MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society
Issue No. 4
October 1996

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
United Kingdom

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

ISSN 1358–8346

Contents

Group News and Information

Letter from the Chairman

Our Group may not appear to have been very active this year but some things have been going on behind the scenes.

Roger Harris has been beavering away on the Group's web pages and I have to restrain him from swamping the BCS with updates. If you have any thoughts about what else should be in these pages please let him know.

Douglas Clarke and Alfred Vella have been struggling to finish off the last few talks for the Proceedings of the 1994 Conference at Cranfield, and I regret I have to apologise again for the delay in publication. Like most new programs, it seems to have been 99% complete for the last year.

Derek Lewis has been busy 'drumming up business' for this Review, which he has produced again with his customary aplomb. Don't forget that our Review could be your opportunity to get your thoughts on machine translation into print quickly without having to go through the long process of a peer review.

You may remember I tried to start a small scale experimental MT project but was unable to arouse sufficient enthusiasm, no doubt because of the complexity of the problem and the time and effort that would be required to achieve anything worthwhile, so I was pleased to see Gareth Evans' contribution in the previous Review in April.

I now wonder if some one or some body could review some of the other MT-related software which is now appearing on the Internet, for publication in the Review. If we cannot write any software for ourselves perhaps we could look at other people's programs and learn something from them.

Initially, I am thinking of parsers such as the LINK system from Carnegie Mellon and the Apple Pie parser from New York University, but there are others. Any volunteers?

Finally, I must thank the speakers, Derek Lewis and Ruslan Mitkov, for their contribution to our events during the past year. The rather small number of talks is mainly due to a change of emphasis towards activities which we believe reach rather more people than before, but they do cost us more and we are currently drawing on our reserves.

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

J.D.Wigg

The Committee

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

David Wigg (Chair)	Tel.: +44 (0)1732 455446 (H)
	Tel.: +44 (0)171 815 7472 (W)
	E-mail: wiggjd@vax.sbu.ac.uk
Monique L'Huillier (Secretary)	Tel.: +44 (0)1276 20488 (H)
	Tel.: +44 (0)1784 443243 (W)
	E-mail: m.lhuillier@vms.rhbnc.ac.uk
Ian Thomas (Treasurer)	Tel.: +44 (0)181 464 3955 (H)
	Tel.: +44 (0)171 382 6683 (W)
Derek Lewis (Editor)	Tel.: +44 (0)1404 814186 (H)
	Tel.: +44 (0)1392 264330 (W)
	Fax: +44 (0) 1392 264377
	E-mail: d.r.lewis@exeter.ac.uk
Catharine Scott (Assistant Editor)	Tel.: +44 (0)181 889 5155 (H)
	Tel.: +44 (0)171 607 2789 X 4008 (W)
	E-mail: c.scott@unl.ac.uk
Roger Harris (Rapporteur)	Tel.: +44 (0)181 800 2903 (H)
	E-mail: rwsh@dircon.co.uk
Correspondent members:	
Gareth Evans (Minority Languages)	Tel.: +44 (0)1792 481144
	E-mail: g.evans@sihe.ac.uk
Ruslan Mitkov	Tel: +44 (0)1902 322471 (W)
	Fax: +44 (0)1902 322739
	E-mail: R.Mitkov@wlv.ac.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 am to 5.00 pm.

Anaphora and Machine Translation

by

Ruslan Mitkov

School of Languages and European Studies
University of Wolverhampton

1. *Anaphora: Basic Notions and Revision of Terminology*

The etymology of the term ‘anaphora’ goes back to Ancient Greek with αναφορα (anaphora), a compound word consisting of the separate words ανα (*back, upstream, back in an upward direction*) and φορα (*the act of carrying*) and denoting ‘the act of carrying back upstream’. For computational linguists starting to explore the field of anaphor resolution, I strongly recommend as a primer Graham Hirst’s *Anaphora in Natural Language Understanding* (Hirst 1981). Although this book may seem a little dated in that it does not include developments in the 1980s and 1990s, it nevertheless provides an excellent survey of the theoretical work on anaphora and of the early computational approaches to the problem; as such it is still very useful reading.

Various definitions of anaphora have been put forward. But I am tempted to paraphrase the classical definition given by Halliday and Hasan (1976), which is based on the notion of cohesion: anaphora is cohesion (presupposition) which points back to some previous item.

The ‘pointing back’ (reference) is called an *anaphor* and the entity to which it refers is its *antecedent*. The process of determining the antecedent of an anaphor is called *anaphor resolution*. The majority of researchers (including until recently the author of this survey) use the term ‘anaphora resolution’, which I consider less accurate than ‘anaphor resolution’. Anaphora is a linguistic phenomenon: what is resolved is therefore *not* the phenomenon but the anaphor (reference) which is initially regarded as ‘unknown’ and whose antecedent must be tracked down. This interpretation compares, for instance, with the use of the term ‘resolution’ in other areas, such as logic; we should note also that ‘pronoun resolution’ is an acceptable term, whereas ‘pronominalisation resolution’ is not).

2. *Anaphor Resolution: Latest Developments*

After considerable and successful initial research in anaphor resolution (Wilks 1975, Hobbs 1976, Webber 1978, Sidner 1979), the late 1980s and 1990s saw a revival of interest when various projects were reported after years of relative silence. We can categorise the best known works either as ‘integrated/knowledge-based’ (i.e. integrating various traditional linguistic constraints, preferences and knowledge sources) or as ‘alternative’ (i.e. drawing on non-traditional techniques and resources, such as corpora, neural networks and uncertainty reasoning). Among the integrated approaches worth mentioning are: D. Carter’s shallow processing approach (Carter 1987); E. Rich and S. LuperFoy’s distributed architecture (Rich and LuperFoy 1988); J. Carbonell and R. Brown’s multi-strategy approach (Carbonell and Brown 1988); C. Rico Pérez’ scalar product coordinating approach (Rico Pérez 1994); and Mitkov’s combined approach (Mitkov 1994). Successful alternative strategies that have been reported include the following: Nasukawa’s knowledge-independent approach (Nasukawa 1994); Dagan and Itai’s statistical/corpus processing approach (Dagan and Itai 1990);

Connolly, Burger and Day's machine learning approach (Connolly, Burger and Day 1994); Mitkov's uncertainty-reasoning approach (Mitkov 1995); Mitkov's practical approach (Mitkov 1996b); and Kennedy and Boguraev's tagger-based approach (Kennedy and Boguraev 1996) . There is also a trend towards developing corpus-based, parser-free, knowledge-independent and practical approaches which are aimed at overcoming the notoriously difficult (and inaccurate) task of knowledge representation and processing (see Dagan and Itai 1990, Nasukawa 1994, Mitkov 1996b and Kennedy and Boguraev 1996).

3. *Anaphor Resolution in MT Systems: Theoretical Issues*

The establishment of the antecedents of anaphors is of crucial importance for correct translation. When translating into languages which mark the gender of pronouns, for example, it is essential to resolve the anaphoric relation. On the other hand, anaphor resolution is vital when translating discourse rather than isolated sentences since the anaphoric references to preceding discourse entities have to be identified. Unfortunately, the majority of Machine Translation systems do not deal with anaphor resolution and their successful operation usually does not go beyond the sentence level.

Anaphor resolution as analysis is a tough problem, but translation adds a further dimension in so far as the reference to a discourse entity encoded by a source language anaphor by the speaker (or writer) has not only to be identified by the hearer (translator or translation system) but also re-encoded in a co-referential/cospecificational expression of a different language.

3.1 *The Translation of Pronominal Anaphors*

In the majority of cases and language pairs the pronouns in the source language are translated by target language pronouns which correspond to the antecedent of the anaphor. However, there are a number of exceptions. In some languages, the pronoun is translated directly by its antecedent. In English-to-Malay translation, for instance, there is a tendency to replace 'it' with its antecedent. Replacing a pronominal anaphor with its antecedent means, however, that the translator (program) must be able to identify the antecedent first.

Very often, pronominal anaphors are simply omitted in the target language. For example, although English personal pronouns have their correspondences in Spanish, they are frequently not translated because of the typical Spanish elliptical zero-subject constructions. Target languages with typical elliptical (zero) constructions corresponding to source English pronouns are Spanish, Italian, Thai, Chinese and Japanese.

Another interesting example is English-to-Korean translation. English pronouns can be omitted elliptically; they can also be translated by a definite noun phrase, by their antecedent, or by one or two possible Korean pronouns, depending on the syntactic information and semantic class of the noun to which the anaphor refers (Mitkov et al. 1994).

3.2 *The Necessity for Anaphor Resolution in MT*

While in most European language pairs anaphor resolution is 'compulsory' (to avoid the risk of quite unacceptable translations in certain cases), there are certain language pairs where anaphor resolution may seem 'optional'. As an illustration, consider the following sentences (taken from Hutchins and Somers 1992):

(1) The monkey ate the banana because it was hungry.

(2) The monkey ate the banana because it was ripe.
(3) The monkey ate the banana because it was tea-time.

In each case the pronoun ‘it’ refers to something different: in (1) the monkey, in (2) the banana and in (3) the abstract notion of time. If we translate the above sentences into German, then anaphor resolution is inevitable (i.e. compulsory) since the pronouns take the gender of their antecedents and the German words *Affe* (masculine, ‘monkey’), *Banana* (feminine, ‘banana’) and *es* (neuter, ‘it’ for expressions of time) all have different genders.

