[*Translating and the Computer 19*: Papers from the Aslib conference… 13 & 14 November 1997]

Multilingual Corpora - Current Practice and Future Trends

Dr. Tony McEnery,
Department of Linguistics and Modern English Language,
Lancaster University,
Bailrigg,
Lancaster,
LA1 4YT.

A.Mcenery@lancaster.ac.uk

Introduction

In this paper I would like to give an overview of multilingual corpus building to date. In doing so, I will review two types of multilingual corpus, parallel and translation corpora. Following this, I will consider what tools are currently available which allow for the exploitation of such corpora in the context of machine/machine aided translation. Throughout I will give a fairly global view of work in this area, but will concentrate largely on work undertaken at Lancaster, a major centre of multilingual corpus construction and exploitation.

Following this overview, I will give an indication of how I see multilingual corpus construction developing, and will specifically highlight the need for corpora of non-autochthonous European languages.

Multilingual Corpus Construction to Date

The resources under discussion, multilingual corpora and their associated exploitation technologies, have a potentially wide impact on at least two major fields, namely machine translation (Sebba, 1991) and bilingual lexicography. To date the exploitation of multilingual corpora, and the claims arising from that exploitation, have been based upon relatively few language pairs. A variety of multilingual corpus exploitation paradigms have developed, such as those of Gale and Church (1992), Church (1993) and Johansson et al (1993). However, because the available multilingual corpora are few, work in the area has reached a point where it is in danger of training, evaluating and assessing itself on a subset (e.g. English/French, English/Spanish, English/German) rather than across a varied sample of world languages. This is a point which will be returned to later. For the moment I would like to discuss the range of data which has been generated to date, and consider the uses it has been put to.

The data developed to date can broadly be divided into two forms – parallel corpora and translation corpora (also known as comparable corpora). Parallel corpora are composed

of a set of L1 texts, and an equivalent set of L2 translations of those texts. Translation corpora, on the other hand, are composed of L1 texts from language *A*, balanced, genre for genre, against L1 corpora from language *B*.[1] There are some terminological problems in specifying the two types of corpora under discussion, however. The definition of parallel and translation corpora presented here are those which are current in computational linguistics. Difficulties arise when we consider work done in contrastive linguistics, e.g. by Johansson and Hofland (1998), where the distinction remains the same, but the terminology is switched. In this paper we will use 'parallel corpus' and 'translation corpus' according to its usage in computational linguistics, but as there is an increasing flow of ideas and work between contrastive and multilingual corpus based computational linguistics, this terminological problem is one well worth being aware of. With this stated, let us now consider what parallel and translation corpora have been developed to date.

Parallel

In the field of applied linguistics, *parallel aligned corpora*[1] have become increasingly important. They are revolutionising research in machine translation (Brown et al, 1991, 1993), second language teaching (Botley, Glass, McEnery & Wilson, 1996) and comparative linguistics (Altenberg & Aijmer, 1993), orienting study towards language in use and away from speculation about usage.

A number of important projects are currently under way generating basic resources for the teaching and automation of translation (see Botley, McEnery and Wilson, 1998). By way of illustrating the types of resources that are being generated, we can look at how corpus construction at Lancaster has developed over the past ten years. Lancaster has been a major centre for the construction and annotation of corpora of Modern British English for more than two decades. This has been demonstrated through pioneering corpus work funded by the SERC/EPSRC, such as the LOB corpus and, more recently, the British National Corpus. Corpus construction and annotation at Lancaster has expanded significantly over time by way of the development, during the 1990s, of CEC-sponsored parallel aligned annotated corpora.

Lancaster has taken a leading role in the construction of parallel corpora on three major CEC parallel corpus building projects:

1. On project ET10/63[2], the Lancaster team was responsible for the cleaning, lemmatisation and part-of-speech annotation of a 1.5 million word corpus of French/English parallel technical texts, used to in automated term extraction experiments by IBM France, as reported by Gaussier, Langé & Meunier (1992), Gaussier (1995) and Daille (1995).

