MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society
Issue No. 2
October 1995

The Machine Translation Review incorporates the Newsletter of the Natural Language Translation Specialist Group of the British Computer Society and appears twice yearly.

The Review welcomes contributions, articles, book reviews, advertisements, and all items of information relating to the processing and translation of natural language. Contributions and correspondence should be addressed to:

Derek Lewis,
The Editor,
Machine Translation Review,
Department of German,
Queen's Building,
University of Exeter,
Exeter,
EX4 4QH
UK

Tel: +44 (0) 1392 264330
Fax: +44 (0) 1392 264377
E–mail: D.R.Lewis@exeter.ac.uk

ISSN 1358–8346

Contents

Editorial

28 September 1995

Multilingual Natural Language Processing (MNLP) Project

In Newsletter 21 of April 1993 Douglas Clarke suggested that there was a need for some commonly required NLP programs which should be freely available to avoid programmers continually re-inventing the proverbial wheel, and that a sub-group should be set up to see what we could do about it.

Some meetings of people interested in supporting this proposal were held and it was decided that a comprehensive list of MNLP functions should be established first before trying to write NLP applications such as morphological analysis, multilingual word-processing, CALL programs, etc.

More detailed proposals were prepared and circulated to those members who had expressed an interest in supporting a software project in one of our questionnaires. However, no significant interest has been shown since in either the proposal or the project.

Finally, in the last Review (April 1995) I summarized the position and made it clear that we would consider proposals from sub-group members, or indeed, from anyone else.

No proposals or enquiries have been received since then. It is true that this period has been over the summer but I think that unless we hear of some positive interest in the next few months we will have regretfully to admit defeat and draw this project to a close.

Therefore, if there is anyone still interested in supporting this project please let us know as soon as possible.

David Wigg

Group News and Information

Letter from the Chairman

The current publication is quite an achievement for our Group. This is the first time we have ever been able to publish twice in one year, which is very encouraging progress.

I believe the Review is important because it reaches about 10 times as many people as the average number of attendees at a London meeting (25).

That doesn't mean that the talks are not important; they are, but for a different reason. The talks bring members together in a more focused way in which some of us can get to know each other and our speakers better.

For those who are able to attend our meetings there is the added bonus of also being able to meet our speakers for a meal afterwards in a local restaurant where discussion can continue in a more relaxed atmosphere. Needless to say, the speaker's meal is paid for by the BCS (via the Group) but we have to pay for our own meals.

The Committee is very keen to expand the base of contributing members, particularly for the Review, but unfortunately we cannot afford to pay for more than one or two committee members to attend meetings from outside London, so we wonder whether we can interest any members in becoming Correspondent Committee members with whom we can build up a closer relationship for our mutual benefit. The expanding use of e-mail, frequently available in academic environments but now more widely used in business and privately, should make this form of co-operation easier.

If you have a particular interest in some aspect of machine translation or even just in multilingual software please share it with us so we can both benefit from the synergy of the group. Please contact me or any other Committee member if you would be interested in joining the Committee on this basis.

Roger Harris, whose official title is Rapporteur (a position requested by the Society to keep BCS HQ informed of what the Group is doing), is assembling a considerable amount of interesting and useful information on the Group's pages on the Society's World Wide Web pages at http://www.bcs.org.uk, so if you have Internet access have a look at this sometime.

All opinions expressed in this Review are those of the respective writers and are not necessarily shared by the BCS or the Group.

PLEASE NOTE

As membership is currently still free we must review our membership list from time to time, and the time has come round again. To renew your membership and to continue to receive the MT Review and notification of meetings please complete the form at the end of this Review (or a copy) and return it to me at the address shown.

J.D. Wigg

The Committee

The telephone numbers and e-mail addresses of the Officers of the Group are as follows:

David Wigg (Chair):	+44 (0) 1732 455446 (Home)
	+44 (0) 171 815 7472 (Work)
	wiggjd@vax.sbu.ac.uk
Monique L'Huillier (Secretary):	+44 (0) 1276 20488 (H)
	+44 (0) 1784 443243 (W)
	m.lhuillier@vms.rhbnc.ac.uk
Ian Thomas (Treasurer):	+44 (0) 181 464 3955 (H)
	+44 (0) 171 382 6683 (W)
Derek Lewis (Editor):	+44 (0) 1404 814186 (H)
	+44 (0) 1392 264330 (W)
	d.r.lewis@exeter.ac.uk
Tania Reynolds (Assistant Editor):	+44 (0) 1444 416012 (H)
Catharine Scott (Assistant Editor):	+44 (0) 181 889 5155 (H)
	+44 (0) 171 607 2789 X 4008 (W)
	c.scott@unl.ac.uk
Roger Harris (Rapporteur)	+44 (0) 181 800 2903 (H)
	rwsh@dircon.co.uk

BCS Library

Books kindly donated by members are passed to the BCS library at the IEE, Savoy Place, London, WC2R 0BL, UK (tel: +44 (0) 171 240 1871; fax: +44 (0) 171 497 3557). Members of the BCS may borrow books from this library either in person or by post. All they have to provide is their membership number. The library is open Monday to Friday, 9.00 to 5.00 pm.

MATCHING WORDS IN A BILINGUAL CORPUS

by

Roger Garside

Department of Computing and UCREL

University of Lancaster

The following account is based on a talk given to the Annual General Meeting of the Group on 6 July 1995. Readers are also referred to the paper ‘The Use of Approximate String Matching Techniques in the Alignment of Sentences in Parallel Corpora’ by A. M. McEnery, M. P. Oakes, and R. G. Garside (Clarke and Vella 1995). The work I am going to talk about is part of a project called CRATER (Corpus Resources and Terminology Extraction), which is being conducted at the University of Lancaster. A further project that is of interest here is UCREL, a joint research group involving the Department of Linguistics and the Computing Department of the University of Lancaster. When it was first set up nearly ten years ago, UCREL stood for Unit for Computer Research into the English Language. The title reflected the fact that the project would be concerned mainly with English. Since then the focus of the research has changed to include the matching of words in a bilingual English-French corpus. However, the acronym UCREL has been retained because it is familiar to most people (I believe it now stands for University Central Computing Corpus Research into Modern Language).

Corpus-based, probabilistic NLP versus rule-based NLP

The NLP research that we do at Lancaster is corpus-based and probabilistic. This approach was fairly rare up until a few years ago, although it was the first to be proposed when NLP began in the 1940s and 1950s. The idea was that, given a large number of possible linguistic items and a more limited set of probabilities that certain items would actually occur, it would be possible somehow to use the probabilities of the context in order to arrive at an analysis of the linguistic structure. At the time the approach did not work because the computing and linguistic resources were simply not available: machinery and resources of any kind were very limited, but this was especially true for dictionaries and corpora that could be machine-processed. As a result, most researchers concentrated on rule-based natural language processing. The standard paradigm for NLP came to be that a group of linguists generated some sort of rule system that would be used to decide what the structure of the language was. Not until the early 1980s, with the emergence of much more powerful computing and linguistic resources, was this paradigm seriously challenged by the probabilistic model.

Underlying probabilistic NLP is the notion of a mathematical model that can be trained and to which probabilities can be assigned for conducting linguistic analysis. The standard — and very popular and successful — example of such a model is word tagging. Tagging involves assigning parts of speech to words in a given sentence. It is performed by training a Hidden Markov Model. This asserts that it is possible to generate sequences of parts of speech according to a given probability pattern. Thus, given a word in context which has a particular part of speech, the category of the following word is more or less probable. For example, the occurrence of either a noun or an adjective after an article is probable, but it is unlikely that the article will be followed by a verb. The model can also be used to predict particular words from given categories. Thus, given a part of speech, we can say the possible words within this category are such-and-such. The model is referred to as 'hidden' because, when analysing input words in sequence, it guesses on the basis of given probabilities what the parts of speech of those words might be: in this sense the categories are 'hidden'.

The advantages of probabilistic processing over rule-based analysis are three-fold:

1. the search space is constrained;

2. the process is robust in the face of language variability;

3. the model is trainable.

First, as far as search space is concerned, one of the main problems of performing any kind of natural language parse is the vast number of possible structures that is produced as a result of the inherent ambiguity in language. Even a simple sentence can generate thousands or millions of possible parses, and it is not clear which is the right one to choose. One of the things that the probabilistic mechanism does is to constrain the search space. In looking for the correct parse it will hopefully provide a home-in sector for isolating a small number of possibilities, in some cases narrowing down the choice to a single parse.

Second, a probabilistic system is robust in the sense that it will always return an answer. It may not give the right answer — of course, it often does not — but it will always give some sort of answer. It is important to chose the right model of probabilistic analysis and also to select the right corpus on which to train the model: models can be general in relation to the text or they can be specific to the genre or to the domain. The appropriate choice in each case will hopefully return more reliable probabilities and therefore a better result.

Third, an advantage of the probabilistic model is that it is trainable. In order words, there are mechanisms for starting the model and for setting probabilities either automatically or semi-automatically. One way is to start out by manually marking up a section of text, using this to generate the initial set of probabilities. There are also unsupervised ways of training the model, including a mechanism for working out the best set of probabilities; these probabilities are then rated repeatedly until a set is eventually arrived at that is actually used.

'Statistical' machine translation

There are various ways in which probabilistic or statistical NLP can be relevant to machine translation. One is corpus alignment. Another is the development of parallel lexica, especially for specialized subject areas.

Corpus alignment involves establishing pointers to matching text units, such as sentences, in translated parallel corpora. If, in a bilingual corpus, the alignment of individual sentences is known, then it is possible to determine automatically the correspondences between the individual words. The method operates as follows. Given an English-French bilingual corpus in which it is known only which sentences correspond to each other, it is presumed that a particular word in an English sentence must correspond to some word in the parallel French sentence. If the corpus contains enough sentences in which the same English word occurs, then it is reasonable to assume that the French translation equivalent will also occur in the parallel French sentences. The method by which the presence of such words may be calculated is known as *mutual information* and is based on the expectation that a particular English word and its French equivalent will occur more often than would be probable by chance alone. In most cases one would expect to see the same French translation for a given English word in the parallel French sentences. But clearly this is not always so in practice. It is more or less true for English and French but applies less to translations of English into highly inflective languages. In such cases the same word in English is unlikely to correspond to a single word in the other language.