Consider now the following translations from English into Korean of the above sentences (1) to (3); a literal English paraphrase of each translation is included for clarification..

(1’) pay.ko.pha.se wuen.swung.i.nun pa.na.na.lul mek.ess.ta
hungry-causal monkey-nominative banana-accusative eat-past,declarative
(2’) ik.e.se wuen.swung.i.nun pa.na.na.lul mek.ess.ta
ripe-causal monkey-nominative banana-accusative eat-past,declarative
(3’) hi.tha.im.i.e.se wuen.swung.i.nun pa.na.na.lul mek.ess.ta
tea time-causal monkey-nominative banana-accusative eat-past,declarative

Note that in the above Korean translations there are no pronouns. These examples might seem to suggest that we could translate from English to Korean, bypassing altogether the tough problem of anaphora resolution. However, such a conclusion would be misleading.

The assumption that anaphoric expressions in the source language can be easily mapped to the corresponding anaphors in the target language, or that they can in many cases be simply ignored in the transfer phase, is unfounded. It is not hard to find English sentences for which anaphor resolution is necessary in order to obtain their correct translation into Korean. Consider the sentences:

(4a) Although programmers usually write good programs, they may still make a mistake.
(4b) Although programs are usually written by good programmers, they may still contain mistakes.

In Korean, there are two types of pronouns corresponding to ‘they’: one for human beings and another for non-human beings. In order to assign the proper Korean pronouns to the English pronoun ‘they’, an MT system must be able to resolve ‘they’ by choosing one of the two possible antecedents for it, viz. ‘programmers’ and ‘programs’.

Anaphor resolution becomes an even more serious business if we are aiming to achieve high-quality translation. The translation of (4a) and (4b) into Korean, with the successful assignment of pronouns, may still sound awkward to Koreans. The reason is that, in Korean, it is stylistically more natural not to mention explicitly anaphors in subordinate clauses that are coreferential with nominal expressions in the main clause. This is somewhat similar to English participle constructions whose subject is ‘understood’. The preferred Korean translation of (4b) may be paraphrased (in English) as follows:

(5) Being usually written by good programmers, programs may still contain mistakes.

Thus the best translation of (4a) and (4b) is arrived at by avoiding altogether the use of overt pronouns. This being so, anaphor resolution is crucial in English-to-Korean MT: clearly the system must first resolve the pronominal form ‘they’ in order to be able to replace it by an appropriate nominal expression.

‘Optional’ anaphor resolution means preserving anaphoric ambiguity in case no anaphor resolution is undertaken. At first sight it may seem that carrying over ambiguities in translation is even more ‘authentic’ from the point of view of having a mirror translation of the source text. In effect, however, not resolving anaphoric ambiguity means that during the translation process the text is not fully understood. After all it should not be overlooked that the goal of the analysis is to produce an unambiguous intermediate representation (Isabelle and Bourbeau 1985). Moreover, a system relying heavily on the ‘ambiguity preservation’ method offers no computational advantage when ambiguity-preserving situations have to be identified dynamically; it is also extremely vulnerable in situations where the lexicon is growing while the system is in use or when additional languages have to be introduced (Nirenburg et al.1992). Every new word sense added to the lexicon carries the potential of ruining the possibility of retaining ambiguity in translation for all previous entries. This means that extra attention must be paid to the maintenance of the lexicons.

3.3 *Further Problems: Gender and Number Discrepancies*

It is worthwhile mentioning that MT adds a further dimension to anaphor resolution. In addition to the discrepancies in target language anaphor selection described in Section 3.1, the additional complexity is due to gender discrepancies across languages, to number discrepancies of words denoting the same concept, and to discrepancies in gender inheritance of possessive pronouns. These issues are discussed below.

First, there are often gender discrepancies across different languages. In French gender is assigned arbitrarily (although not as arbitrarily as in German), so that translation problems occur where the gender of the source does not carry over to the target language.

French: Je prends le livre. Il est bon.
German: Ich nehme das Buch. Es ist gut.

Since English has natural gender, this may create problems in translation into languages that assign gender grammatically.

English: I take the 9 o’clock train. It is usually late.
German: Ich nehme den 9-Uhr-Zug. Er ist gewöhnlich zu spät.

Second, there may be discrepancies between a noun which is referred to by a singular pronoun in one of the languages and by a plural noun in the other.

German: Die Informationen sind falsch. Sie verursachten aber eine Panik
English: The information is wrong. But it (not they) caused a panic.
English: The police are coming. They are just in time.
German: Die Polizei kommt gerade. Sie kommt gerade noch rechtzeitig.

Third, there may be problems involving discrepancies of gender inheritance across languages; examples of this are found between French and German:

French: Jacques a détruit sa voiture.
German: Jacques hat sein (*not:* ihr) Auto zerstört.

3.4 *Translation or Reconstruction?*

It is valid to ask the question whether it is generally worthwhile aiming at the translation of pronouns at all. Is it not rather irrelevant how pronouns translate? This consideration could arise from their status as units that are not autonomous in their meaning/function but dependent on other units in the text. In view of this a more natural way to treat pronouns in MT would be based on the following principles.

First, analysis has to determine the reference structure of the source text; in other words, co-reference/co-specification relationships between anaphors and antecedents have to be determined.

The second principle is that this is the only information that is conveyed to the target language generator.

Third, the target language generator generates the appropriate target language surface expression as a function of the target equivalent of the source antecedent according to the rules of this language.

Such a treatment is, in general, rather a reconstruction of referential expressions (in this case pronouns) than a translation of such expressions. A similar approach has been adopted in by Mitkov et al. (1995; see also section 4.3 below).

4. *Anaphor Resolution in Machine Translation: Research to Date*

Owing to the fact that the majority of MT handles one-(simple)-sentence input, not an extensive amount of work has been reported on anaphor resolution in MT. In the following we will briefly outline all work published so far in the field.

4.2 *An English-to-Japanese MT Program*

Wada (1990) reports on an implementation of discourse representation theory in an LFG-based English-to-Japanese MT program. His anaphor resolution mechanism consists of three functional units:

- construction of the DRS (discourse representation structure)
- storage of the salient element
- search for the antecedent

The first module constructs the DRS's compositionally, while the second module stores the most salient, focused element in the current discourse for processing the next sentence. In order to find the most salient NP, this module sets three kinds of filters (grammatical function, use of pronominal and syntactic constructions) and checks all the NP's appearing in the current sentence with respect to the three filters.

The third module incorporates three functions. The first function is 'search', which searches for an antecedent by testing the accessibility on the DRS and syntactic constraints such as gender and number as well as binding features. If the search in the DRS fails, a further search in the 'storage of the salient element' module is performed. According to the result of 'search', three classes of pronominals are distinguished: (1) an antecedent is found in the current DRS; (2) an antecedent is not found in the current DRS, but is controlled by a discourse focus; and (3) an antecedent is not found either in the DRS or as the salient element.

The second function sets an unique anaphoric index in case (1), whereas the third function assigns 0 to the pronominal in (2) and undertakes a default word-for-word translation in (3).

4.2 *English-to-Chinese MT*

Chen (1992) describes the interpretation of overt and zero anaphors in English-to-Chinese and Chinese-to-English MT and justifies the importance of anaphor resolution. He outlines a 'reflexive resolution algorithm' (based on c-command constraints and some semantic features), a 'pronoun resolution algorithm' (based on c-command constraints and some simple semantic features) for overt anaphors in English and proposes an algorithm for the use of zero anaphors in Chinese. In addition, with a view to applying the results in Machine Translation transfer, he investigates the statistical distribution of anaphors (zero and pronominal) and their antecedents in both languages.

4.3 *Resolution of Japanese Zero Pronouns*

Nakaiwa reports in various papers on the resolution of Japanese zero pronouns in Japanese-to-English MT (see Nakaiwa and Ikehara 1995, Nakaiwa et al. 1994, Nakaiwa et al. 1995, Nakaiwa and Ikehara 1992). He uses semantic and pragmatic constraints such as conjunctions, verbal semantic attributes and modal expressions to determine intrasentential antecedents of Japanese zero anaphors (Nakaiwa and Ikehara 1995). His tests suggest that such antecedents can be resolved correctly in 98% of cases.

4.4 *Portuguese-to-English MT*

H. Saggion and A. Carvalho (1994) describe pronoun resolution in a Portuguese-to-English MT system which translates scientific abstracts. They use syntactic agreement and c-command rules to solve intrasentential anaphors and syntactic analysis of the immediately preceding sentence (Reinhard 1983) plus a history list of previous antecedents to solve intersentential anaphors (Allen 1987).

4.5 *An English-to-German MT System (KIT-FAST)*

Hauenschild, Mahr, Preuß et al. (1993) and Preuß, Schmitz, Hauenschild et al. (1994) describe work on anaphor resolution in the English-to-German MT system KIT-FAST, developed at the University of Berlin. Their approach uses two levels of text representation — structural and referential — as well as various 'anaphor resolution factors', viz. proximity, binding, themehood, parallelism and conceptual consistency.

The structural text representation includes information about functor-argument relations (e.g. between nouns, verbs and adjectives and their complements), semantic roles of arguments (agent, affected, attribuand, associated, location, aim), thematic structure of sentences, semantic features that express local or temporal conceptualisation, and grammar and anaphoric relations represented by co-indexation.