---

[1] A parallel aligned corpus takes a parallel corpus, and says, for example, at the level of the sentence, which sentences in the original text translate into which sentences in the translated text.

[2] Developed under the Eurotra programme, project number ET10/63.

2. The CRATER[3] project, of which Lancaster was the co-ordinating partner, had several aims related to the production of parallel corpora:
   a. to make a public domain version of the ET10/63 corpora.
   b. to add a third language to that corpus, namely Spanish.
   c. to generate a part-of-speech tagger for Spanish, and to use that tagger for the part-of-speech tagging of the Spanish corpus.
   d. to generate sentence, word and term alignment software, and
   e. to hand correct the tagging of the English, French and Spanish corpora.

   The CRATER project achieved all of these aims in the eighteen months for which it ran, and the corpus resources generated have been used by Sharp Laboratories of Europe in the training of more accurate part-of-speech taggers for English, French and Spanish (Ellworthy, 1995).
3. On MULTEXT[4], Lancaster was subcontracted to produce alignment software for use with all nine languages covered by the project, and to develop a bilingual English/French corpus of 200,000 words, morphosyntactically tagged and aligned. Both the tagging and alignment were hand verified.

In addition, Lancaster has also worked in the development and exploitation of Chinese/English parallel aligned corpora (McEnery & Oakes, 1996, McEnery, Piao & Xu, 1998), and is currently collaborating with the University of Lodz, Poland, on the construction of English/Polish multilingual corpora.

So parallel corpora are being generated, and European languages now have an increasing range of parallel resources available to them. But what of translation corpora?

Translation

Translation corpora are currently being constructed throughout Europe. For example:

- The CEC sponsored PAROLE project is undertaking corpus collection which could broadly be described as translation corpus collection for all official EC languages.
- The English-Norwegian parallel (the term being used here in its contrastive sense) corpus contains translation corpora for English and Norwegian in a range of genres.
- Lancaster and Lodz University are currently gathering a translation corpus of Polish to match the British National Corpus category for category.

Parallel corpora, as noted in the previous section, are being exploited for a variety of translation related tasks. But translation corpora have been developed to overcome problems of artificiality and error which are sources of potential problems with corpora

---

[3] Developed as part of the Corpus Resources and Terminology Extraction project (CRATER) of the MLAP programme (grant number MLAP 93-20). of which Lancaster was the co-ordinating partner.
[4] Funded under the CEC's LRE Programme, this project ran from 1/1/94 to 31/8/96.

incorporating translated material. A good example of this can be found in a corpus of healthcare documentation currently under construction at Lancaster.

A recent study at Lancaster of English/Polish translation in healthcare documentation has revealed that the translated text (Polish) reads very much as a translation, and not as L1 Polish text[5]. While the language in the translation is grammatically and pragmatically correct, it is quite obvious that the text is a translation. This is most noticeable in the use of direct Polish equivalents of English words which would not be used in similar contexts by Polish speakers – the connotations are inappropriate:

English: *People who are overweight experience difficulties*
Polish Translation: *Osoby z nadgwa doswiadczaja trudnosci z*
Polish: *Osoby otyle odczuwaja*

Overweight is translated as *with overweight* rather than as *otyle (overweight).* Also, the direct Polish translation of *experience*, *doswiadczyc*, is more to do with external experiences rather than internal/bodily ones, for which one would use *odczuc.* Further to this, some unusual lexical preferences occur, such as:

English: *Currently*
Polish Translation: *Wspolczesnie*
Polish: *Obecnie*

English: *Every/each year*
Polish Translation: *Kazdego roku*
Polish: *Co roku*

English: *contains information*
Polish Translation: *zamieszcza informacje*
Polish: *zawiera informacje*

The evident flavour of a translation is present here. Unusual syntactic choices are evident in the use of finite clauses as opposed to the participial clauses of natural Polish, the use of prepositional constructions instead of inflectional ones. There are also too many analytical constructions in general in the translations provided.