Despite these drawbacks the relatively simple approach outlined above achieves remarkably good results. Indeed, researchers have often started out with an astonishingly simple statistical approach which has been subsequently refined by applying some linguistic knowledge.

To illustrate the approach, consider how purely statistical techniques can be used to extract multiword units. The technique is fundamentally the same for both monolingual and multilingual corpora. For, say, a monolingual English corpus in a subject domain, the first stage is to ask whether some of the words in the corpus occur next to each other more often than might be expected by sheer chance. For a multilingual corpus — English-French, for instance — the same question is asked for individual word pairs in each language. The next stage is to ask whether there is a pair of words in the English sentence that tends to correspond to a pair of words in the parallel French sentence. If two such pairs are found, then it is plausible that the word pairs are translations of each other. For the method to work the corpus for the domain must be very large, but the basic technique is the same. There are various statistical tests for establishing (a) whether a given pair of words occurs more frequently in English than in French, (b) whether a corresponding pair occurs more frequently in French than in English, and (c) whether both pairs occur more often than might be expected in the aligned sentences.

In the first phase of the CRATER project extensive use is made of statistical techniques. We determine the number of times two words occur together in a pair of sentences, the number of times one word occurs without the other, the number of times the second word occurs without the first, and the number of sentences in which neither word occurs at all. Given these four numbers various statistical tests are carried out to try to establish whether it is possible to extract the more interesting pairs of words.

The next stage is to refine the process by applying linguistic knowledge. By using a linguistic filter it is possible to concentrate the search on the most interesting word pairs. For example in French, useful pairs of words may have a particular sequence of categories: ‘noun d’, ‘noun d noun’, or ‘noun d determiner noun’, or ‘noun adjective’. The corresponding pattern in English might be ‘noun noun’ or ‘adjective noun’. In this way a common translation pattern can be established which states that the ‘noun adjective’ sequence in French corresponds to an ‘adjective noun’ sequence in English. The filter is applied on top of the purely statistical techniques. In effect the filter says: do not look at all possible sequences of words, only at those sequences which match a particular pattern. The same technique can be applied to three-word sequences. In the CRATER project we have found that, of the 500 two-word phrases ascertained by the statistical test as translations, about 80% are valid, that is, four out of five. Clearly the technique is not 100% successful and some manual work is required at the end.

Work on statistical machine translation proper began at IBM (Yorktown Heights), although the original team has since been dispersed. The project, called CANDIDE, was based on a corpus of English and French parliamentary texts contained in the Canadian Hansard. The object was initially to build a dictionary of translation equivalents. The first step involved taking a French sentence and guessing at some of the corresponding English words in the translation. For this to work it is important to maximize the probability of given French words corresponding to certain English words. There are two probability distributions involved here. The first is what sequences of English words are likely to occur (some sequences are more probable than others). The second is, given a French word, what are the likely English words generated from it? This is a well established and reliable mechanism and, although it is possible to improve considerably on the technique, the basic method remains the same. The approach can also be used to decide on the correct sense of a word in context. Thus by extracting statistically what object the French verb ‘prendre’ takes, it is possible to determine whether it should be translated as ‘take’ or ‘make’.

Corpus alignment

As mentioned earlier, the object of corpus alignment is to identify the corresponding units of language in a translated parallel corpus. The unit may be a paragraph, a sentence, or a single word. Alignment at word level effectively means developing a bilingual lexicon. However, most current work is carried out at sentence level. Generating a sentence alignment for a bilingual corpus is a useful end in itself. Monolingual corpora have been used as aids in teaching English at Lancaster, while parallel multilingual corpora have possible applications in second language learning.

There are two reasons why corpus alignment is a non-trivial activity. First, there is in practice never a one-to-one correspondence between sentence units in bilingual corpora. If each sentence in one language corresponded to a single sentence in another language, then alignment would present no problem. The difficulty is that some sentences in one language will correspond to more than one sentence in the other. The correspondence will vary from genre to genre: it is less predictable in narrative fiction, for example, than in technical language. The absence of one-to-one correspondences means that we require a matching algorithm that aims to align the disparate text units.

The second reason why alignment is non-trivial is that entire sections of text will be missing. If, for instance, an attempt is made to align a large corpus on the basis of paragraph markers alone, it is likely that some paragraph markers or even parts of the text will be left out. There are such missing sections in the alignment performed by IBM on the Canadian Hansard. On a really large corpus containing tens of millions of words the process of alignment is far from straightforward. Although it is possible to go through a small corpus by hand and to identify the corresponding paragraphs, this is not feasible with large corpora.

The Gale and Church Algorithm

There are two well known algorithms for performing an alignment. One, called the Gale and Church Algorithm, is publicly available and is used at Lancaster. The second is the Kay and Röscheisen Algorithm. A third, less well known algorithm is used solely by IBM. The Gale and Church and Kay and Röscheisen Algorithms are described in the March 1993 issue of *Computational Linguistics* (Gale and Church 1993; Kay and Röscheisen 1993).

The Gale and Church Algorithm employs dynamic programming as an efficient search technique for establishing the optimal alignment. The basic method is as follows. The starting-point is a long sequence of words, that is, a text in language A (say, English), which has to be matched up with a text in language B (say, French). There is always a number of different possible alignments: for example, the first sentences in each corpus might match one another; the first of the following two sentences in English might match the second sentence in French; the fourth French sentence might match the fifth English sentence, and so on. The goal is always to find the best possible alignment. ‘Best possible’ here means best in a probabilistic sense.

There are two main elements of probability in the Gale and Church Algorithm. The first element has to do with sentence length and is based on the strong likelihood that a long sentence in English will correspond to a long sentence in French; similarly a short sentence in one language will correspond to a short sentence in the other. Roughly speaking, given that the average lengths of sentences in French and English are known, it is possible to set up a distribution of possibilities from this information.

The second element of probability is based on the lack of a one-to-one correspondence between sentences. A sentence in language A may correspond to zero, one, two, or even seven sentences in language B. In the Gale and Church Algorithm a given sentence in one language corresponds to one, zero or two sentences in the other language. Thus one can have two sentences in language A matching two sentences in language B, but one cannot have three sentences in language A matching two in language B. The only permitted matches are 1–1, 2–2, 1–0, 0–1, 1–2, or 2–1.

From their own data, Gale and Church assessed their algorithm as providing valid 1–1 correspondences in 89% of cases. Thus in nine out of ten cases the sentences corresponded 1–1. Gale and Church achieved these results with economics texts. On the Canadian Hansard they achieved 91%. The figure for 1–0 correspondences was 1%, for 1–2 correspondences around 9% (in either direction), and for 2–2 correspondences 31%. The probabilities in Gale and Church's original paper were obtained from hand-aligned data. On the basis of figures thus derived the authors counted what the probabilities should be for the distribution of sentences.

Evaluation of the Gale and Church Algorithm

Gale and Church give the following figures for the performance of their algorithm: 94.2% success for English to French, and 97.3% for English to German. After performing the alignment they gave it to speakers of French and German to check the correspondences and to comment on whether or not they believed them to be correct. The checkers were themselves tested with economics texts to ensure that their assessments could be relied upon.

The algorithm performs best on 1–1 alignments. Separated out from other alignments, error rates on 1–1 matches are 5%, 5–6%, and 3–4%. These results have led to claims that the algorithm works only on the 1–1 alignments and that it is practically useless on the others. However, the assessment depends on what the aims of the user are. If the object is to produce a fully aligned corpus, then clearly it will be necessary to go through and to correct manually the alignment that has been produced automatically. This was done in the CRATER project. On the other hand the user may not wish to have the entire corpus aligned. If he is using the corpus to develop a word lexicon, for instance, he may need only a limited number of parallel sentences, in which case he will throw away all the non1–1 alignments because the error rate for these is too high; for his purposes the corpus consists only of those sentences that are correctly aligned. At Lancaster we obtained 98% alignment for 1–1 correspondences in English–French texts and 93.2% for English-Spanish; the domain of the corpus was telecommunications and the translation a fairly literal one. Tests on various English-Polish parallel texts returned results in the range 65–85%.

The IBM Alignment Algorithm

The IBM Alignment Algorithm, which to my knowledge has not been made public, returns very similar results to the Gale and Church Algorithm in terms of its success in finding 1–1 alignments.

The first thing to note is that the IBM algorithm uses a different definition of length from that of Gale and Church. Clearly, long sentences in one language will be long in another language by whatever definition of length is applied. But whereas Gale and Church measure sentence length by characters, the IBM alignment method is based on tokens (words). There are arguments for both approaches, but it is safe to say that, in general, it is better to base sentence length on tokens rather than on characters.

The IBM researchers also adopted a different methodological approach. Whereas Gale and Church took some linguistic data, aligned it by hand, and then used the figures thus obtained in order to generate the values of the distributions, the IBM workers devised a method called *unsupervised training*. This assumes a mechanism for guessing initial probability estimates which are used to determine the likelihood of an alignment. These values are then used to generate some new (and more accurate) estimates. The process is repeated over and over again. Once the initial guestimates of the probabilities are decided on, an alignment is generated — the most likely one given the probabilities used. From this, new probability estimates are derived, and, in turn, a new and probably more accurate alignment is performed.