The referential text representation contains aspects of the text content, namely the discourse referents and the conceptual relations between them. Co-reference relations are represented by one discourse referent, and every relation that holds for an antecedent is also valid for an anaphor that refers to it. The referential information is represented in a terminological logic with the help of the knowledge representation system BACK (Weisweber 1994).

As far as the factors for anaphor resolution are concerned, proximity accounts for the fact that personal pronouns are most likely to have their antecedents in the superordinate or preceding sentence, while possessive pronouns are more likely to refer to a noun occurring in the same sentence. Binding excludes as antecedent all sisters of a pronominal argument in the structural text representation, whereas themehood defines structurally prominent constituents such as, for example, the subject or the topic of a sentence. Since the factors refer to the structural and referential representations, they have no access to purely syntactic information such as the subject. For this reason a notion of semantic subject is defined on the basis of the structural text representation and given preference. Moreover, the most topical candidate is preferred, whereas free adjuncts are considered as bad antecedent candidates for possessive and personal pronouns.

The parallelism factor is expressed in agreement and identity of roles; the conceptual consistency factor checks for compatibility between antecedents and anaphors.

4.6 *An English-to-Korean MT System (MATES)*

Mitkov, Kim, Lee and Choi (1994) describe an extension to the English-to-Korean MT MATES system which allows it to resolve pronominal anaphors. MATES is a transfer-based system, which does an English sentence analysis, transforms the result (parse tree) into an intermediate representation, and then transforms it into a Korean syntactic structure to construct a Korean sentence.

The paper argues that pronouns cannot be ignored in an English-to-Korean translation and that one cannot bypass the task of resolving anaphoric references (however deceptive it may be in some cases) if aiming at good quality and natural translation. The work explores this problem in detail and suggests practical rules for some tough cases of English-to-Korean anaphor translation, including different types of complex sentences, general quantifiers, human ‘it’, non-human ‘she’, and inanimate ‘they’. On the basis of the results reported, the authors propose general lexical transfer rules for English-to-Korean anaphor translation and outline an anaphor resolution model for the English-to-Korean MATES system.

The anaphor resolution model proposed is a simplified version of the model proposed by one of the authors in Mitkov (1994). Full implementation of the latter, including a centre tracking inference engine, appears to be too costly for the immediate goals of the operational English-to-Korean MT system.

4.7 *An Extension of CAT2*

R. Mitkov, S.K.Choi and R. Sharp (1995) describe an extension of the unification-based MT system CAT2. The latter was developed at IAI, Saarbrücken, as a sideline implementation of the Eurotra Project. The translation strategy is based on tree-to-tree transduction, where an initial syntactico-semantic tree is parsed, then transduced to an abstract representation (‘interface structure’), designed for simple transfer to a target language interface structure. This is then transduced to a syntactico-semantic tree in the target language, whose yield provides the actual translated text.

The current version of the anaphor resolution model implemented in CAT2 is based exclusively on syntactic and semantic constraints and preferences. Syntax constraints include the almost obligatory agreement of the pronoun and its antecedent in number, person and gender as well as c-command constraints as formulated in Ingria and Stallard (1989). Syntactic preferences used are syntactic parallelism and topicalisation. Syntactic parallelism gives preference to antecedents with the same syntactic role as the pronoun, whereas topicalised structures are searched first for possible anaphoric referents.

Semantic constraints and preferences include verb case role constraints, semantic networks constraints (semantic consistency) and semantic parallelism preference. Verb case role constraints stipulate that if filled by the anaphor, the verb case role constraints must be satisfied also by the antecedent of the anaphor. Semantic networks indicate the possible links between concepts as well as concepts and their attributes; semantic parallelism preference favours antecedents which have the same semantic role as the pronoun.

The project concentrates primarily on intersentential anaphor resolution, but since CAT2, like most other MT systems, only operates on single sentences, we simulate the intersententiality by conjoining sentences, as in the following:

The decision was adopted by the council; it published it.

The task is to resolve the pronominal references in the second ‘sentence’, with the antecedents in the first. The implementation, which successfully handles pronominal resolution in the context of English-to-German translation, can be carried over to multiple-sentence input.

The noun phrase features relevant for anaphor resolution are collected in the complex feature ‘anaph’, consisting of two additional features, ‘ref’ (referential features) and ‘type’. The referential features include the noun’s agreement features (person, number, gender), lexical semantic features (e.g. animacy), and its referential index; the type feature indicates whether the noun is a pronoun or not:

anaph={ref={agr=A,sem=S, index=I},type=T}

All non-pronominal nouns receive a unique index by a special index generator within CAT2; each pronoun takes its index value by way of unification with its antecedent, as outlined below.

Anaphor resolution in CAT2 occurs in two main steps. First, all ‘anaph’ features within a sentence are collected in a ‘cand’ (candidates) feature and percolated to the S node, so that anaphor resolution between sentences can take place locally under the topmost node (intrasentential anaphor resolution will already have occurred). Then, for each pronoun in the second sentence, its ‘ref’ feature is resolved with those of the antecedents in the first sentence. Backtracking provides for all combinations, under the condition that the ‘ref’ features agree, i.e. unify.

Illustrations of correct pronoun resolution are the translations by CAT2 of the following sentences:

English: The council adopted the decisions; the commission published them.
German: Der Rat verabschiedete die Beschlüsse; die Kommission veröffentlichte sie.
English: The council adopted the law; it published it.
German: Der Rat verabschiedete das Gesetz; er veröffentlichte es.

English: The commission published the law; it was adopted by the council
German: Die Kommission veröffentlichte das Gesetz; es wurde von dem Rat verabschiedet.

English: The decision was adopted by the council; it published it.
German: Der Beschluss wurde von dem Rat verabschiedet; er veröffentlichte ihn.

References

Carbonell, J. and Brown, R. (1988) 'Anaphora Resolution: a Multi-strategy Approach', *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88)*, Budapest, August, 1988

Carter, D. M. (1987) *Interpreting Anaphora in Natural Language Texts*, Chichester: Ellis Horwood

Chen, H. H. (1992) 'The Transfer of Anaphors in Translation', *Literary and Linguistic Computing,* Vol. 7, No. 4

Connoly, D., Burger, J., and Day, D. (1994) 'A Machine-learning Approach to Anaphoric Reference', *Proceedings of the International Conference on 'New Methods in Language Processing'*, UMIST, Manchester, 14–16 September 1994

Dagan, I. and Itai, A. (1990) 'Automatic Processing of Large Corpora for the Resolution of Anaphora References', *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90),* Helsinki, 1990

Halliday, M. and Hasan, R. (1976) *Cohesion in English*, Longman English Language Series 9, London: Longman

Hauenschild, C., Mahr, B., Preuß, S. et al. (1993) *Anapherninterpretation in der maschinellen Übersetzung*, Final Technical Report KIT-108, Technical University of Berlin

Hirst, G. (1981) *Anaphora in Natural Language Understanding*, Berlin: Springer Verlag

Hobbs, J. (1978) 'Resolving Pronoun References', *Lingua*, Vol. 44

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

Ingria, R. and Stallard, D. (1989) 'A Computational Mechanism for Pronominal Reference', *Proceedings of the 27th Annual Meeting of the ACL*, Vancouver, British Columbia, 26–29 June 1989

Isabelle, P. and Bourbeau, L. (1985) 'TAUM-AVIATION: its Technical Features and Some Experimental Results', *Computational Linguistics*, Vol. 11

Keenan, E. (ed.) (1975) *The Formal Semantics of Natural Language,* Cambridge University Press

Kennedy, C. and Boguraev, B. (1996) 'Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser', *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, Copenhagen, Denmark, 5–9 August 1996

Mitkov, R. (1994) 'An integrated model for anaphora resolution', *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94),* Kyoto, Japan, 5–9 August 1994

Mitkov, R. (1995) 'An Uncertainty Reasoning Approach to Anaphora Resolution', *Proceedings of the Natural Language Pacific Rim Symposium*, Seoul, Korea, 4–7 December 1995

Mitkov, R. (1996a) 'Anaphor Resolution in Natural Language Processing and Machine Translation', *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution*, Lancaster, UK,17–19 July 1996 (keynote speech)

Mitkov, R. (1996b) ‘Pronoun Resolution: the Practical Alternative’, *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution*, Lancaster University, UK, 17–19 July 1996

Mitkov, R. and Schmidt, P. (1996) ‘On the Complexity of Anaphor Resolution’, *International Conference on Mathematical Linguistics (IMCL’96),* Tarragona, Spain, 2–4 May 1996

Mitkov, R., Choi, S.K., and Sharp, R. (1995) ‘Anaphora Resolution in Machine Translation’, *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 5–7 July 1995

Mitkov, R., Kim, H.K., Lee, H.K. and Choi, K.S. (1994) ‘The Lexical Transfer of Pronominal Anaphors in Machine Translation: the English-to-Korean Case’, *Proceedings of the SEPLN’94 Conference*, Cordoba, Spain, 20–22 July 1994

Nakaiwa, H. and Ikehara, S. (1992) ‘Zero Pronoun Resolution in a Japanese-to-English Machine Translation System by Using Verbal Semantic Attributes’, *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP’92)*, Trento, Italy, 1992

Nakaiwa, H., Yokoo, A. and Ikehara, S. (1994) ‘A System of Verbal Semantic Attributes Focused on the Syntactic Correspondence between Japanese and English’, *Proceedings of the 15th International Conference on Computational Linguistics (COLING’94),* Kyoto, Japan, 1994