Another point worthy of mention is that Subject/Verb/Object and Subject/Verb ordering are more common than one would expect in the Polish, giving evidence again for a noticeable translation effect. The effect of this to a Polish reader is that new information is rendered via a preverbal subject much more often than would be the case in natural Polish, where such a subject is much more likely to be post-verbal, as noted by Siewierska (1993), who documents that whereas only 24% of the subjects in SVO clauses convey new information, of the clause final ones 79% are new.

---

[5] My thanks to Professor Anna Siewierska for her help with this study.

So the reasons why people may wish to compare and exploit L1 texts only is clear – incorporating L1 texts within any multilingual corpus based system would be to permit the possibility of incorporating inaccurate and/or unrepresentative data. Yet parallel corpora continue to be the subject of research and construction. Of importance to understanding why this is the case, is considering the sustainability of projects aimed at the construction of parallel and translation corpora. While translation corpora are highly attractive because of the 'naturalness' of the data they contain (all material within such a corpus being L1 material) populating such a corpus can be difficult – Johansson et al (1993) found that it was not possible to populate a fully balanced corpus of Norwegian/English, because some genres of L1 Norwegian writing did not exist[6], and may only exist as L2 translations from English. Also, parallel corpora are attractive as via alignment they bring the possibility of bootstrapping example-based machine translation systems.

It is clear that translation quality is an issue which affects the exploitation of parallel corpora, and this is a factor which must be taken seriously when parallel corpora are being constructed. Yet while their intrinsic usefulness remains high, it is unlikely that translation corpora will supplant them entirely.

Multilingual corpus exploitation

We have already mentioned at least one important exploitation tool which increases the usefulness of parallel corpora – corpus alignment. Alignment can occur at many levels, but at the moment sentence and word alignment are by far the most common forms of alignment available[7]. In order to exemplify current work in alignment I intend once more to outline work at Lancaster, and then to set this work in a broader context by a wider review. Following this I will consider how corpus retrieval tools (concordancers) are developing to allow humans to exploit multilingual corpora.

Alignment

As well as constructing parallel corpora, Lancaster has also exploited those corpora. **Sentence alignment software** has been developed and tested, based upon a variety of techniques as described by Gale and Church (1992), Kay and Röscheisen (1993) and Garside, Hutchinson, Leech, McEnery & Oakes (1994) to allow users to align the corpora as a prelude to exploitation in, for example, lexicographic research. Table one gives a series of results obtained at Lancaster using the Gale and Church technique for sentence alignment (taken from McEnery and Oakes, 1996). Effective and efficient **word alignment software** has been developed as a further aid to corpus exploitation, based upon both approximate string matching techniques and co-occurrence statistics. Table 2 gives an example of the results achieved on the English/Spanish language pair using Dice's similarity coefficient (McEnery & Oakes, 1995). **Multi-word unit alignment software** has also been developed, based upon the work of Gaussier, Langé & Meunier (1992), whose work has been extended to cover the English/Spanish and French/Spanish

[6] This problem is exactly the same as Shastri (1986:xii-xiii) reported in trying to build an Indian English equivalent of the Brown and LOB corpora, the so called Kohlapur Corpus.
[7] Though sentence alignment is much more effective than word alignment.

language pairs. Some results which illustrate this are given in figure one below (taken from McEnery, Nieto-Serrano & Smalley 1996).