The method is not without its drawbacks: it requires large volumes of data and is very time-consuming. It is a hill-climbing technique, guaranteed to give a local minimum. Thus, if one thinks of the possibilities as a hilly surface, the method will ensure that one climbs the local peak. While it will not necessarily ascend Everest, it will conquer at least some of the lesser summits. IBM concentrated on French-English because of the availability of the large (French-English) Canadian Hansard corpus. IBM claimed 99% success in correct alignment, although only the 1–1 correspondences were counted. It has also been suggested that the Canadian Hansard is a particularly literal translation, which would help account for the remarkably high alignment figure; the algorithm might work less well with a freer translation. It should be noted that IBM did not carry out a 2–2 alignment, whereas Gale and Church performed 1–1, 1–0, 1–2 and 2–2 alignments. It should also be noted that IBM performed a two-part alignment. This involved first aligning the texts at paragraph level and then discarding about 10% of the material that did not match, either because the paragraphs did not align or because parses and paragraphs were missing. The intention was to generate a parallel corpus with different menus for extracting various probabilities, ultimately for use in machine translation.

The Kay and Röscheisen Algorithm

The Kay and Röscheisen Algorithm is a much more complicated affair than the Gale and Church and IBM algorithms. It may characterized as a sort of relaxation technique. The method presupposes a so-called *alignable sentence table* (AST). The AST is a section of sentences in the corpus which may possibly match: thus, for instance, the first sentence in language A may match the first, second, or even the third sentence of language B; it will not, however, match the 99th sentence of language B. Similarly, the last sentence in language A is unlikely to match the first sentence in language B. Thus one can say that any particular sentence in language A is likely to be matched with only a selection of sentences in language B. If this is the case, one can also make some guesses at which individual words might be aligned. The technique operates as follows: if a particular word occurs in language A and another word also seems to occur in at least one of the sentences in language B, then it is possible to say that they might be translations of each other. In this way one generates from the initial AST a *word alignment table* (WAT) of words similarly distributed in the alignable sentence and therefore presumed to be translation equivalents. By assuming that certain words have been aligned, we assume that particular sentences must also be aligned: if a word is a translation of another word, then the sentences that contain those words are also likely to correspond. The result is a sentence alignment table (SAT), that is, a list of sentences containing words that have been aligned. Thus alignments of sentences are generated from alignments of words, although it must be noted that this applies only to sentences that are fairly certain to correspond. The process is then repeated, leading to adjusted versions of the initial alignable sentence table and the associated SAT. Once it is decided that some sentences are definitely aligned, then constraints begin to operate within the text. Since there can be no crossing over of alignments, it can be assumed that certain other sentences must match. If sentence 27 in language A matches sentence 29 in language B, then we can be confident that some of the sentences preceding the former must match some of the sentences preceding the latter; by the same token, one or two of the sentences that follow must match each other.

The Kay and Röscheisen Algorithm clearly works very well, but there are some problems with it. Since the authors have not provided a paper for it, we at Lancaster have written our own implementation. This has actually proved to be very difficult, and the program that we have produced operates very slowly. As a result we have not done very much with it. Kay and Röscheisen claim to achieve very good results with four iterations. Indeed, with very small files it is possible to get very good alignments. The advantage of the approach is, of course, that it does not just align whole sentences: it also aligns partial sentences and words. The other algorithms guarantee to produced an alignment of all sentences, even if the alignment contains 1–0 correspondences. The Kay and Röscheisen program does not necessarily do this, but it does produce a word alignment table.

Anchor points

The above techniques all make use of what may be called *anchor points*. These are positions in one text which, with a degree of certainty, seem to match up with positions in a parallel text. Obvious anchor points are paragraph markers. Given a very long text, there is going to be a large number of such markers, so it is unrealistic to assume a close alignment from these alone. In view of the increase in popularity of electronic publishing a proliferation of markers of all kinds is inevitable, regardless of the language involved. Thus if a text in language A contains a mark-up for a section in italic script, then the corresponding section in language B is likely to have the same mark-up. It is interesting to note that the IBM team performed a first-pass alignment at the paragraph level and then a second-pass alignment at the sentence level.

The IBM team also used elements of the text as anchor points. They were able to assume that the phrase ‘Mr Speaker’ in the English text of a parliamentary debate would correspond to the French ‘M. le Président’ and that the phrase ‘some time later’ would match ‘temps plus tard’, and so on. Such elements are highly specific to the corpus. Other elements might include numbers, proper names, and dates. It may even be possible to use certain features of punctuation as anchor points: an example is question marks, since a question in one language is likely to correspond to a question in the other language.

Cognates

Work on the CRATER project at Lancaster has concentrated on identifying word cognates as anchor points. The first question here is whether it is at all possible to find cognates by automatic methods. For instance, are there orthographic marks in verbs which represent tactic correspondences and which can be used to determine automatically whether, say, a ‘telephono’ in an Italian corpus matches the word ‘telephone’ in an English corpus?

CRATER stands for Corpus Resources and Terminology Extraction MLAP 93/20; partners in the project are UCREL (Lancaster), C2V and IBM (Paris), and the Universidad Autónoma (Madrid). The corpus used is the ITU (Intenational Telecommunications Union) corpus. The English and French parts of the corpus were processed separately from the Spanish section. The corpus contains about one million words in each language and the subject domain is telecommunications; since the corpus is the text of an official CCITT book, it is available in various languages. The English text was tagged using a system developed at Lancaster which assigns 150 parts of speech. The French text was processed using an IBM tagger developed in Paris; this assigns about 100 tags. The Spanish text has been tagged using a publicly available Hidden Markov Model tagger; its complexity is such that is employs about 500 tags. The English and French texts have already been manually post-edited; the Spanish text is in the process of being post-edited.

The first thing to note is that lexical cognates can be identified non-automatically. Working from English to Italian, for example, it is clear that in some cases English ‘ph’ appears as ‘f’ in Italian. This means that if English word contains the sequence ‘ph’, we may immediately consider the possibility that the sequence can be replaced by ‘f’ in Italian. Examples of French to Spanish include ç (c + cedilla) to ‘z’, ‘c’ or ‘cc’. In certain phonemic contexts double consonants in French go to single consonants in Spanish: thus French ‘recommander’ becomes Spanish ‘recomendar’. Finally, an initial ‘s’ followed by a consonant in French might correspond to initial ‘es’ plus the consonant in Spanish: thus ‘scène’ could be matched to ‘escena’.

The Dice Similarity Coefficient

Work at Lancaster has concentrated on *approximate string matching*, a widely used technique in information retrieval for natural language processing. One way of carrying out string matching is based on the Dice Similarity Coefficient. There are other methods, but the Dice method is the most popular. Essentially, the technique assumes a number of possibilities for the two items that are to be matched. Thus, there is a certain number of possibilities for the first element that is to be matched, and another number for the second element. The first stage is to count the total number of character bigrams (that is, sequences of two characters), both matched and unmatched, in a pair of words.

As an example consider the words ‘telephone’ in English and ‘telefono’ in Italian. The bigrams for English are: ‘te’, ‘el’, ‘le’, ‘ep’, ‘ph’, and so on. For Italian they are ‘te’, ‘el’, ‘le’, ‘ef’, etc. We count the total number of bigrams in both words, the total number of bigrams that match (such as ‘te’, ‘el’, etc), and apply the following formula

$$\frac{2 \times b_m}{b_i}$$

where b_m is the number of matching bigrams, and b_i is the total number of bigrams (both matching and non-matching) in the two words (that is, $b_1 + b_2 + ... + b_n$). This provides a measure of the average match, that is, the number of real matches over the number of possible matches. If the result of applying the formula is 1, then there is an exact match; if the result is 0, then there is no match at all. In our example the result is 0.47, which is a fairly average figure. For the two words in this particular case one might have expected a higher value. This would indeed be returned if we applied linguistic information and changed the Italian 'f' to 'ph': in this case the only bigram that would not match would be the final 'ne' and 'no'.

There are particular problems with three-letter acronyms. Our corpus contains a large number of three-character acronyms in one language which seem to match four-character words in the other language. The only way to eliminate this problem would be to start matching at four characters. Since the ITU corpus is lemmatized, it would be possible to match, not on characters, but on tokens and lemmas.

Here are some results of percentage accuracy rates for different dice coefficients:

Dice=	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Tokens	0	7	22	49	82	97	100
Lemmas	0	5	23	48	71	95	100

Truncation

An even simpler technique used in information retrieval is truncation. This is based on the assumption that a pair of words are cognates if the words match on the first *n* characters. The technique does not work well for low values of *n*. One way of overcoming this problem is to generate stoplists. These are lists of 'false friends' — that is, words that are known not to match beyond *n* characters — for each value of *n*. At Lancaster we took a thousand pairs of words and looked through them to see where the technique worked and where it did not: for example, we would identify the cases in which, where $n = 5$, we would not want words longer than *n* to be seen as cognates. The percentage rates for truncation without human intervention are as follows:

Length	3	4	5	6	7	8	9	>10
Raw	1	8	14	69	93	98	100	100
Stop	5	20	74	94	98	100	99	100

In this table the top row denotes the different truncation lengths, that is, the value of *n* chosen for the truncation; the bottom row is the accuracy rate achieved after the 'raw' results (given in the second row) are corrected against the stoplist of 'false friend' cognates. The table indicates that 98% of cognates in the corpus are found automatically and without human intervention by matching the first eight letters of each pair. It is important to note that these cognates are only possible alignments or translations of each other, not necessarily actual or real translations.

Dynamic Programming

A third method involves Dynamic Programming, which has been mentioned earlier in the context of the Gale and Church Algorithm. Dynamic Programming is based on the situation in which two elements partially match and in which it must be decided what has to be done in order to make the elements match completely. Essentially the same method is used in spell-checking. We begin by asking how much we have to change one element in order to match the other, and what the cost is of achieving the match. Thus, for instance, we might consider a word A to be a misspelling of word B. The objective is then to modify word A with minimal cost in order to make it match word B. There is a given repertoire of possible operations for modifying or rewriting a word, each of which is associated with a certain cost. Typical operations include insertion, deletion, and substitution. For example, to match the English word 'telephone' with the Italian 'telefono', we would have to substitute the 'p' for an 'f', delete the 'h', and substitute the 'e' for an 'o'. Assigning to each operation the cost value of '1' — that is, each operation is considered to be equally costly — we arrive at '3' as the total cost of making the match; this value is divided by the length of the longer word in order to arrive at a normalized score that can be used as a basis of comparison.