Nakaiwa, H., Shirai, S., Ikehara, S. and Kawaoka, T. (1995) ‘Extrasentential Resolution of Japanese Zero Pronouns Using Semantic and Pragmatic Constraints’, *Proceedings of the AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*

Nakaiwa, H. and Ikehara, S. (1995) ‘Intrasentential Resolution of Japanese Zero Pronouns in a Machine Translation System Using Semantic and Pragmatic Constraints’, *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI’95),* Leuven, Belgium, 1995

Nasukawa. T. (1994) ‘Robust Method of Pronoun Resolution Using Full-text Information’, *Proceedings of the 15th International Conference on Computational Linguistics (COLING’94)*, Kyoto, Japan, 5–9 August 1994

Nirenburg S., Carbonell J., Tomita M., and Goodman K. (1992) *Machine Translation: a Knowledge-based Approach*, San Francisco: Morgan Kaufmann Publishers

Ramm, W. (ed.) (1994) *Studies in Machine Translation and Natural Language Processing,* Volume 6, *Text and Content in Machine Translation: Aspects of Discourse Representation and Discourse Processing*, Office for Official Publications of the European Community, Luxembourg

Rico Pérez, C. (1994b) ‘Resolución de la anáfora discursiva mediante una estrategia de inspiración vectoral’, *Proceedings of the SEPLN’94 Conference*, Cordoba, Spain, 20–22 July 1994

Preuß S., Schmitz B., Hauenschild C., and Umbach C. (1994) ‘Anaphora Resolution in Machine Translation’, in Ramm (1994)

Rich, E. and LuperFoy, S. (1988) ‘An Architecture for Anaphora Resolution’, *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, 9–12 February 1988

Saggion, H. and Carvalho, Ar. (1994) ‘Anaphora Resolution in a Machine Translation System’, *Proceedings of the International Conference ‘Machine Translation, Ten Years On’,* Cranfield, UK, 12–14 November, 1994

Sidner, C.L. (1979) *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*, M.I.T. Artificial Intelligence Laboratory Technical Report No. 537

Wada, H. (1990) ‘Discourse Processing in MT: Problems in Pronominal Translation’, *Proceedings of the 13th International Conference on Computational Linguistics (COLING’90)*, Helsinki

Webber, B.L. (1978) *A Formal Approach to Discourse Anaphora*, New York: Garland

Weisweber, W. (1994) ‘The Experimental MT System of the Project KIT-FAST — Complementary Research for Eurotra-D’, *Proceedings of the International Conference ‘Machine Translation, Ten Years On’,* Cranfield, UK, 12–14 November, 1994

Wilks, Y. A. (1975) ‘Preference Semantics’, in Keenan (1975)

Ruslan Mitkov may be contacted at the School of Languages and European Studies, University of Wolverhampton, Stafford Street, Wolverhampton WV1 1SB. E-mail: R.Mitkov@wlv.ac.uk

New Testament Greek:
Towards a Two-level Description

by

Gary Stringer

Pallas (Computing in the Arts)
University of Exeter

Introduction

In recent years, there has been a growing recognition of the value of using a morphological parser as part of any natural language processing system. This is true even of English-based systems. English is generally regarded to be poor in its inflectional system; such a parser is almost essential in languages with more complex inflections. With a language such as Greek, where the affixes of a word carry much of the syntax of a phrase, and where a high proportion of the non-lexical content of a phrase undergoes inflection, a morphological processor does a great deal of the work constituting a full syntactic analysis.

This paper describes the design and implementation of a system to morphologically analyse New Testament Greek. The system is based upon the two-level model, first described by Koskenniemi (1983) and implemented in the form of PC-KIMMO (Antworth 1990), and adds a simple post-processor to provide further disambiguation and enhanced readability of output. Some of the problems encountered in applying the model to Hellenistic Greek are discussed, together with initial solutions and a description of future work.

Textual Analysis

In studying Hellenistic Greek (in this case the text of the New Testament), the focus is upon text preserved by ancient documents, rather than living, spoken language; the phonology of the language is interesting but generally less relevant to most students. However, a morphological processor will give us an immense advantage when studying texts, allowing us to search intelligently for words as well as providing feedback on the grammatical properties of words.

The most ancient documents preserving the text of the New Testament are written without punctuation, accents or capitalisation, and this is the form taken by the parsing mechanism as its input. Additionally, the original text is written in *scriptio continuo* (i.e. without word divisions). The parser currently requires the insertion of word divisions, though the problem of determining these divisions holds sufficient interest to be considered in future work.

The initial aim of the work was to provide a parsing mechanism which would allow simpler access to dictionary headwords, by supplying the root of any given inflected form, since inflectional prefixes make dictionary access difficult, especially for beginners. This was then extended to provide limited morphological information, which could be generated with little extra effort by the parser. The post-processor was added later, in order to translate the raw parsed analysis into something more readable.

The main parsing mechanism employs the PC-KIMMO implementation (version 1.0; Antworth 1990), and takes the shape of grammars and lexicons in this form. This program is based on the work of Koskenniemi (1983), Karttunen (1983) and others, and employs the well-known two-level model to describe morphological processes. Post-processing was initially performed by hand and later generated using programs written in Turbo Pascal.

The complexity of the Greek morphological system, and constraints on time, have necessitated restricting the scope of the parser to a subset of verbal inflection; though the end result is not a comprehensive working system, it provides the basis for further work in attaining this goal.

The Structure of the Lexicon

The morphology of the Greek language is largely concatenative in nature, and this in most cases is restricted to the simple addition of affixes to a single stem. Occasionally a word form will take one of a range of alternate stems, usually closely related though sometimes quite different in form. In a few cases though, the concatenative process involves some more complex change, usually modifying the area in which the morphemes collide. There are also some reduplicative phenomena, which can be seen to introduce some problems for the two-level model.

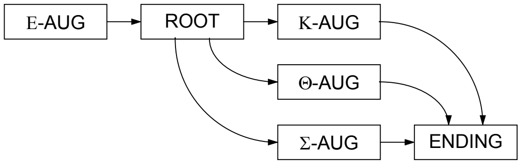


Figure 1: Lexicon Structure

In accordance with the PC-KIMMO implementation, the morpheme lexicon is structured as a set of continuation classes, a simplified map of which is shown in Figure 1. This structure implies a simple left-to-right nature to any affixation, which is a major defect in currently available implementations, and causes two problems for the parser. Firstly, some of the processes in the morphology operate on a right-to-left basis, notably the reduplicative prefix which forms, for example, the perfect indicative active (e.g. lu-w → le-lu-ka). This can be worked around, using the two-level rules, as described later.

Secondly, the implication is that the nature of each morpheme may only depend upon that of its immediate predecessor; there is no ‘memory’ or ‘look-ahead’ which would provide more efficient parsing by reducing any unnecessary backtracking. We can see this again in the form of le-lu-ka where the tense is determined by the combination of both the prefix (le-) and the suffix (-ka) which surround the stem (lu). This problem is overcome by employing a post-processor to link the disconnected morphemes once the main parse is completed.

Two-level Rules

Whilst the continuation classes of the morpheme lexicon are capable of handling the purely agglutinative aspects of the morphology, changes often occur at the boundaries of affixation which cannot be explained by simply adding morphemes to one another. The two-level rules define the mechanisms of these changes, and describe the change between the surface form of a word and its lexical form, broken into morphemes. To do this, it uses computationally formal devices known as finite-state transducer networks, which indicate the possible mappings between lexical and surface forms with a type of flow chart.

As a typical example of a morphological process implemented as a transducer, we can examine the contraction over a morpheme boundary where the preceding morpheme (usually a verbal stem) ends in an epsilon (e). This contraction is exhibited in several forms, for example in the paradigm of φιλεω (‘to love’):

file-ete	fileite	e + e become ei	e+e→ei
file-omen	filoàmen	e + o become ou	e+o→ou
file-w	filî	e + long vowel deletes e	e+w→w
file-h	filĺ	e + long vowel deletes e	e+h→h
file-ei	filei	e + diphthong deletes e	e+ei→ei
file-ousi	filousi	e + diphthong deletes e	e+ou→ou

From this data, the correspondence between the lexical form and the surface form is apparent, and this is compiled into a transducer mechanism (Figure 2), which will perform the translation in the parser itself. Note the presence of the morpheme boundary symbol (+) in the network, since this change may only occur over the concatenation of two morphemes. It should be possible to trace a route through the transducer network for each pair of lexical/surface forms in the table above, describing the correspondence between them.

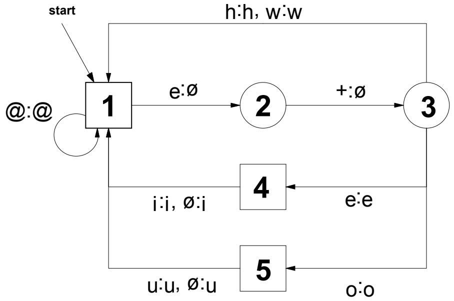


Figure 2: Finite-state transducer for epsilon contraction

In the two-level model, a number of these transducers operate in parallel, in order to reduce the computational complexity required by multiple successive applications of morphological rules, as in traditional generative morphology. This gives a very efficient parser, which can operate in real-time, opening up uses in many areas of text retrieval and analysis. Further explanation of the operation of the two-level rules can be found in both Antworth (1990) and Stringer (1996).