Language Pair	*Domain*	*Paragraphs Tested*	*% Correct*
English-French	Telecommunications	100	98.0
English-German	Economics	36	75.0
English-Polish	Fiction	89	100.0
English-Spanish	Telecommunications	222	93.2
Chinese-English	Newspaper	171	54.5[8]

Table 1: The success of sentence alignment

Dice Score	0.4-0.49	0.5-0.59	0.6-0.69	0.7-0.79	0.8-0.89	0.9-0.99	1.0
Accuracy	.411	.927	.902	1.00	1.00	1.00	1.00

Table 2: The success of word alignment in English/French using Dice's similarity coefficient

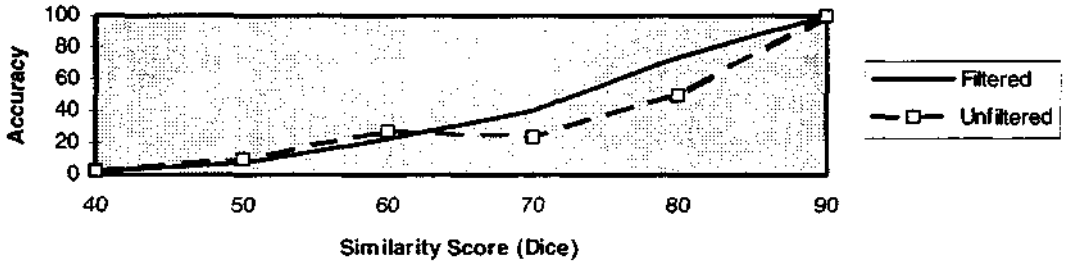

Figure 1: The success of compound noun alignment between English and Spanish using finite state automata and similarity data both with and without co-occurrence measures as an additional filter.

Alignment is seen as a key process in the exploitation of parallel corpora for a variety of purposes, including the following:

- Example-based machine translation
- Statistically-based machine translation (Brown et al 1990)
- Cross-lingual information retrieval (Melamed, 1996)
- Gisting of World Wide Web pages (Melamed, 1996)
- Computer-assisted language learning (Catizone *et al*, 1989, Warwick-Armstrong & Russell, 1990)

[8] Note that this relatively poor score for English/Chinese underlines our argument that a language pair sensitive approach to alignment is necessary.

All of the above applications of parallel corpora are, however, dependent to lesser or greater degrees on appropriate alignment software being available. Currently, alignment proceeds:

- On the basis of a statistical heuristic
- On the basis of linguistic rules
- On a combination of the two above.

Let's briefly consider these three approaches.

Statistical alignment techniques employ empirically justified heuristics to achieve alignment. Often simple quantitative measures can be used to determine sentence-level translation equivalents, given two parallel translations, with good results. For example, the approaches of Brown et al (1991) and Gale & Church (1991) depend on relative sentence lengths, based on the premise that long sentences in one language are more likely to be translations of long sentences in the other, while short sentences in one language are more likely to be translated by short sentences in the other. Melamed (1997) uses techniques originally developed in automated image processing in order to achieve sentence and word alignment.

The alternative approach is to employ linguistically motivated methods. These approaches are rationalistic, often inspired by what we might do intuitively when manually aligning texts. Linguistic methods are generally based on pairing lexical units which make up phrases, eventually accompanied by their dependency structures.

The statistical and linguistic approaches are not mutually exclusive, but are complementary and can be usefully hybridised. Statistical methods tend to work better for large corpora, since they are relatively rapid, while linguistic methods can be better for small corpora (Debili & Sammouda 1992). Further, the two techniques can be combined, for example, Chen (1993) found that the most reliable indicators for alignment were "critical parts of speech, namely nouns, verbs and adjectives" which could be used to support an otherwise statistical alignment method.

While the overview of alignment techniques presented here is, of necessity, brief and sketch like, it is possible to see from the work presented that a range of techniques have been developed to achieve alignment. Further to this, sentence alignment technology has proved to be fairly reliable on the language pairs it has been tested on so far. While word alignment is still not as reliable, advances have been made towards this goal also, and work on phrasal alignment is under way. Aligned data is clearly of use for machine translation tasks, but what of machine aided translation? At least part of the answer to this question lies in the development of parallel concordancers which allow translators to navigate parallel text resources.