The percentage scores for the accuracy of the Dynamic Programming method are as follows:

Score	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Accuracy	1	2	12	39	73	95	100

The technique can be refined by adjusting the costs of the modification operations in the light of the language pairs involved. For instance, the cost of substituting English 'ph' by Italian 'f' should be much lower than, say, for changing the 'ph' to an 'x' or an 'n'.

Restricted regions

Up to now we have assumed that cognates are generated from the words in the entire text. A different approach is to restrict the region of search and to look for correspondences in particular areas of the texts to be aligned. Gale and Church distinguished between 'hard' regions (paragraphs) and 'soft' regions (sentences). Given an alignment at paragraph level or a preliminary alignment at sentence level, we might choose to limit the generation of cognates to just those regions.

Here are some dice coefficients for matches in restricted regions:

Dice	0.4	0.5	0.6	0.7	0.8	0.9	1.0
All	0	7	22	49	82	97	100
Hard	5	33	66	94	96	99	100
Soft	15	69	71	97	100	100	100

The ‘best match’ criterion

The ‘best match’ criterion was adopted by the IBM team working in Paris. It proceeds from the assumption that a word *x* in language A might match a number of different words in language B. Conversely, a word *y* in language B might match a number of different words in language A. We first identify the best of all the matches for *x* in language B and the best match for *y* in language A. If *x* and *y* turn out to be the best matches for each other, then these are the matches that are retained and the others are discarded. In this way the parallelisms are restricted to those that in an automatic way seem to afford the best possibilities.

The following table shows the percentage accuracy rate for both dice and the best match criterion.

Dice	0.4	0.5	0.6	0.7	0.8	0.9	1.0
All	0	7	22	49	82	97	100
Hard	5	33	66	94	96	99	100
H/BM	26	72	81	100	100	100	100
Soft	15	69	71	97	100	100	100
S/BM	41	93	90	100	100	100	100

The figures show that the accuracy rate improves for hard and soft regions. The best results of all are obtained with soft regions, although the problem here is that there may not be enough matches to apply the ‘best match’ criterion; in this case it is necessary to balance the accuracy achieved with the number of cognates that actually turn up.

The incorporation of cognates in the Gale and Church Algorithm

In conclusion, I shall discuss how cognates can be incorporated into the Gale and Church Algorithm. The first stage is to obtain a set of possible cognate pairs using one of the mechanisms, such as the Dice Coefficient mentioned above. A particular cognate pair will be associated with a probability estimate which is a measure of the likelihood of its being the correct pair. The Gale and Church Algorithm also states that there is a probability that the sentences are in some sort of alignment, that is of 1–1, 2–1, and so on. There is also a third measure of probability: we can say that there are likely to be cognates in one language which occur in a parallel sentence, while there will also be cognates in the second language, where the expected match is not present.

The above approach has been tried out at Lancaster on an English-Spanish corpus. Although the processing has proved to be slower than expected, we can say that the accuracy of alignment has improved considerably. The technique is still being refined and has yet to be tested fully.

References

Clarke, D. and Vella, A. (1995) *Machine Translation — Ten Years On*, proceedings of an international conference held at the University of Cranfield 12 – 14 November 1995: Cranfield University Press (forthcoming)

Gale, W. A. and Church, K. W. (1993) ‘A Program for Aligning Sentences in Bilingual Corpora’ in *Computational Linguistics*, No. 1 (March), Vol. 19: 75–102

Kay, M. and Röscheisen, M. (1993) ‘Text-Translation Alignmnent’ in *Computational Linguistics*, No. 1 (March), Vol. 19: 121–42

Lexical Resources for MT: a Survey

by

Adam Kilgarriff

Research Fellow

Information Technology Research Institute

University of Brighton

The following survey first explains why we should be interested in monolingual as well as bilingual dictionaries. It then discusses the potential providers and the sorts of resources they have, such as the following:

machine translation companies;

dictionary publishers;

CD-ROM dictionaries;

NLP research groups working on machine-readable dictionaries;

other NLP and psycholinguistic lexicographic work.

The survey concludes with an annotated list of sources.

Bilingual and monolingual, 'direct' and 'transfer'

At a first pass it might appear that it is bilingual dictionaries alone which will be of interest to machine translation (MT). The task is to get from one language to the other, so why would an account of a word given in the same language as the original be of any use? This line of argument is akin to the earliest, 'direct' approach to machine translation: you start from the premise that you simply look the words up in a bilingual dictionary and then swap source word for target word; you do something more sophisticated only when that fails. The shortcomings of this approach are well known (see Hutchins and Somers 1992:71–77 for a discussion), and most recent systems have been 'transfer' systems. Such systems do some grammatical analysis of the source language text in order to produce an intermediate representation (usually a tree structure); they then transfer this into an equivalent target language representation, and from this they generate the target language text. In such a scenario the relevance of monolingual dictionaries becomes evident. If they provide the best grammatical descriptions of words, then they will be the best resources to use for the analysis and generation stages of the process, with the bilingual dictionary being used just for the transfer. Thus MT systems need bilingual lexicons for stating mappings between lexical items in the source and target languages (or, for interlingua systems, between the interlingua and each language). But they also need a source for language-specific facts about words, possibly a monolingual dictionary.

Machine translation companies

Researchers and system-builders in MT have always been aware that lexicons need to be big. Indeed commercial MT companies have been building up their lexicons for as long as they have been in the business. The evidence suggests that it takes at least ten years for an MT product to reach the market, and much of those ten years are spent on lexicography. Lexicons are MT companies' most valuable resources, representing very large quantities of detailed information about words, formally and consistently. Some of the firms are currently collaborating with university research groups.

Informal approaches to some of the leading companies suggest that they are indeed interested in academics experimenting with their lexicons, though this would be on the basis of a licence which confirmed that their resources (or software derived from them) would not be redistributed.

Machine-readable dictionaries (MRDs)

Until relatively recently most work in NLP took place on a 'toy' scale. Such work was done in the laboratory, exploring properties of some interesting aspect of grammatical behaviour, and there was no reason to have more than a few dozen words in the lexicon. By the mid-1980s much of the theoretical basis for NLP — chart parsing, feature-based formalisms, unification — was in place and the question, 'how do we apply it?,' came to the fore. A prerequisite for many possible applications is 'scaling up'. This is particularly important in relation to the lexicon, where scaling up means moving from dozens of words to tens of thousands of words. For the last ten years scaling up has been a very active research area.

From within NLP the first major push was to exploit machine-readable copies of dictionaries. This involved first getting hold of the data from dictionary publishers and then decoding it in order to turn a typesetters' tape into some form of database where the different fields of information — headword, part-of-speech, definition, etc. — were retrievable. How to do this and what can be achieved are explored and reviewed in Boguraev and Briscoe (1989) and in Byrd et al. (1987). Much of the leading-edge research was performed on monolingual English dictionaries, though some groups, notably the IBM group at the T. J. Watson Research Center, have been busy gathering and decoding dictionaries (monolingual and bilingual) for a number of languages.

One goal of machine-readable dictionary research was simple: to produce 'no-semantics' lexicons containing spellings, morphology, and parts of speech — possibly also spelling variants, and a code for the domain a word was likely to occur in. In these aims the enterprise has been fairly successful, with the ANLT lexicon (see the list of resources given below) as one useful output.

A more ambitious goal was the construction of a lexical knowledge base containing formalized information about the meanings of words. The outcome here was less successful. The EU-sponsored project ACQUILEX explored the issues and problems involved. Researchers started to see diminishing returns from effort spent on dictionary processing as they moved from orthography, through syntax, to semantics. For the follow-on project, ACQUILEX-2, there was a greater emphasis on working with lexicographers and corpora to produce dictionaries with the right information in a well-structured form in the first place. At an ACQUILEX-2 review workshop last year, one key paper was subtitled 'Have we wasted our time?' (also published as Ide and Veronis 1995*)*. It noted that errors, inconsistencies, and circularities continue to be the bane of the MRD community and undermine hopes of producing a usable, wide-coverage LKB via the largely automatic processing of MRDs.

In the last few years dictionary publishers have realized the benefits of writing and storing dictionaries as databases. As a result we now find that most large dictionary publishers have started to use computerized dictionary-writing environments. Such environments are built on the premise that dictionary entries are highly structured entities and that the dictionary entry is a database entry from the moment it flows from the lexicographer's metaphorical pen.

Licences

Dictionaries are very rich in information. This richness corresponds to a large number of person-hours expended on the making of dictionaries. Since the cost of writing a state-of-the-art dictionary runs into millions of pounds, the product is not given away lightly.

Various kinds of licence arrangements are possible, but one that allows the licensee to pass the dictionary itself (or software derived from it) to anyone else is sure to be very expensive. Since academics are typically poor and typically acquire fame and glory through their work being adopted and developed by others, this is a fairly severe constraint.

The high cost of licences is one reason why existing MT products do not use MRDs. Another reason is the long time lag between the fixing of the architecture for an MT system and its arrival on the marketplace. The design of the MT systems currently on the market pre-dates most MRD research. This may be set to change. At least one MT group is now extensively using MRDs: Sharp Laboratories Europe are partners of Cambridge University Press (CUP) in a project called 'Integrated Language Database' and are using CUP's new dictionary CIDE in their development work.

The licence fee for an MRD for a new dictionary is variable, but often seems to be in the region of £1,000. This figure would cover a licence only for research use in a university and typically comes with an obligation to keep the publisher informed of what is being done with it. Where resources developed in academia are nonetheless based on published dictionaries, a licence with the publisher, under similar terms, is usually still required.

There are, of course, many free resources, though in most cases these will not be as accurate or as complete as a commercial dictionary.