Again, these transducers operate in a left-to-right sequence, which favours a suffixing language to some extent and creates excessive backtracking where right-to-left processes occur. This is particularly problematic in the case of prefix reduplication, where the morphological change is wholly dependent upon the morpheme following the prefix (as in le-luka, ge-grafa, pe-futeuka, etc.). Here the process is described using a separate transducer for each consonant which could possibly be reduplicated (Figure 3 shows one such transducer, for l).

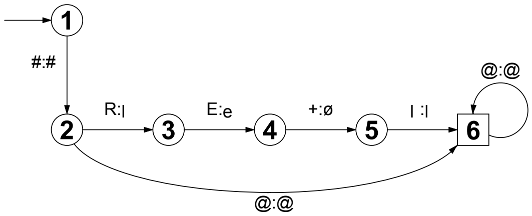


Figure 3: Finite-state transducer for lambda reduplication

This solution is not the most elegant, but it provides a practical method of dealing with reduplication within the two-level model; and though it creates some otherwise avoidable backtracking, this does not significantly impact performance.

Conclusions

The production of a morphological parser for Hellenistic Greek is a large project, simply because the morphology is complex and often irregular. While this study has concentrated on a relatively small subset of Hellenistic Greek morphology, the problems and solutions discovered are more generally applicable to the language as a whole. One of the initial aims of the work was to evaluate the feasibility of a morphological parser, and through concentrating work on the perceived difficulties of the language, it would seem both possible and desirable to construct a more fully comprehensive system on the basis of these findings.

A major area which would require some further improvement in a system with wider coverage, and which has not so far been discussed, is the method of defining exceptions to the morphological rules described within the system. The PC-KIMMO implementation has little scope for this, which in such an irregular language is a major drawback. The simplest solution, that of providing explicitly parsed forms and a ranking system to override superfluous parses, could be incorporated into the post-processor. However, it may also be helpful to advanced users not to filter out ‘bad’ parses, as these may indicate the productive use of language or unusual irregular forms, especially where the texts examined are outside the well-researched texts of the New Testament canon.

Though the implementation of PC-KIMMO used has proved adequate for designing the parser detailed above, recent improvements in the latest version of this program have suggested more elegant solutions to some of the problems. In particular, the use of a unification grammar to replace the simple continuation classes within the lexicon will prove useful in dealing with the problem of disconnected morphemes; the need for a post-processor is removed. Future work will concentrate on this enhanced implementation, and on expanding the lexicon to handle less regular forms.

References

Antworth, E.L. (1990) ‘PC-KIMMO: A Two-level Processor for Morphological Analysis’, Dallas, TX: Summer Institute of Linguistics

Karttunen, L. (1983) ‘KIMMO: a General Morphological Processor’, Texas Linguistic Forum 22:163–186

Koskenniemi, K. (1983) ‘Two-level Morphology for Morphological Analysis, *Proceedings of IJCAI-83*: 683–85

Stringer, G.B. (1996) *Two-level Computational Morphology and its Application to New Testament Greek*, unpublished MA dissertation, University of Exeter

Gary Stringer may be contacted at Pallas (Computing in the Arts), Queen's Building, University of Exeter, EX4 4QH, e-mail: G.B.Stringer@exeter.ac.uk.

The Power Translator: an Evaluation of a PC-based MT System

by

Derek Lewis

Department of German

University of Exeter

The Power Translator (PT), marketed by Globalink, first appeared in the UK in the late 1980s. A PC-based MT system, the PT is described in the User's Guide as a package for producing 'draft translations that you can edit into final form'. According to the suppliers, the program 'translates quickly, taking only seconds for a short document and minutes for a long one'. The system is available for a number of European languages; at the time of writing each installation cost around £200 (for each stand-alone PC) and provided bi-directional translation for a single language pair.

The following account is based on experiences of using the PT in an educational environment. Undergraduate students in their final year of a Modern Languages degree at the University of Exeter research the system and write up an evaluation of its performance as part of a course in Natural Language Processing Applications. The language directions are English-German and German-English. Of course, it is one thing to use an MT system over a prolonged period in a commercial or a professional environment; it is quite another to introduce it to undergraduates during a single 10-week university term (which is all the time that the course allows). Having said that, the course has quite specific aims: to introduce language students to an important technological tool in translation and to enable them to assess its potential and its limitations; students are also expected to investigate and evaluate the tool independently, working with the technical documentation in German that is supplied with the system. In some respects the student is in the position of a potential purchaser who is expected to assess the potential usefulness of an MT system; he might, for instance, be working on behalf of a translation bureau or of a company that wants to know what level of automation it can introduce into the processing of its foreign language documentation. The following paper will describe the main features of the PT system and evaluate aspects of its performance; the benefits of introducing MT into the curriculum will be the object of a future paper.

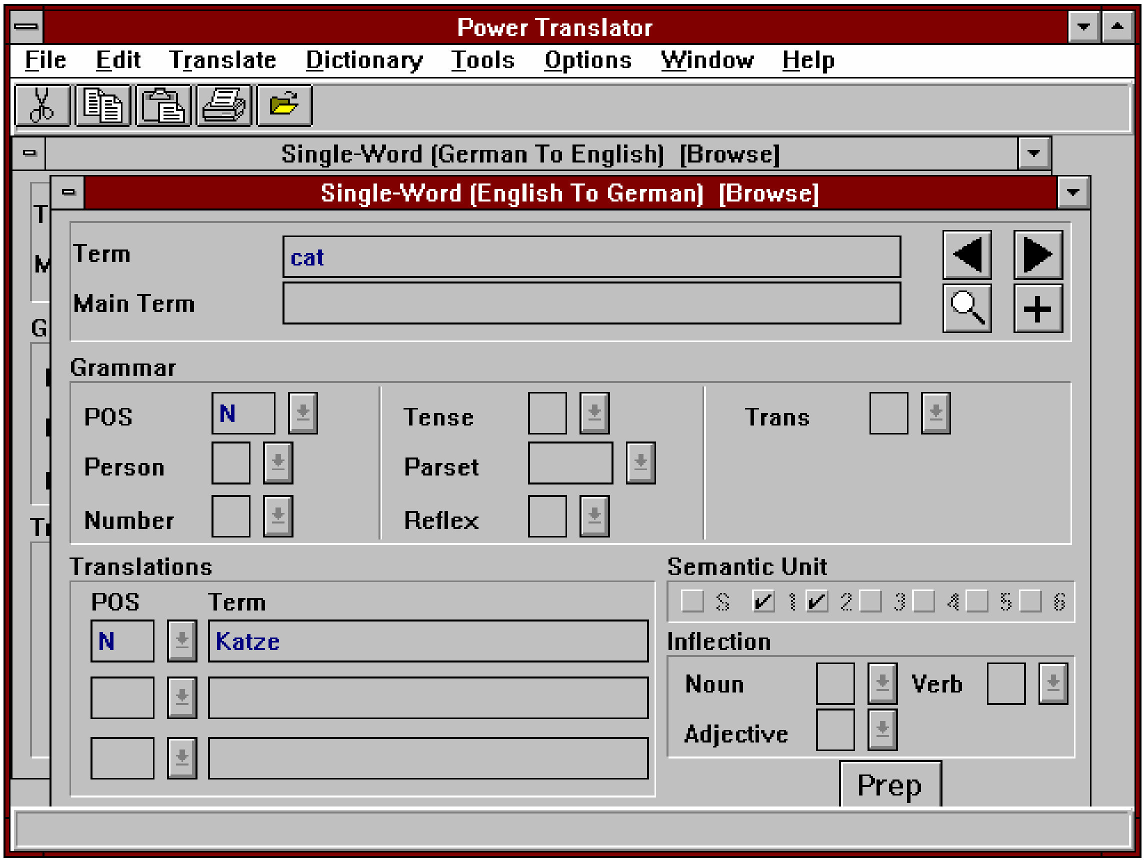
The basic specifications for the Windows 2.0 version of the PT are an 80286 or 386 PC with 2 Megabytes of RAM and 23.5 Megabytes hard disk space (of these 20 MB are taken up by the core dictionaries and 3.5 MB by the processing programs. There is a DOS version of the PT , which is slightly cheaper, and which was the initial basis for my own work with the system at Exeter. Unfortunately, the DOS version for English and German proved to have a bug that ensured that the system ignored any new term and its translation that the user entered in the dictionaries; the program thus translated only those words and phrases in the core dictionaries that were supplied by the manufacturers. The most interesting thing about this experience is that the existence of the bug came as something of a surprise to the suppliers; few other users had reported the same problem. If we assume that the bug was restricted to the English-German version, it raises the question of whether most purchasers do in fact attempt to customise the system themselves. Do they buy and use the program as an off-the-shelf ready-to-run MT package? If so, does it fulfill their needs? Although unable to answer these questions, I can say that, once they were made aware of the bug, the suppliers responded swiftly and positively. They provided promptly and at no extra charge a bug-free version for Windows 3.1. They have also been extremely helpful in approving limited copying of the system's documentation for educational purposes.