Concordancing

Concordancing has long been the mainstay of corpus linguistics. A concordancer allows a string and/or related strings to be searched for in a corpus, and retrieved with an associated context. To make parallel corpora easy to exploit it is clear that we require a new type of concordancer, one which can deal with parallel aligned corpora. Although one could imagine carrying out two monolingual searches through the L1 and L2 texts of a parallel corpus, it would be of greater advantage to carry out a search in, say, L1, and as part of the retrieval of relevant context the program displayed both the L1 contexts and their L2 translations. This would be a way of providing translators with an on-line translation memory of sorts

With such a need identified, it is hardly surprising that multilingual concordancers are becoming available. The WordSmith program devised by Mike Scott at the University of Liverpool contains a rudimentary alignment algorithm. More sophisticated is the MultiConc program produced by Woolls (1998) at Birmingham. This uses a modified version of the Gale and Church alignment algorithm to align texts 'on the fly' as they are presented to the system. The system is also capable of working in a wide range of fonts. Less sophisticated is the ParaConc program deve oped by Barlow (1998) which allows multilingual Concordancing but only on texts which have been explicitly pre-encoded with alignment information. ParaConc has no on-line alignment facility. Of some interest is the growing number of web based multilingual corpus browsers, such as that of Peters, Picchi and Biagini (1998), which allow remote Concordancing of multilingual texts. All of these systems, especially those with on-line text alignment facilities, represent a possible way of translators exploiting available corpus data and data they themselves generate as a form of translation memory.

Multilingual corpora of the future

Having reviewed multilingual corpus building and exploitation in the present, I would now like to describe the programme of work I see ahead of corpus builders and computational linguists if the full promise of the work undertaken in this field to date is to be realised. Central to my vision of what must occur is the point I raised in the introduction – we need data in a wider range of languages.

The main problem with the exploitation of parallel corpora is that there is a major bottle-neck in the provision of suitable corpora (as discussed by McEnery and Oakes, 1996). There are *too few parallel corpora in existence*, covering *very few language pairs.* The net effect of this is that the efficacy and benefits of the work undertaken so far can only be assessed for a small group of language pairs where there are suitable resources. CEC projects promise partly to remedy this, but in terms of languages beyond Europe, and languages within Europe without official status, the available resources are close to non-existent and are likely to remain so: the CEC is planning no further multilingual corpus building activity in the near future as it has now constructed corpora for each of the official CEC languages.

A second problem is that the techniques developed to exploit such corpora have not been widely tested on a broad range of language pairs. As a consequence, a significant research effort is required to examine how techniques which work well in aligning sentences between languages closely related both genetically and typologically, for example, French and English, work in aligning more distant language pairs. McEnery, Piao & Xu (1998) has shown that in the case of English/Chinese a substantial re-think is needed in the process of alignment. It is easy to see that the same may be true of other combinations – a project is needed which develops the resources for such questions to be examined, and proceeds to examine them.

There is no immediate evidence of the data problem being solved outside of Europe and North America. For example, it is strange to relate but true that countries such as India and China have a tradition of corpus construction, but that most of the data generated to date has been English Language data. Shastri (1986) in India constructed an Indian English corpus which matched the LOB and Brown corpora for balance. Jiao Tong University in Shanghai, China, has been working on English language corpora with John Sinclair's Birmingham team for some time, and the English language Ghuangzhou Petroleum Corpus[9] was constructed in China. Away from English language corpus construction, however, the work is patchier - work has been undertaken to try to deal with specific problems, e.g. word segmentation (the team at the computer science department in Quing Hua University, China has been particularly active here), word alignment (Ker & Chang, 1997) and sentence alignment (Wu, 1994), but in terms of the construction of publicly available corpora the picture is at best patchy and at worst desolate.

Some corpus resources are known to exist – e.g. rudimentary English/Chinese Parallel corpora (Wu, 1994). Further corpus building is planned, e.g. in China, where the Chinese Academic Society is planning to construct a Chinese language equivalent of the BNC in conjunction with the Beijing University of Language and Culture. Douglas Biber has worked on building and exploiting corpora of Somali, Tuvaluan and Korean (Biber 1995), the Korean Advanced Institute of Science and Technology is developing the Korean National Corpus and Daniel Ridings at Gothenburg is working on TEI conformant corpora of African languages[10]. So there is some interesting data around, even if some of it (as is the case of the corpus of Wu, 1994) is not available for general use.