CD-ROM dictionaries

In the last few years the market for dictionaries on CD-ROM has opened up. As a result there are now machine-readable dictionaries available for many languages and language-pairs which can be bought for a modest sum. There are, of course, legal constraints on what you can do with the information you buy when you purchase such a CD-ROM, as well as the practical difficulty of extracting the information from the medium in a usable form. If you embedded the lexical information in an MT system which was then redistributed without the publisher's consent, this would infringe copyright; but to build such a system for yourself does not. The team at ISI in California has done just that for the CD-ROM MRD format used by the Electronic Book Catalogue; one of the team's researchers, Matthew Haines, has devoted many person-hours to decoding between fifteen and twenty dictionaries. Thus one way to get hold of lexical databases for languages and language-pairs available on CD-ROM is to buy the CD-ROM itself and set about decoding it. Haines declares: 'each dictionary is a new project. It takes a lot of time to reduce an electronic book to a database, but with individuals serious about investing that time, I will be happy to share my programs and offer advice.'

Research lexicography

It is worthwhile noting that the research community itself has developed some substantial resources. Foremost amongst these is Wordnet (Miller 1990). This is a dictionary-like resource produced by lexicographers and students at Princeton University, whose motivation was to pursue various hypotheses from psycholinguistics about the mental lexicon. The outcome is an online dictionary, available entirely free for research purposes over the Internet. Wordnet is now being very widely used in the NLP and information retrieval communities. An EU project which is about to begin aims to extend Wordnet to include words from a number of European languages. Another important resource for English that is being developed specifically for reuse in NLP research and commercial applications is COMLEX (Grishman et al. 1994); this is being produced under contract to the Linguistic Data Consortium.

The Message Understanding Conference (MUC) initiative in the United States has nurtured the development of application-specific lexicons on short timescales, making imaginative use of corpora of the appropriate genre for automatic and semi-automatic 'lexicography'. A wide range of approaches and in some cases public resources is described in the MUC-5 proceedings and in Boguraev and Pustejovsky (1993). Recent work includes the automatic 'learning' of translation pairs from corpora in two different languages (Fung 1995, Wu 1995).

Annotated resource list

The following contact names and e-mail addresses have been gleaned from various sources. Where I have not yet been able to confirm the e-mail address by eliciting a reply, I have marked the address as unconfirmed.

Commercial MT companies

LOGOS
Friederike Bruckert,
LOGOS Computer Integrated Translation GmbH,
Mergenthallerallee 79–81,
D-65760 Eschborn/Ts., Germany,
Tel: +49 (0)619 659030
Fax: +49 (0)619 6590315
E–mail: bruckert@logos.de (unconfirmed)

INTERGRAPH
The organisation has recently contracted to sell its service over the Compuserve network.
Susan Moore
Tel: +1 205 730 3315
E–mail: sjmoore@com.ingr
Also web page http://www.intergraph.com

SYSTRAN
Boasts the largest repertoire of language pairs and the largest dictionaries.
Tel: +1 619 459 6700
Fax: +1 619 459 8487
E–mail: info@systranmt.com (unconfirmed)

METAL
Geert Adriaens,
Siemens-Nixdorf,
Centre Software de Liège,
Rue des Fories 2,
4020 Liège,
Belgium;
E–mail: gad@csl.sni.be (unconfirmed)

Dictionary Publishers

LONGMAN
The Longman Dictionary of Contemporary English (LODCE) is the single most widely used dictionary in NLP research. Most work has been based on the first edition of 1978. A corpus-based third edition is now available in SGML. Other monolingual English learners' dictionaries and thesauri are also available.

Steve Crowdy,
Longman Dictionaries,
Longman House,
Burnt Mill,
Harlow,
Essex CM20 2JE
Tel: +44 (0)1279 623816
E–mail: 100425.3057@com.compuserve

CAMBRIDGE LANGUAGE SERVICES is the commercial wing of CUP. They have recently produced the Cambridge International Dictionary of English. The Cambridge University NLP group and the Sharps Laboratories Europe MT group were involved in some aspects of its production, and CLS have widely advertized that they intend to make the database readily available for research use.

Paul Proctor,
Sue Allen-Mills
E–mail: sallmill@uk.ac.cam.cup

HARPER COLLINS
The 1978 Collins English Dictionary is readily available on the $25 ACL-DCI CD-ROM. A version of the Collins English-Spanish dictionary for MT has been produced at Carnegie-Mellon University and may be made available to other researchers.

Contact:
Bob Frederking
E–mail: ref@cs.cmu.edu

For other Collins monolingual and bilingual dictionaries contact:
Lorna Sinclair-Knight
E–mail: lornas@reference.collins.co.uk

For COBUILD dictionaries contact:
Gwyneth Fox
E–mail: gwyneth@cobuild.collins.co.uk

OXFORD UNIVERSITY PRESS
OUP publish various monolingual (learner, concise, shorter, full OED) and bilingual dictionaries in machine-readable form.

Simon Murison-Bowie,
Electronic Publishing,
Walton Street,
Oxford OX2 6DP

CD-ROMs
The Electronic Book Catalogue lists dictionaries (monolingual and/or bilingual) for English, French, Spanish, German, Japanese, Danish and Dutch.
Tel: +44 (0)171 561 9590
Fax: +44 (0)171 561 9591

Academia and the NLP Research Community

Distributors of Lexical Resources:

LINGUISTIC DATA CONSORTIUM (LDC)
This is a membership organisation which exists explicitly for the purpose of developing and redistributing linguistic resources for NLP. The LDC distributes the ACL-DCI CD-ROM, CELEX, COMLEX, and others. Anyone wishing to gain access to the LDC resources either pay for one specific resource or pay for a year's membership which then entitles them to take copies of all resources published in that year. A year's membership fee for a university is $2,000.

E–mail: ldc@unagi.cis.upenn.edu
WWW: ftp://ftp.cis.upenn.edu/ldc_
www/ldc_catalogue

The CONSORTIUM FOR LEXICAL RESOURCES is a library of lexical resources with some overlap with the LDC, although the emphasis is on distributing rather than developing resources. Most resources at LDC are available free.
E–mail: lexical@nmsu.edu

WWW: ftp://crl.nmsu.edu/CLR/catalog

Other resources:
WORDNET
This is available free by ftp from:
clarity.princeton.edu/pub/

ANLT LEXICON
This was produced from LDOCE (1978) as part of the UK-funded Alvey Natural Language Tools Project. It contains fairly detailed syntactic information, with subcategorisation codes extracted from LDOCE and then extensively checked and corrected; there is no semantic information. ANLT is available under licence through Lynxvale (the trading arm of Cambridge University) for 500 ECU.

http://www.cl.cam.ac.uk/Research/NL/anlt.html

CELEX
This contains detailed morphology and some syntactic information for English, German and Dutch. It is available on CD-ROM through the LDC for $150.

Bibliography

Boguraev, B. K. and Pustejovsky, J. (1993) *Acquisition of Lexical Knowledge From Text: Workshop Proceedings*, Ohio: ACL Special Interest Group on the Lexicon.

Boguraev, B. K. and Briscoe, E. J. (1989) *Computational Lexicography for Natural Language Processing*, Longman: Harlow

Brown, P., Cocke, J., Della Pietra, S., Jelinek, F., Lafferty, J. D., Mercer, R. I., and Roossin, P. S. (1990) 'A Statistical Approach to Machine Translation' in *Computational Linguistics*, No. 6, Vol. 2: 79–86.

Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S., and Rizk, O. A. (1987) 'Tools and Methods for Computational Lexicology' in *Computational Linguistics*, Vol. 13: 219–40.

Fung, P. (1995) 'A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora' in *Proceedings, 33rd Annual Meeting of the Association of Computational Linguistics*, MIT: 236–43.

Grishman, R., MacLeod, C., and Meyers, A. (1994) 'Comlex Syntax: Building a Computational Lexicon' in *COLING 94*, Tokyo

Hutchins, J. and Somers, H. (1992) *Introduction to Machine Translation*, London: Academic Press.

Ide, N. M. and Veronis, J. (1995) 'Knowledge Extraction from Machine-readable Dictionaries: an Evaluation' in *Machine Translation and the Lexicon*, Springer Verlag, Lecture Notes in Artificial Intelligence 898: 19–34

Miller, G. (1990) 'Wordnet: An On-line Lexical Database', *International Journal of Lexicography* (special issue), No. 4, Vol. 4: 235–312

Wu, D. (1995) 'An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words' in *Proceedings, 33rd Annual Meeting of the Association of Computational Linguistics*, MIT: 244–51

A Corpus-based Bilingual Dictionary: Why and How?

by

Marie-Hélène Corréard

Oxford University Press

In the following talk I shall describe how the Oxford-Hachette French Dictionary (OHD) was written with the help of French and English corpora and how the use of this corpus data has helped the editors to make it a better dictionary.

First I shall look at the general set-up of the project. The dictionary was written by two teams of native speakers in one location (Oxford), each working in their own language and cross-checking with each other for accuracy. The method chosen to compile the dictionary, which is a completely new text and not a revision of an existing one, made good use of the most recent developments in language research combined with the traditional craft of the lexicographer. The dictionary was produced in three stages: compilation, translation, and editing; but its distinguishing feature was the recourse to corpus data at all stages, more particularly at the editing stage.

Next I shall examine why it was necessary to use corpora, what sort of corpora were available for editors, and when they were most useful. I shall also look at concrete examples of how we used the corpora.

No human being, even with all the time in the world, can be sure of remembering all the ways of using one word, let alone how other people use it. Traditional dictionaries, however well researched, have the same problem and do not always represent all the ways in which a word is used. In addition to this, languages evolve and grow. Some words appear while others go out of use. Simply because of the time it takes to write them, dictionaries do not always reflect these changes. Computerized corpora, however, allow editors to do just that.