The PT's suppliers indicate that over 15,000 copies of the system have been purchased in the UK since it first appeared (around 1990). Most purchasers are business executives in companies of all sizes; translators and academics have not proved to be the main buyers. Once again, interesting questions arise from this. In particular: are the purchasers also the end-users? how extensively is the package ultimately used by translators who have been consulted by their managers? The pricing of the PT suggests that the suppliers are aiming to put MT on a similar level to other PC-software: relatively inexpensive and with a large user-base. There is no doubt that Globalink has achieved remarkable success in marketing low-cost MT. According to a recent report it has captured 90% of this market and enjoys a $17 million turnover. Furthermore, with its so-called 'Telegraph' software, it has concluded a contract with a large insurance company, the Commercial Union, to provide rough translations, especially of 'junk mail', between English and French: 'it's cheaper than paying a translator 25p a word' (*The Guardian*, 5.12.1996:24).

As an off-the-peg system the PT is straightforward to use. With a little knowledge of the standard graphical Windows environment, the user can translate a text almost immediately. The source text is simply opened as a file, and either all the text or a marked portion of it designated for translation. The source text can be typed in directly or it can be imported as a Windows text file (this is understood to be a file that has been saved in a Windows application but without the internal formatting codes that are specific to that application). The source text and its translation are both displayed on the screen, and any words or phrases that the program has been unable to translate (because it has not found them it its dictionaies) are clearly marked. The package has a number of facilities and can be customised in various ways, but we shall concentrate here on the dictionaries and their operation. Apart from modifying the input text, customising the dictionaries is the most important way in which the user can improve the translation process.

The core or default dictionaries that are supplied with the English-German system comprise about 2.5 Megabytes; they are known as the GENERAL dictionaries and contain about 250,000 entries. The filenames of the dictionaries indicate their various functions: there are separate morphological dictionaries for verbs and nouns in German (VGERMAN.MOR, NGERMAN.MOR), an English-German morphology dictionary (ENGGER.MOR), English-German synonym dictionaries (ENGGER.SYN, GERENG.SYN), semantic unit dictionaries (EGSUGEN.DIC, GESUGEN.DIC), and bilingual dictionaries for transfer (e.g. EGMICGEN.DIC, GEXTRGEN.DIC). When the user adds his own terms and translations, these are stored in separate USER dictionaries, whose size grows along with the number of entries. The core or general dictionaries cannot be modified. The difference between the user and the general dictionaries is invisible to the user, even when he is adding new terms or modifying existing ones. If the user changes a term in the general dictionary, the alteration is recorded in the user dictionary only, which is the first to be consulted by the system during translation; if a match is found, the program looks no further. Ready-made subject-specific dictionaries are available from the suppliers; these range in size from 10,000 to 180,000 entries (the latter for the 'Russian polytechnical dictionary').

To see what kind of information is stored in a typical lexical entry and how it is presented to the user, consider the entry for the English word *cat* in the English to German Single Word (SW) dictionary:


Power Translator
File Edit Translate Dictionary Tools Options Window Help
Single-Word (German To English) [Browse]
Single-Word (English To German) [Browse]
Grammar
Translations
Semantic Unit
Inflection
Prep

Term: cat

Main Term:

POS	N	Tense		Trans	
Person		Parset			
Number		Reflex			

POS	Term
N	Katze

□ S ☑ 1 ☑ 2 □ 3 □ 4 □ 5 □ 6

Noun		Verb	
Adjective			

Term: this is the field for the term.

Main Term: if the term is an inflected form, such as *shown*, then the main or stem form (*show*) is entered here. It should be noted that, for irregular forms, in particular nouns and verbs, each inflected item must be entered individually as a separate term.

The are two main fields or boxes for further information: a **Grammar** box for grammatical information about the term; and a **Translations** box for information about the translation(s) of the term.

In the **Grammar** box we have the following fields:

POS: this is the Part of Speech field (here N for Noun). Other parts of speech are adjective, adverb, conjunction, (also: subordinating), interjection, noun, proper noun, numeral, preposition, pronoun, and verb.

Person: this can be 1, 2, or 3, and is entered for inflected forms of pronouns and the inflected forms of irregular verbs (e.g. 1 for *we*, 3 for *flies*).

Number: this is entered for irregular noun and pronoun forms; e.g. S for the singular form of an irregular noun (*house*) or P for the plural form (*houses*, *children*).

Tense: this is used for irregular verbs, such as *speak*, 6 is entered for the past participle (*spoken*), 2 for the simple past (*spoke*).

ParSet (for **Paradigm Set**): the code entered here determines the translation of a term in a particular context; thus 300 is used for verbs and nouns followed by a to-clause (*an* *opportunity to learn German)* and tells the program to use the German *zu* instead of *um...zu* in translation (*eine Gelegenheit, Deutsch zu lernen*).

Trans (for **Transitive**): possible codes here are: 'None' for verbs that are transitive or that can be transitive or intransitive; I for instransitive-only verbs; D for ditransitive verbs (i.e. that take a direct and an indirect object); and C for adjectives that form comparative and superlative forms with -er and -est.

Reflex (for **Reflexive**): the code R is entered for an English verb whose German translation is reflexive (e.g. *refer — sich beziehen*; the *sich* is not entered in the German translation field below).

Object: codes are entered for German verbs and prepositions to indicate the case of the noun that follows the term (e.g. for *wegen*, the code is G for Genitive, since this preposition governs nouns in the genitive case; other cases are dative and accusative.

In the **Translations** box we have the following:

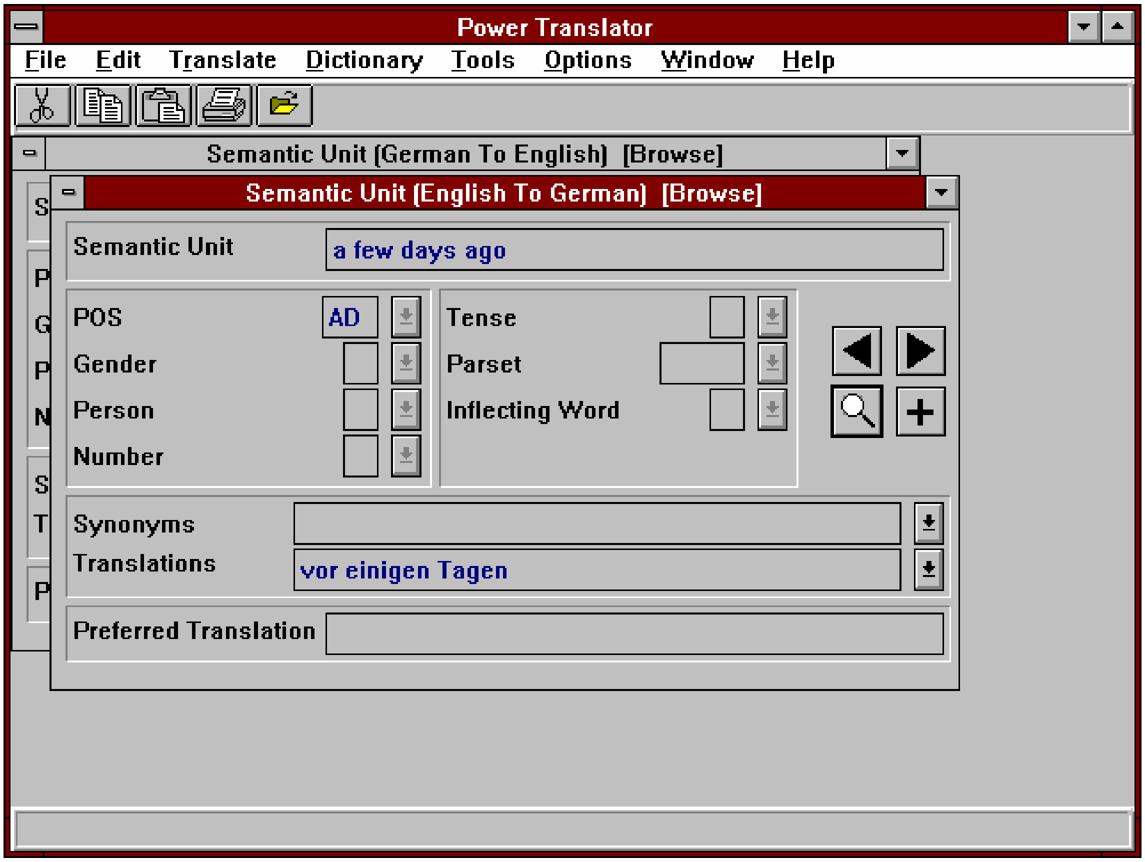
POS and **Term**: the translation and part of speech for the source language term are entered here; up to three translations may be entered as long as they are different parts of speech. Provision is made for English adjectives to be translated either as adjectives or as nouns in German. Thus *monetary* is translated as the adjective *geldlich* if the code A is entered along with the translation *geldlich*; but if *monetary* occurs in a noun phrase in the source text and the code Z has been entered for the translation term *Währung*, then *monetary* is translated as the noun *Währung* (*monetary policy* → *Währungspolitik*).

Semantic Unit: a Semantic Unit (SU) is a series of two or more words that is handled as a unit of meaning; the series cannot be translated word for word. Examples are *recommended price* (German: *Richtpreis), fresh water*, (*Süßwasser*), *table of contents* (*Inhaltsverzeichnis*), *a few days ago* (*vor einigen Tagen*). If the translation of a source language single-word term, such as the German *Inkubationszeit*, is a SU (English: *incubation period*), then the S box in the Semantic Unit field is marked. The SU field can also be used to mark the position of a single-word source language term within a semantic unit. Thus if the source term is *period*, then position 2 is marked, because the word *period* also occurs in the second position of a SU (viz. *incubation period*); by the same token the SU field in the dictionary entry for the single-word term *incubation* will be set to 1, because *incubation* is the first element of a SU. It should be clear from this that the individual words that occur in SU's are entered separately in the SW dictionaries: these entries contain pointers in the SU field that tell the program to consult the SU dictionary during translation.

