large Natural Language Corpora: The Penn Treebank. Unpublished manuscript.

McEnery, A. and Daille, B. (1993). Database Design for Corpus Storage: The

ET 10-63 Data Model. *Unit for Computer Research on the English Language Technical Papers* 1.

Oakhill, J. and Garnham, A. (1992). Linguistic Prescriptions and Anaphoric Reality. *Text* 12:2, 161-182.

Oakhill, J., Garnham, A., Gernsbacher, M. and Cain, K. (1992). How Natural Are Conceptual Anaphors? *Language and Cognitive Processes* 7:3/4, 257-280.

Sadler, V. (1989). The Bilingual Knowledge Bank — A New Conceptual Basis for MT. Utrecht: BSO-Research.

Tsujii, J., Ananiadou, S., Carroll, J. & Sekine, S. (1991). Methodologies for the Development of Sublanguage MT System II. CCL, UMIST Report No. 91/11.

Wilson, A. and Rayson, P. (1991). The Automatic Content Analysis of Spoken Discourse. Paper presented at 12th International Conference on English Language Research on Computerized Corpora, Ilkley, May 1991.