A dictionary of modern language should reflect everybody's use of language, not just the language of one particular editor or group of editors. Using a corpus allows editors to see how a broad and varied range of people have used the word they are describing.

The editors had access to two corpora of ten million words each, one in French and one in English. With the help of search and display tools they could, in a matter of seconds, draw out all the occurrences of a particular word. Further, they could customize the search to suit their needs by choosing whether they wanted to alphabetize the results on the keyword, on the word following it, or on the word preceding it, according to the type of entry and the language they were working on. These corpora included the text of letters, books, journals, and newspapers, as well as transcripts of conversations, lectures, and discussions.

Deliberately chosen for their contemporary contents, our corpora were not suitable for illustrating words that are to be found more in classical literature than in the contemporary language. This does not mean that such words were not included, but simply that we had to use other sources to describe them, notably monolingual dictionaries and other corpora. The fact that a word was not in the corpus was not an argument for discarding it from the dictionary altogether. Conversely, the fact that a word was in the corpus did not mean it had to be in the dictionary; it could be a highly specialized term generated by an event much covered by the media, for example, the Challenger disaster and the subsequent enquiry. When working with corpus data, editors always had access to the source of the citations and were therefore able to moderate some interesting finds if it turned out that they were all from one author.

The corpora were most useful for core vocabulary, new senses, and new words. We used corpus data for several tasks:

1. when working on the source language: to improve the coverage of core vocabulary and to ensure that all essential complementation patterns (one of the greatest pitfalls for non-native speakers) were given;

2. when working on the target language: to improve the quality of translations;

3. when checking the finalized entry: to ensure that our user would be well equipped to use correctly all the information provided in as broad a range of contexts as possible; this was achieved by having lots of contexts against which to test both source and target language elements.

Finally I shall consider the results of using corpora in the compilation of the dictionary. In general it is true that the dictionary is a better product than it would have been without the use of corpora. The benefits can be summarized as follows:

1. The dictionary is more comprehensive: the wordlist is more extensive; entries provide more examples of common phrases and the most common senses are covered.

2. The dictionary is more accurate: it reflects more accurately the way French, for example, is used by native speakers, since all the example sentences are taken from real examples of French in our corpus.

3. The dictionary is more reliable: all the translations are authentic because they have been checked in the corpus.

4. The dictionary is more user-friendly: all the material the user needs is to be found in the entry; furthermore the translations have been checked against a wide range of real contexts in which they might be used.

5. The dictionary is safer: because it is possible to verify all the usages in real language, the restrictions that accompany certain usages can be clearly indicated; all essential complementation patterns are shown.

6. The dictionary is more up-to-date: the corpus was updated throughout the entire editing period and new text was added up to the last minute.

In conclusion we may note that we have used corpus data for other tasks, notably for writing lexical notes and compiling and editing grammar words. The use of corpora brings a new dimension to lexicography, making dictionaries more reliable and more comprehensive. It is particularly valuable for bilingual dictionaries because the same corpus is used for different tasks at the various stages of the actual writing of the dictionary. Also it enables editors to cater better for all the different users of a bidirectional dictionary. Since the publication of the unabridged OHD, we have published a unidirectional dictionary; once again, the corpora were essential for selecting the most important items to be represented in the dictionary.

Report on the
Second Language Engineering Convention
London, 16–18 October 1995

When receiving notice of this Convention one at first wondered whether it was a Convention on Second Language Engineering or the Second Convention on Language Engineering. Perhaps the ambiguity was intended by the organisers as a test of the perception of language engineers and others gathering for the Convention!

The Convention, organized by the D.T.I. and the European Commission, included a programme of presented papers and an exhibition of some of the latest language engineering systems.

The papers presented covered a wide range of themes, including dialogue analysis, speech recognition, CALL, aids for the disabled, and standards. A significant number of papers dealt specifically with Machine Translation (MT); others had relevance to MT; most were undoubtedly of interest to each of the delegates there.

The importance of language engineering (LE) was emphasized in the address of welcome by Mr Roberts Cencioné of the European Commission. If Europe was to be truly multilingual, with no one language becoming dominant, this would put a considerable burden on language technology to provide facilities in a variety of areas including language engineering and international business.

'It Would be Good to Talk' was the title of the first main paper, by Professor Peter Cochrane of BT Laboratories. This paper reflected the expected spread of speech-processing systems in many aspects of everyday life, in person–with–machine communication or in person–through–machine–with–person communication. It also reflected much of the scope of the other papers presented. Such systems would involve a considerable degree of linguistic analysis which, he warned, could introduce artificialities such as pauses into dialogue. Such pauses could lead to unintended misinterpretation, as for example:

> Wife to husband on the phone: 'Are you with another woman?'
>
> Pause.
>
> Husband: 'No'.

Applications of LE to telephony were considered in a number of other papers including one entitled 'Back to the Future ...' by Dr N. Fraser of Vocalis Ltd. He pointed out that telephone design reflects current technology rather than users' needs. LE would allow the incorporation of a 'virtual operator' into a telephone system and bring back all the advantages that users enjoyed in the early days of telephones — sans dial, sans key-pad, sans telephone directory.

Telephony, of course, is one context in which MT could be applied. Such a speech–to–speech translation system, developed at SRI International (Cambridge, UK), was described by Dr M. Rayner. He discussed the need for such a system to be portable, reconfigurable and customisable, but also stressed some of the difficulties, for example, the variations between even Mexican Spanish and Puerto Rican Spanish in an English–with–Spanish interpreting system. Also there was the problem of how to check that the listener had correctly recognized what the speaker had said.

Other papers covered topics on the future of translation services, particularly on the information highway. One such paper, entitled ‘No Highway without Service Stations: TELELANG’, was presented by Mr Z. Karssen of EUROLANG, France.

Minority languages were not overlooked and an interesting presentation, on applications of LE to Welsh, was given by Mr M. Williams of the Welsh Joint Education Committee.

The importance attached to standards was indicated by a whole session being devoted to this subject. This, interestingly, re-echoes the concern shown within the NLTSG about the question of standards.

Considerable time was also devoted to the Third Language Engineering Call for Proposals, initiated by the European Commission.

The exhibition included demonstrations of a number of speech recognition systems. Some were demonstrated as components of speech input word-processors. Even with pauses between words, these are clearly more satisfactory than those in which each letter of a word has to be keyed in. The days of QWERTY Keyboards may well be numbered! Other demonstrations were of continuous speech recognition systems, which could well be incorporated in many of the LE systems discussed in the Convention.

Conventions, like conferences, are undoubtedly far from easy to organize. Dr Richard Sharman, of IBM (UK), and his associates are to be congratulated on the excellence of the planning and organizing of this Convention.

[Further information about the Convention is available from the NLTSG].

Douglas Clarke

Degree Courses

Some readers may be interested to know where they could start to prepare themselves for a career in Machine Translation. An examination of the UCAS Handbook for 1996 entry reveals the following Universities offering Joint Degrees in Computing and Linguistics;

University of East London
Barking Campus
Longbridge Road
Dagenham
Essex RM8 2AS

BA/BSc and BA/BSc Honours in Information Technology and Linguistics

The University of Edinburgh
Edinburgh EH8 9YL

MA Joint Honours in Linguistics and Artificial Intelligence

The University of Essex
Wivenhoe Park
Colchester CO4 3SQ

BA Single Honours in Computational Linguistics

BSc Joint and Combined Honours in Computer Science with French, German, Russian and Spanish

The University of Kent at Canterbury
Canterbury
Kent CT2 7NZ

Combined Honours Courses in Computing and Linguistics

Lancaster University
The University
Lancaster LA1 4YW

BSc/BA, with Honours, in Computer Science and Linguistics

University of Leeds
The University
Leeds LS2 9JT

BA Honours in Linguistics and Computing

Luton College of Higher Education
Park Square
Luton
Beds LU1 3JU

BA and BSc Honours in Computer Science and Linguistics

The University of Manchester Institute of Technology
Manchester M60 1QD

BSc Honours in Computational Linguistics with French, German, Japanese or Spanish

Queen Mary and Westfield College
(University of London)
Mile End Road
London E1 4NS

BA Honours (4 yrs) in Computer Science and Linguistics with French or German

The University of Sheffield
Sheffield S10 2TN

BSc Honours in Computer Science with Modern Languages

University of Southampton
Southampton SO17 1BJ

BSc (Honours) in Computer Science and Modern Languages

University of Sussex
Sussex House
Falmer
Brighton BN1 9RH

BA Honours in Linguistics in Cognitive and Computing Sciences

University of Ulster
Coleraine
Co.Londonderry
N.Ireland BT52 1SA

BSc Honours in Computing and Linguistics

University of Wolverhampton
Wulfruna Street
Wolverhampton WV1 1SB

BA/BA Honours or BSc/BSc Honours in Computing with Linguistics

Other Degrees

There were too many Universities offering joint degree courses in a computing subject with one or more foreign language to list them individually! If you are interested you can get a copy of the UCAS Handbook for 1996 entry by telephoning +44 (0) 1242 227788.

JDW

Book Reviews

Bernice Sacks Lipkin (1994) ***String Processing and Text Manipulation in C,***
Prentice Hall. (Sbk) £27.50. ISBN 0-13-121443-8.

This book satisfies the apparent American criteria for quantity. It has 433 pages of closely packed information in an A5 format and a thirteen-page index.

Of the seven chapters, the first is a very useful review of the basic concepts of C which actually takes more than its fair share of the book — 78 pages. The book assumes a general knowledge of programming, and although it disclaims any pretence of being a general text for the C language or grammar, its coverage is comprehensive, not being confined to just text processing.

Chapter Two plunges into a detailed analysis of text data, referred to as strings — how these are declared, defined, and loaded into preset or dynamic storage where they can be processed.

The next chapter explores the use of pointers in text processing and explains some of the finer points of C syntax and semantics. Lots of detailed examples are provided.