Inflection: depending on the part of speech of the source language term (noun, verb, or adjective), one of these fields is set to indicate how the term is inflected.

The **Prep** (Preposition) box is set to determine the target language translation of a preposition that may follow a term. For example, the English verb *depend* is normally followed by the preposition *on*; the **Prep** box enables *on* to be translated by the German *von* (as in *abhängen von*). Of course, a term may take more than one preposition. The system allows for this, permitting a number of alternative prepositions and their translations to be entered (example: *think of* or *think about* which corresponds to German *denken an, denken über*). If the source language preposition is not to be translated at all, then the option 'None' is selected from the **Prep** box (example: English: *he showed the machine to the visitors*; German: *Er zeigte den Besuchern die Maschine*).

We shall now consider an example of an entry in the Semantic Unit (SU) dictionary. SU's can be up to six words long. A typical entry is that for the phrase *a few days ago*.



Semantic Unit: a few days ago

POS: AD

Translations: vor einigen Tagen

The **POS** field indicates the syntactic function of the phrase as a whole. If this is not appropriate (for instance, for SU's that are whole sentences), the code SU is entered. In this case the code is AD, i.e. for an adverbial.

The **Gender** field applies to SU's in which a head word is a noun.

The **Number** field indicates verbs that inflect for either singular or plural (but not both) in the SU.

The **Tense** field identifies the tense of an inflected verb in the SU.

The **Parset** code denotes some aspect of the syntactic behaviour of the SU. For instance, the code 200 is entered for a noun, verb, or adjective that takes a *that* complement.

The **Inflecting Word** field identifies the position of the most significant word in a SU that changes form according to how the SU is used in a text. For instance, in the SU *tear down* the inflecting word is *tear,* so the code is 1.

Inflection Code: this appears only for German SU's and is set for compound nouns.

Synonyms: items entered here will be given the same translation as the main SU itself. Thus if the SU in the German SU dictionary is the phrase *ab und zu* and the translation of this (in the English SU dictionary) is *now and again*, then entering *hin und wieder* in the synonym field will result in *hin und wieder* also being translated as *now and again*.

Translations: this field may contain a list of possible alternative translations for the SU. The first in the list is the primary translation and is the one that the system will normally choose. For instance, the German *aktive Schuld* can be either *account receivable* or *active debt*; if *account receivable* is the first in the list, then this is how the phrase will be translated. The facility allows the user to store alternative translations for, say, different texts. If he is not happy with the first translation in the list, then he simply designates another translation by selecting it from the **Preferred Translation** field that follows. This facility provides for some variation of translation for a SU, possibly to take account of text type. Alternative translations mut, however, be set manually before translation.

When adding a SU the user must bear a number of points in mind. First, each word in the SU must have a separate entry in the SW dictionary. Second, if you add a SU to the source language SU dictionary, then its translation must also be added to the target language SU dictionary (even if the translation is a single word and would normally appear only in the SW dictionary; however, the translation field for the target language SU must not be completed). Third, German SU's containing a verb must be entered twice: first with the verb (infinitive form) at the end and linked with its English translation, and second with the German verb at the beginning of the phrase; thus *aktiv dienen* and *dienen aktiv* (English: *be on active duty*). It is interesting to note that the core SW dictionaries do not contain single word entries alone. Phrases such as *go over*, *go back*, and *go into service* are all included as units in the SW dictionary for English; these entries do not include translations, only their inflection codes (here: V for verb). Although this is not clearly documented, it suggests that SU's have to be entered, not just as single words in the SW dictionary, but also as phrasal units.

It should be clear from the above that entering new single word terms is relatively straightforward. Much more complex is the procedure for entering SU's. This is not so much on account of the grammatical information required for the SU, but because it is necessary to provide the system with pointers to the SU in the other dictionaries. To see that this is so, consider the stages required to add the SU *drive away* and its German translation *wegfahren*.

1. Ensure that the words *drive* and *away* are already present in the English SW dictionary. Since they are common words, they should in this case be there already; otherwise the user must add them. For the word *drive*, mark the S box in the Semantic Unit box; mark also the box number 1. Do the same for *away*, but marking box number 2. The numbers indicate the position of each word in the phrase.

2. Check that *wegfahren* is entered in the German SW dictionary. Mark S in the Semantic Unit box to denote that the translation of *wegfahren* is a SU.

3. Add *wegfahren* to the German SU dictionary, even though it is not a SU. The Translations Field must not, however, be completed.

4. Add the phrase *drive away* to the English SU dictionary; in the Translations Field for this entry add also the German translation *wegfahren*.

5. Finally, add the phrase *drive away* to the English SW dictionary, marking the inflection (with V for verb), but entering nothing in the translation box.

The entry is now complete. On the whole, viewing and updating the dictionaries is greatly assisted by the Windows interface; more than one dictionary window can be open at the same time and is accessible via a mouse click. The procedure for entering linguistic data is menu-driven, which has both advantages and disadvantages. The advantage is that the system is easy to use and that the input is controlled. The disadvantage is that you may input only the information that is requested, which is not always the information that is needed for a good translation. Overall the Windows-based interface works well and is user-friendly. A somewhat tiresome drawback is that it is not possible to open more than one source file at a time for translation: when a file is closed the entire PT program is automatically shut down; it must be reloaded for a new file to be opened.

The documentation for the PT is well laid out and informative, although I have found the manual for the DOS version to be clearer and more explicit on one or two points than that for the Windows version (although there is considerable overlap). The authors go to considerable lengths to ensure that the user of MT does not have unrealistic expectations of MT: ‘No machine translation program can produce a perfect translation’ (p. 37). They emphasise the need for post-editing and devote several pages of the User’s Guide to listing common English/German translation errors (pp. 42–47). For an English source text these relate to the incorrect or non-idiomatic use of prepositions, missing or incorrect pronouns, incorrect resolution of noun/verb ambiguity, word order, misplaced commas, tense errors, incorrect passive structures, and mistranslation of case. A similarly comprehensive catalogue is provided for German source text. Specific guidance is given on how to pre-edit text in order to get the best results (for example: use short declarative sentences and avoid ambiguous words and complex punctuation).

I cannot claim to have carried out an exhaustive evaluation of the PT. What I can do is provide the preliminary results of how the system has translated part of an extensive suite of English test sentences compiled by Hewlett-Packard Laboratories. The whole suite, as it stood in 1987, comprised several hundred sentences. The compilers of the suite deliberately included malformed English sentences (for instance: *Did Abrams be interviewed by Browne?*). These I have discounted: the PT merely translates malformed English sentences into equally malformed German sentences, which serves only to show that the system returns a translation every time, despite the quality of the source text. An interesting project would be to construct a similar suite of test sentences for German (I am unaware of the existence of any such suite).

The following results (which are just a selection) speak for themselves. I have not attempted to quantify what proportion of sentences are ‘correctly’ translated. On the whole, the results are positive and demonstrate that the system can cope well with short, syntactically well-formed sentences. The PT prefaces items not found in its dictionaries with the string @@. I have noted mistranslations, or deviant forms, in bold print.

Selection of Results Obtained from Translating the HP Suite of Sentences for Evaluating NLP Systems

Structure: SUBJECT + VERB + ADJUNCT

English:	Abrams **works**.
German :	@@Abrams **Werke**.
English :	He works.
German :	Er arbeitet.
English :	Abrams works well.
German :	@@Abrams arbeitet gut.
English :	Abrams works every day.
German :	@@Abrams arbeitet jeden Tag.

English : Abrams works for me.
German : @@Abrams arbeitet für mich.

Note in the above sentences the syntactic ambiguity in the source sentences between the third person singular and plural noun forms. The PT has a better chance of parsing these forms correctly if the source sentence contains an additional element, such as a pronoun subject, a direct object, or an adverb, that provides more syntactic information. That this is not always the case, however, is clear from the following:

English : She **works** for whom?
German : Sie **Werke** für das?
English : She **works** for who?
German : Sie **Werke** für das?
English : She manages whom?
German : Verwaltet sie wen?

Structure: SUBJECT + VERB + DIRECT OBJECT

English : Abrams hired Brown.
German : Abrams **ein**stellte Braun.
English : He hired me.
German : Er **ein**stellte mich.

Note here how German separable verb prefixes remain attached to verb; this is a consistent failure of the PT, although it is easily corrected by post-editing.

English : The woman evaluated her own staff.
German : Die Frau bewertete ihren eigenen Stab.

Structure: SUBJECT + VERB + OBJECT + INDIRECT OBJECT

English : Abrams showed the office to the man.
German : Abrams zeigte das Büro **zu** dem Mann.
English : Abrams showed the man the office.
German : Abrams zeigte dem Mann das Büro.
English : Abrams showed the office to him.
German : Abrams zeigte das Büro **zu** ihm.
English : Abrams showed him the office.
German : Abrams zeigte **ihm** das Büro.

The last sentence in the above list is the result of modifying the dictionary entry for *show* (i.e. the parser is told to ignore the preposition *to* in translation.

Structure: VERB ‘TO BE’ + COMPLEMENT

English : It is true that Abrams hired Browne.
German : Es ist wahr, dass Abrams Browne einstellte.