Even if individual countries such as China do construct corpora, however, there will be some problems in terms of corpus resources for some UK domestic translation needs. When we consider specific languages which are used in the UK, but which do not have a strong political presence in their home country we see that British initiatives may be needed to produce corpus resources for that language. A good example of this is Cantonese. In China official language policy is geared towards Putonghua, and the

[9] A joint effort by the Hong Kong Science and Technology University and the Guangzhou Foreign Languages Institute.
[10] Specifically, one million word corpora of Shona and Ndebele.

corpus resources planned by the Chinese government will cover only that variety. The over-riding philosophy is that Putonghua is Chinese, and at the written level there is no difference between Putonghua and dialects such as Cantonese. This later statement is clearly not true – written Cantonese is different from written Putonghua (see Liu, 1990, Wah-Wei, 1993) and is also the majority form of Chinese spoken in Britain.

So while the advances made in multilingual corpus building and exploitation over the past decade are more than worthy of praise, this must not blind us to the fact that there remains a great deal of work to be done. A massive diversification in L1 and L2 language corpora is needed, and that expansion may well lead us to re-evaluate the work in multilingual corpus exploitation which has been undertaken to date.

Conclusion

In this paper my aim has been to give an informative overview of the state of the art in multilingual corpus linguistics both in terms of construction and exploitation. The relevance of multilingual corpora to machine based translation is clear. Yet unless that relevance is expanded rationally by the continued development of multilingual corpora and further advances in alignment technology, the full promise of multilingual corpora will not be realised.

Bibliography

Altenberg, B. & Aijmer, K. 1995. "Engelskan i ett kontrastivt perspektiv", in L.-G. Andersson and F. Börjeson (eds.) *Språkundervisning på universitet: rapport från ASLAs höstsymposium Göteborg, 11-13 november 1993,* Association Suedoise de Linguistique Appliquée, pp. 3-11.

Barlow, M. 1998. "Parallel texts in language teaching", in S. Botley, A. McEnery & A. Wilson (eds).

Biber, D. 1995. *Dimensions of Register Variation,* Cambridge University Press, Cambridge.

Botley, S., Glass, J., McEnery, A. & Wilson, A. 1996. *Proceedings of TALC 96,* UCREL Technical Papers, Lancaster University.

Botley, S.P., McEnery, A. & Wilson, A. 1998. *Multilingual Corpora – Teaching and Research,* Rodopi, Amsterdam.

Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation", *Computational Linguistics,* 19:2, pp. 263-312.

Brown, P.F.,Lai, J. & Mercer, R. 1991. "Aligning Sentences in Parallel Corpora", in *Proceedings of ACL-91.*

Catizone, R., Russell, G. & Warwick-Armstrong, S. 1989, "Deriving Translation Data from Bilingual Texts", in *Proceedings of the First International Lexical Acquisition Workshop,* Detroit.

Chen, S. F. 1993. "Aligning Sentences in Bilingual Corpora Using Lexical Information", in *Proceedings of ACL-93,* Columbus Ohio.

Church, K. W. 1993. "Char_align: A Program for Aligning Parallel Texts at the Character Level", in *Proceedings of ACL-93,* Columbus Ohio.