Chapter Four starts to introduce application-oriented functions such as the 'setbreak' utility and 'bitmapping' functions which the newcomer to C may be forgiven for thinking are part of the C language. In fact, the author has slipped into describing particular functions 'transcribed' from a particular text processing system called TXT and written in another language called SAIL.

The fifth chapter reverts to some in-depth analysis of techniques for storing and handling text data in two-dimensional arrays, as in tables.

Chapter Six introduces the theory and practice of structures and linked lists in C. It shows how they can be implemented for text processing, as always, with numerous examples.

The final chapter concentrates on a set of loosely defined functions for processing individual strings, said to be a selection taken from TXT; these are also on the disk supplied with the book. In fact all the examples of programs shown in the text are on the disk and the few that I tried worked correctly.

I think this book would be most useful for inexperienced programmers who find some of C's features, such as pointers, difficult to use. At the same time they may have to spend some time going through the numerous examples in order to find the ones they want. Experienced C programmers, on the other hand, would probably soon lose interest in the text, but they might be interested in picking up the coded functions supplied on the disk; these are described in considerable detail in the text and should be reasonably easy to tailor for one's own use.

To summarize, this is a rather long text for the subject matter, but some beginners and some experienced programmers might just find what they are looking for here. My advice is to have a good look at it first before paying the asking price.

David Wigg

A. M. McEnery (1992) *Computational Linguistics — a Handbook and Toolbox for Natural Language Processing*, Wilmslow: Sigma Press. Paperback. £14.95. 263 pages. ISBN 1-85058-247-5.

This book has been written with clear aims: to present in an accessible style some of the principles of NLP; to provide examples of working systems; to enable the reader to implement some of these systems on a standard basic PC; and to encourage the reader to embark on further research and to experiment by writing his own application programs.

The book begins by explaining the nature and scope of computational linguistics: concerned on the one hand with linguistic theory and on the other with the practical requirements of implementing commercial language processing systems, the discipline faces in two opposite directions — what the author calls the Janus Principle. The core fields of speech processing and text processing (which includes machine translation) are reviewed in terms of their main concerns and their typical application areas. The author divides approaches to computational linguistics into cognitive (holistic) and probabilistic (reductionist) ones. While the first approach tries to model language in terms of how the whole human mind perceives and handles information, the second concentrates on simulating individual language phenomena by using statistical processing techniques to achieve specific goals. The distinction is partially mirrored by the difference between declarative and procedural programming languages: to illustrate this, the author contrasts the operation of a short backtracking PROLOG programme with a routine written in C.

Starting out from the principles of information theory, the author describes a simple algorithm for calculating the statistical probability and information content of an event. Transferring these principles to language, he outlines the operation of Markov models and transition probabilities; there is a detailed exposition of how such models operate for individual words, based on a ‘database’ of two short sample sentences. The explanatory approach and the presentation of the process for setting up tables of probabilities for words are excellent. But the more interesting procedure for deciding what syntactic functions a word can have in syntagmatic relationships may not be so transparent to the lay reader. The data presented in illustrative tables would benefit greatly from more explanation for those unfamiliar with statistical methods: one transition item appears twice for no apparent reason, and it is unclear how the sum of probabilities is arrived at (page 53). The copy editing could have been improved: there is, for instance, some confusion between the terms digrams, diagrams and equations (pages 50–51). Nevertheless, the reader can gain a general idea of how Markov models operate.

Markov models are used extensively in the Lancaster University-based UCREL project. UCREL (Unit for Computer Research on the English Language) was established to provide automatic grammatical tagging of the one million word Lancaster-Oslo-Bergen (LOB) corpus of English texts. Various versions of the CLAWS (Constituent-Likelihood Automatic Word-Tagging System) software system, which incorporates a Markov model, were used to carry out the tagging. Most recently UCREL has concentrated on using a stochastic Markov-based syntactic parser (developed from the LOB corpus) to support an acoustic speech processing system: the idea is that the parser resolves ambiguities in spoken input that the acoustic component (which is also Markov-based) is unable to by itself. The book gives a detailed account of the some of the processes and problems involved in operating a probabilistic tagging mechanism. Although CLAWS tags individual words only, some progress has been made towards developing a Constituent Likelihood Grammar that arranges tagged words into structural units, or phrases, between word and sentence level.

As an example of a cognitive parsing mechanism, the book describes the operation of a small Augmented Transition Network written in PROLOG. A simple knowledge base in which objects and properties are arranged as linked attribute-value sets is also presented. Maintaining his balance between theoretical research and practical considerations, the author points out that, despite the relative ease with which such grammars can be written in PROLOG, the language is inefficient in terms of memory requirements and processing speed, especially for large applications.

Two large application systems, both implemented in PROLOG, are presented in some detail. The first is 'The Schoolboy' — an extensive database of verb forms which has been used in teaching environments to illustrate various aspects of linguistic systems design, including the principles of morphological analysis and lexicon construction. The second system is the Hierophant Project (1984–89). The object of this project was to implement a natural language querying system for accessing and returning responses from a database of social security information. The developers considered three methods for producing anglicized output: super slot attributes, semantic attributes, and an ATN frame grammar. They finally chose the frame technique, which uses pattern matching to see if an input sentence matches a predefined 'frame' or structural template. Natural language output is also derived from the progressive selection of appropriate frames: a so-called application-sensitive corpus-based approach is used to set up a sub-grammar of frames that are finely tuned to the characteristics of the corpus. The author goes to considerable pains to explain how the system works, although the material can become rather technical on occasion, as the following sentence shows: 'The savings would be gleaned from the system as there would be no need to find the real value after it has been ascertained that it is not the target value, as the mechanics of objects only able to inherit a single item in a slot will give that information implicitly in combination with logic, thus the data item may be deduced as a virtual item' (page 148). The developers of Hierophant were especially concerned with providing informative output in natural language form: the author has some particularly interesting observations on the communicative aspect of language-based interaction between the computer and the human being.

The final section discusses the so-called Bench-Capon and McEnery (BCM) Hypothesis, which warns against regarding interpersonal linguistic interaction between human beings as a model for building computerized natural language interfaces. According to the BCM the user should not be lured into pretending he is interacting with another human being but acknowledge that the computer is a medium or channel of intercourse like many others. Drawing on insights from intensional linguistics and pragmatics theory, the author maintains that such an awareness would help avoid misunderstandings and errors on the part of the human user and enable appropriate codes of communication to be developed.. The BCM Hypothesis is contrasted with the Barlow, Rada and Diaper Antithesis, which regards the (inanimate) computer as an equal interlocutor in the user's mind operating on the same communicative level.

In conclusion, this book is impressive for its range of coverage and its attempt to integrate the theory of linguistic communication with programming practice. The appendices contain listings of the systems presented. The book should most certainly act as a stimulus for its reader to explore the field further.

Derek Lewis

W. John Hutchins and Harold L. Somers (1992) ***An Introduction to Machine Translation*****, London: Academic Press. Hardback. £19.95. 362 pages. ISBN 0-12-362830-X.**

As the authors of this book state in their introductory chapter, Machine Translation is an interesting combination of a number of disciplines, drawing on such diverse subjects as linguistics, computer science, artificial intelligence, and translation theory. It also consists of an unusual combination of an abstract, theoretical, highly technical side and a commercial side. Thus the field of MT has seen a number of real life, implemented systems which correspond in some cases very closely to a theoretical model and in other cases considerably less so. The two authors embody these two contrasting aspects: John Hutchins is well known and highly respected throughout the world of MT for his knowledge of the history of the subject and for his overview of current systems; Harold Somers enjoys a similar reputation as a result of the leading-edge practical and theoretical work that he conducts at UMIST. This book is the result of their combined skills and knowledge of the subject.

It seems to me that this eighteen-chapter volume goes considerably beyond its modest title of an introduction to the subject: in my own experience of using it as a textbook for a final year undergraduate module in MT, it comes as close as any to being the complete textbook for the course, and as such has no direct competitor in the field. The first three chapters provide the essential underpinnings of the subject for total beginners: a general context and a brief history; an introduction to the linguistic background; followed by an introduction to the computational background.

Each chapter introduces itself with some useful comments to help the reader find his way around the topic: for example, suggestions as to who might not need to read the chapter, or which bits the more experienced reader might wish to skip; and each chapter is completed by a short section on further reading, which is a nice way of saying, ‘well, if you did not find what you were hoping for in our book, here are some other places you could look for it.’ The authors provide an interesting mix of factual information and opinion in these sections, an informal note which contrasts agreeably with some of the drier, technical sections.

The next group of chapters, numbers four through to nine, cover the various theoretical approaches and break down the theory of MT into digestible chunks. There is a chapter on basic strategies (Chapter Four), followed by one each on analysis (Five) and generation (Seven); Chapter Six discusses the problems of transfer and interlingua. As a general rule the authors use examples extensively, drawing from a good selection of familiar European and less familiar language groups; this approach is highly effective in fleshing out the points being made. Each chapter can be read in its entirety or can be tackled in parts by less experienced readers, who are encouraged to feel they can return to grapple with a more difficult point at a second attempt. Chapter Eight offers an account of the practical uses of MT systems and covers issues which will be much more familiar to translators. This is followed by a chapter on the evaluation of MT systems, which is the final part of the general material in the book.

The next eight chapters of the book provides a more detailed account of MT systems. These are in turn: Systran, SUSY, Meteo, Ariane (GETA), Eurotra, Metal, Rosetta, and DLT. At the time of writing (1992) these chapters present an up-to-date description of both the commercial and the research aspects of major recent work in MT; they further provide both a counter-balance and some illustration to the preceding chapters of the book. Each system is placed in its historical context; an account is provided of the particular features and technical methods used and each report offers a section of summary, conclusions and discussion, which makes no apology for its subjective commentary. These are very useful for the newcomer. each chapter has the usual excellent suggestions on further reading for those interested in looking in more detail.

The final chapter looks to the future, acknowledging that contemporary MT research offers a confusing picture. The concluding comments apply not only to that chapter, but in fact to the whole book. The enthusiasm of the authors for their subject comes over here as throughout the volume.

I would happily recommend this book, both for the kind of undergraduate introductory course where students have to come to terms very quickly with the unfamiliar terminology and theory of Machine Translation, and also to the interested layperson, who is setting out to explore what the subject encompasses. I find the informal approach and language in parts makes the complexity of some of the material much easier to absorb, and I cannot think that I have come across a more thorough introduction to any other subject at such a reasonable price. It is almost within the purchasing capacity of a student!

Catharine Scott

Conferences and Workshops

The following is a list of recent or forthcoming conferences and workshops. Telephone numbers and e-mail addresses are given where known.

2–4 October 1995
First AGFL Workshop on Syntactic Description and Processing of Natural Language
AGFL Secretariat
Department of Computer Science, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands.
Tel: +31 80 653456, fax: +31 80 553450, e-mail: www-agfl@cs.kun.nl

16–18 October 1995
2nd Language Engineering Convention
Queen Elizabeth II Conference Centre, London, UK.
Mrs Dion Bassett
Concorde Services Ltd, 10 Wendell Road, London, W12 9RT, UK.
Tel: +44 (0) 181 7433106, fax: +44 (0) 181 7431010

23–27 October 1995
AVINTA, UNEG, UNEXPO, CVG
CNIASE 95 VIII National Conference on Artificial Intelligence
Ciudad Guayana, Venezuela.
http://www.cc.gatech.edu/ai/cniase95/home.html

2–3 November 1995
2nd ‘SPEAK!’ Workshop: Speech Generation in Multimodal Information Systems and Practical Applications
John Bateman
GMD/IPSI, Darmstadt, Germany.
Tel: +49 6151 869 826, fax: +49 6151 869 818, e-mail: bateman@gmd.de

9–10 November 1995
Translating and the Computer 17: Conference and Exhibition
Nicole Adamides, Events Manager, Aslib,
The Association for Information Management, 20-24 Old Street, London, EC1V 9AP, UK.
Tel: +44 (0) 171 253 4488, fax: +44 (0) 171 430 0514, e-mail: pdg@aslib.co.uk
http://www.aslib.co.uk/aslib/

1–2 December 1995
AMLaP-95
Architectures and Mechanisms for Language Processing
Edinburgh, Scotland.
E-mail: amlap@cogsci.ed.ac.uk

4–6 December 1995
Third Natural Language Processing Pacific-Rim Symposium
Sofitel Ambassador Hotel, Seoul, Korea.

6–8 December 1995
First AMAST Workshop on Language Processing
Algebraic Methods in Language Processing
University of Twente, Enschede, The Netherlands.

27 and 28 December 1995
PACLIC10: The 10th Pacific Asia Conference on Language, Information and Computation
Language Information Sciences Research Centre
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong.
Fax: 852 27889443, e-mail: rlpaclic@cityu.edu.hk

29 January–2 February 1996
Computers in Engineering Symposium
Session on Natural Language in Human-Computer Interfaces, Houston, USA.
Susan Haller
Computer Science and Engineering Department
University of Wisconsin – Parkside, Kenosha, WI 53141-2000.
Tel: +414 595 2343, fax: +414 595 2114, e-mail: haller@cs.uwp.edu

17–20 March 1996
TARK VI: Sixth Conference on Theoretical Aspects of Rationality and Knowledge
Department of Mathematics and Computer Science, Utrecht University
Padualaan 14, De Uithof, 3508 TB, Utrecht, The Netherlands.
Tel: +31 30 534117, fax: +31 30 513791, e-mail: peter@fwi.uva.nl/jj@cs.ruu.nl

26–27 March 1996
CLAW (Controlled Language Applications)
Centre for Computational Linguistics
Katholieke Universiteit Leuven, Maria-Theresiastraat 21, B-3000 Leuven, Belgium.
Tel: +32 16 325088, fax: +32 16 325098, e-mail: claw96@ccl.kuleuven.ac.be

1–2 April 1996
Society for the Study of Artificial Intelligence and Simulation of Behaviour (SSAISB)
AISB-96
Alison White
School of Cognitive and Computing Sciences, University of Sussex, Brighton BN1 9QH, UK.
Tel: +44 1273 678448, fax: +44 1273 671320, e-mail: alisonw@cogs.susx.ac.uk
http://www.cogs.susx.ac.uk/aisb/aisb96/cfw.html

15–17 April 1996
SDAIR ’96: Fifth Annual Symposium on Document Analysis and Information Retrieval
Alexis Park Resort, Las Vegas, Nevada, USA.
University of Nevada, Las Vegas, 4505 Maryland Parkway.
Box 454021, Las Vegas, NV 89154-4021

22–26 April 1996
PAP & PACT: PAP '96: The Fourth International Conference on the Practical Application of PROLOG
London, UK.
Al Roth
Tel: +44 (0) 1253 358081, fax: +44 (0) 1253 353811, e-mail: info@pap.com
http://www.demon.co.uk/ar/PAP96/index.html

9–11 May 1996
International Translation Studies Conference
Dublin City University
Dublin 9, Republic of Ireland.
Fax: +353 1 7045527

4–7 June 1996
ICCC '96: International Conference on Chinese Computing '96
Institute of Systems Science
National University of Singapore, Singapore 0511.

13–15 June 1996
INLG '96: 8th International Workshop on Natural Language Generation
Herstmonceux, Sussex, UK
Tel: 01273 642900, e-mail: inlg96@itri.brighton.ac.uk

25–29 June 1996
Association for Literary and Linguistic Computing,
Association for Computers and the Humanities
Joint International Conference ALLC-ACH '96, University of Bergen, Norway.
Espen Ore
Norwegian Computing Centre for the Humanities
Harald Haarfagresgt. 31, N-5007 Bergen, Norway.
Tel: +47 55 21 28 65, fax: +47 55 32 26 56, e-mail: Espen.Ore@hd.uib.no
http://www.hd.uib.no/allc-ach96.html

17–18 July 1996
DAARC96 – Discourse Anaphora and Anaphor Resolution Colloquium
Lancaster University
Dr Tony McEnery
Department of Linguistics and MEL, Lancaster University, Lancaster LA1 4YT, UK
Tel: 01524 65201 ext. 3024, fax: 01524 843 085, e-mail: mcenery@comp.lancs.ac.uk

4 August 1996
WVCL-4 (Fourth Workshop on Very Large Corpora)
ACL Special Interest Group Workshop (SIGDAT)
University of Copenhagen, Copenhagen, Denmark.
E-mail:WVLC-4@ling.umu.se
http://www.ling.umu.se/SIGDAT/WVLC-4.html

5–9 August 1996
COLING 96: International Conference on Computational Linguistics
University of Copenhagen, Denmark.
Prof. B. Maegaard
Centre for Sprogteknologi, Njalsgade 80, DK-2300 Copenhagen S, Denmark.

12–23 August 1996
ESSLI: European Summer School in Logic, Language, and Information, Prague
Malostranské nám. 25
118 00 Praha 1
Czech Republic.
Tel. +42 2 24510286, fax: +42 2 532742, e-mail: essli@ufal.mff.cuni.cz

12–16 August 1996
12th European Conference on Artificial Intelligence
Budapest, Hungary.
Dr Elisabeth Andre, Workshop Coordinator
German Research Center for AI, DFKI GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany.
Tel: +49 681 3025267, fax: +49 681 3025341, e-mail: ecai-96-ws@dfki.uni-sb.de

13–18 August 1996
EURALEX ’96: Seventh Euralex International Congress
University of Gothenburg, Sweden.
Department of Swedish, Section of Lexicology
412 98 Gothenburg, Sweden.
Tel: +46 317734544, fax: +46 31773 44 55 (Att. EURALEX), e-mail:
gellerstam@svenska.gu.se

11–13 September 1996
NeMLaP-2: International Conference on New Methods in Natural Language Processing
Bilkent University
Bilkent TR-06533, Ankara, Turkey.
http://www.cs.bilkent.edu.tr/~nemlap2/nemlap.html.

25–27 September 1996
RECITAL’96 Rencontre des Etudiants-Chercheurs en Informatique pour le Traitement
Automatique de la Langue
Ferrari Stephane
GT SemLex, Groupe Langage et Cognition
Dpt Communication Homme-Machine, LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France.
Tel: +33 1 69858018, fax: +33 1 69858088, e-mail: ferrari@limsi.fr

MEMBERSHIP RENEWAL

As membership is currently still free we must review the membership list from time to time. To renew membership please fill in the form below (or a copy) and return to:

Mr.J.D.Wigg
BCS-NLTSG
72 Brattle Wood
Sevenoaks
Kent TN13 1QU
UK. Date: ....../....../......

Name: ....
Address: ....Postal Code:...
Country: ....E-mail: ....
Tel.No: ....Fax.No: ....

Questionnaire

We would like to know more about you and your interests and would be pleased if you would complete as much of the following questionnaire as you wish (Please delete any unwanted words).

1. a. I am mainly interested in the computing/linguistic/user/all aspects of MT.
   b. What is your professional subject? ....
   c. What is your native language?...
   d. What other languages are you interested in? ....
   e. Which computer languages (if any) have you used? ....

2. What information in this Review (No.2, Oct. '95) or any previous Review, have you found:
   a. interesting? Date ....
   b. useful (i.e. some action was taken on it)? Date ....

....
....
....

3. Is there anything else you would like to hear about or think we should publish in the MT Review?

....
....
....
....

4. Would you be interested in contributing to the Group by,

   a. Reviewing MT books and/or MT/Multilingual software
   b. Researching/listing/reviewing public domain MT and MNLP software
   c. Designing/writing/reviewing MT/MNLP application software
   d. Designing/writing/reviewing general purpose (non application specific) MNLP procedures/functions for use in MT and MNLP programming
   e. Any other suggestions?

....
....
....
....

Thank you for your time and assistance.

MT Review Oct.'95