Daille, B. *1995.Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering.* UCREL Technical Papers Number 5, Department of Linguistics, Lancaster University.
Debili F. & Sammouda E. 1992. "Appariement des phrases de textes bilingues français-anglais et français-arabe ", in *Proceedings of COLING-92,* Nantes.
Ellworthy, D. 1995. *Using the CRATER Corpora to Train a Tagger,* CRATER project internal report.
Gale, W.A., & Church, K.W. 1993. "A Program for Aligning Sentences in Bilingual Corpora". *Computational Linguistics* 19:1, pp. 75-102.
Garside, R., Hutchinson, J., Leech, G.N., McEnery, A.M. & Oakes, M.P. 1994. "The exploitation of parallel corpora in projects ET10/63 and CRATER", in D.Jones (ed) *New Methods in Language Processing,,* UMIST, pp. 108- 115.
Gaussier, E., Langé, J.-M., & Meunier, F. 1992. "Towards Bilingual Terminology", *Proceedings of ALLC/ACH Conference,* Oxford, Oxford University Press, England, 121-124.
Gaussier, E. 1995. *Some Methods for the Extraction of Bilingual Terminology,* Ph.D. Thesis, Université de Paris Jussieu.
Johansson, S. & Hofland, K. 1998. "The English-Norwegian parallel corpus project: current work and new directions", in S. Botley, A. McEnery & A. Wilson (eds).
Johansson, S. & Hofland, K. 1993, "Towards an English-Norwegian Parallel corpus", in U. Fries, G. Tottie, and P. Schneider (eds.), *Creating and Using English Language Corpora,* Rodopi, Amsterdam, pp. 25-37.
Kay, M., & Röscheisen M. 1993. "Text-Translation Alignment", *Computational Linguistics,* 19:1, pp. 121-142.
Ker, S.J. & Chang, J.S. 1997. "A Class-based Approach to Word Alignment", *Computational Linguistics,* 23:2, pp 313-344.
Liu, K.L.P. 1990. "Language, power and education in Hong Kong : a sociolinguistic history", unpublished MA thesis, Lancaster University.
McEnery, A. & Oakes, M. 1995, "Sentence and word alignment in the CRATER project: Methods and assessment", in S. Armstrong-Warwick and E. Tzoukerman (eds.) *Proceedings of the EACL-SIGDAT Workshop,* Dublin, pp. 77-86.
McEnery, A.M. & Oakes, M.P., 1996 "Sentence and word alignment in the CRATER project", in J. Thomas & M. Short (eds) *Using Corpora for Language Research,* Longman, London, pp. 211 -231.
McEnery, A.M., Nieto-Serrano A. & Smalley, J.P. 1996 "Cognate Extraction Using Approximate String Matching Techniques",CRATER Project Internal Report.
McEnery, A.M., Piao, S.L. & Xu, X. 1998. "Parallel Alignment in English and Chinese", in S. Botley, A. McEnery, & A. Wilson (eds).
Melamed, I.D. 1996. "A Geometric Approach to Mapping Bitext Correspondence", to appear in the *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Philadelphia.
Peters, C., Picchi, E. & Biagini, L. 1998. "Parallel and comparable bilingual corpora in language teaching and learning", in S. Botley, A. McEnery, & A. Wilson (eds).
Sebba, M. 1991 "The Adequacy of Corpora in Machine Translation". *Applied Computer Translation* 1:1, pp. 15-27.

Shastri, S.V., Patilkulkarni, C.T. & Shastri G.S. 1986 *Manual of Information to Accompany the Kolhapur Corpus of Indian English for Use with Digital Computers*. Department of English, Shivaji University, Kolhapur.
Siewierska, A., (1993) "Syntactic weight versus information structure and word order variation in Polish", *Journal of Linguistics*, 29, pp 233-265.
Wai, P.W. 1993. "150 years on : linguicism in Hong Kong educational context – past, present and future", unplublished MA thesis, Lancaster.
Warwick-Armstrong, S. & Russell, G. 1990, "Bilingual Concordancing and Bilingual Lexicography", in *EURALEX 4th International Conference*, Malaga, Spain.
Woolls 1998 "From Purity To Pragmatism; User-Driven Development Of A Multilingual Parallel Concordancer" in S. Botley, A. McEnery, & A. Wilson (eds).
Wu, D. 1994. 'Aligning a parallel English-Chinese corpus statistically with lexical criteria', in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp 80-87.