Structure: ‘THERE ARE’ + PRESENT PARTICIPLE

English : There are programmers working for Devito.
German : Es gibt Programmierer, die für Devito arbeiten.

Structure: MODAL VERBS

English : Abrams may hire Browne.
German : Abrams mag Browne einstellen.

English : Abrams might hire Browne.
German : Abrams könnte Browne einstellen.
English : Abrams can hire Browne.
German : Abrams kann Browne einstellen.
English : Abrams could hire Browne.
German : Abrams könnte Browne einstellen.
English : Abrams shall hire Browne.
German : Abrams wird Browne einstellen.

Structure: MORE COMPLEX ENGLISH VERB FORMS

English : Abrams has hired Browne.
German : Abrams hat Browne eingestellt.
English : Abrams is hiring Browne.
German : Abrams **ein**stellt Browne.
English : Abrams is to hire Browne.
German : Abrams soll Browne einstellen.
English : Abrams **had better hire** Browne.
German : Abrams **hatte besseren Lohn** Browne.
English : Abrams could have hired Browne.
German : Abrams hätte Browne einstellen können.
English : Abrams could be hiring Browne.
German : Abrams könnte Browne einstellen.
English : Abrams could have been hiring Browne.
German : Abrams hätte Browne einstellen können.
English : Browne was interviewed by Abrams.
German : Browne wurde von Abrams interviewt.
English : Browne could be interviewed by Abrams.
German : Browne könnte von Abrams interviewt werden.
English : Browne has been interviewed by Abrams.
German : Browne ist von Abrams interviewt worden.
English : Browne is being interviewed by Abrams.
German : Browne wird von Abrams interviewt.
English : Browne could have been interviewed by Abrams.
German : Browne hätte von Abrams interviewt **worden** können.
English : Browne could be being interviewed by Abrams.
German : Browne könnte von Abrams interviewt werden.
English : Browne has been being interviewed by Abrams.
German : Browne ist von Abrams interviewt **geworden**.
English : Browne could have been being interviewed by Abrams.
German : Browne hätte von Abrams interviewt werden können.
English : Abrams had Browne hired by Chiang.
German : Abrams liess Browne durch Chiang einstellen.

Structure: 'THAT' CLAUSE

German : Abrams wusste, wer Browne einstellte.
English : Abrams knew that Browne hired Chiang.
German : Abrams wusste, dass Browne Chiang einstellte.

Structure: INTERROGATIVE SUBORDINATE CLAUSE

English : Abrams knew who to hire.
German : Abrams wusste, **wer** man einstellt.
English : Abrams knew who Browne hired.
German : Abrams wusste, wer Browne einstellte.

The wrong case marking for the interrogative *who* is not always resolved by using English *who/whom*:

Structure: INTERROGATIVES

English : Who does Abrams work for?
German : Für **wer** arbeitet Abrams?
English : Which office does Abrams work in?
German : In welchem Büro arbeitet Abrams?
English : Whom did Abrams show an office to?
German : Zu **wen** zeigte Abrams ein Büro?
English : Who is Browne managed by?
German : Durch **wer** wird Browne verwaltet?
English : Which department is Abrams the manager of?
German : Welche Abteilung ist Abrams der Leiter **von**?
English : Whose department does Abrams work in?
German : In wessen Abteilung arbeitet Abrams?

Structure: RELATIVE CLAUSE

English : Abrams hired a woman who was competent.
German : Abrams einstellte eine Frau, die qualifiziert war.
English : Abrams hired women who were competent.
German : Abrams einstellte Frauen, die qualifiziert waren.
English : Abrams interviewed a woman who Browne showed an office to.
German : Abrams interviewte eine Frau, zu der Browne ein Büro zeigte.

Samples of MT Text Output

Below are sample illustrations of machine translated output (from German into English). The source, *TEXT 1*, is in fact slightly simplified from the original. The translation, *TEXT 1a*, is the result after modification of the dictionaries. For the reader's convenience, those parts of the translation which, for one reason or another, are deviant are highlighted in bold type.

TEXT 1: German Source Text

Struktur und Gestalt der Städte verändern sich schnell im neunzehnten und zwanzigsten Jahrhundert. Diese Veränderungen entsprechen der Wandlung der städtischen Gesellschaft im Industriezeitalter. Die Konzentration von Wohnungen gehört seit je zu den Kennzeichen der Stadt. Sie ist zur Erfüllung der städtischen Funktion notwendig. Das Gegenbild, die Gartenstadt, ignoriert die tatsächliche Situation und die Rolle der Großstadt. Sie kann diese Funktionen nur ergänzen. Das Ziel muß sein, die Großstadt aufzulockern. Man soll die Konzentration preiszugeben. Man soll zugleich die Landschaft vor der Zersiedelung bewahren.

TEXT 1a: English Translation after Dictionary Update

Structure and **figure** of the cities change quickly in the nineteenth and twentieth century. These changes correspond to **the** change of the urban society in the industrial age. The concentration of apartments belongs from time immemorial to the **marks** of the city. She/it is necessary for the fulfillment of the urban function. The opposite, **which** garden city, ignores the actual situation and the role of the metropolis. She/it can complete these functions **only**. The goal **be must, to loosen** the metropolis. One should to relinquish the concentration. One should protect at the same time **the** landscape from **the** overdevelopment.

The second text, *TEXT 2*, is a short extract from a German computer magazine; it describes fitting instructions for an item of equipment. *TEXT 2a* is the raw English output; *TEXT 2b* the result after modification of the dictionaries (i.e. entering unknown terms and semantic units; the term ‘feature-connector’ was left unmodified in this instance because the translator was uncertain of its English equivalent).

TEXT 2: German Source Text
Entfernen Sie alle Stecker aus Ihrem PC, beginnend mit dem Netzstecker. Stellen Sie das Gehäuse auf einen freien Tisch und entfernen Sie den Deckel von Ihrem PC. Lokalisieren Sie einen freien Steckplatz in der Nähe Ihrer Grafikkarte, damit das Verbindungskabel zum Feature-Connector paßt. Entfernen Sie die Blende von diesem Steckplatz.

TEXT 2a: English Raw Translation
Remove all plugs from your PC, beginning with the @@Netzstecker. Put the casing on a **free** table and remove **you** the cover of your PC. Localise a free outlet nearby your @@Grafikkarte, so that the @@Verbindungskabel fits with the feature - @@Connector. Remove them **Blind** from this outlet.

TEXT 2b: English Translation after Dictionary Update
Remove all plugs from your PC, beginning with the mains plug. Put the casing on an empty table and remove **you** the cover of your PC. Localise a free outlet nearby your graphics card, so that the connecting cable fits with the feature-@@Connector. Remove the hood of this outlet.

We may conclude that the PT performs more than adequately for the price and size of system. Experiences confirm the well known parameters of successful MT usage: texts of limited syntactic complexity and with a uniform vocabulary are translated to reasonable standards of comprehensibility and as a general rule are suited to post-editing, especially by trained persons familiar with the subject area; occasional recourse to the source text is required. In terms of the speed its performance and the user-friendliness of its interfaces, the PT scores highly.

Derek Lewis may be contacted at the Department of German, University of Exeter, Exeter EX4 4QH, e-mail: D.R.Lewis@exeter.ac.uk

The Natural Language Translation Specialist Group's Web-Site

by

Roger Harris

A description of the Natural Language Translation Specialist Group's web-site and its contents was published in the previous issue (No. 3, April 1996, p. 32) of *Machine Translation Review*. Since then I have added a further quantity of machine translation and linguistic reference material. Very approximately, some eighty A4 pages of information are now directly available. The amount of cross-linked machine translation and linguistic information available on other sites must surely amount to many thousands of pages. The cross-linked information is located at computer sites around the world and their electronic addresses, as displayed on your computer screen, are 'live': i.e. if you click on the address you will be automatically connected to that site. You do not need to type in the address.

Most of the electronic addresses of individuals are also 'live' and a mouse click will automatically transfer you to a simple word-processor screen which allows you to compose and send a message whilst on-line.

If the information in a file looks as if it might be useful, then you can download the file into your mailbox or set your communications software to log all data which is displayed on the screen. Downloading might be available on your system as a 'print' option, i.e. you 'print' the file to your mailbox or to a file in your on-line root directory. Downloading in this manner may not automatically keep a record of the full address path of the data, so you should consider jotting down the details.

The full address (URL) of the NLTSG web-site is as follows:

http://www.bcs.org.uk/siggroup/sg37.htm

The information which you will find there is listed under the following headings:

1. Committee (names, functions, e-mail addresses, telephone numbers).

2. Meetings (the most recent meeting was held in October 1996 when Dr. Ruslan Mitkov from the School of Languages, University of Wolverhampton, spoke on 'Machine Translation and Anaphora.').

3. *Machine Translation Review* (published twice yearly, in April and October, by the NLTSG). In addition to the topics listed below, each issue also contains book reviews and details of conferences and workshops.

   a. No.3 – April 1996 (summarised). Topics include:
      i. Implementing an Efficient Compact Parser for a Machine Translation System, by J. Gareth Evans.
      ii. Using Icon for Text Processing, by David Quinn.
      iii. The NLTSG's Web-Site, by Roger Harris.

   b. No.2 – October 1995 (summarised). Topics include:
      i. Matching Words in a Bilingual Corpus, by Roger Garside.
      ii. Lexical Resources for MT: a Survey, by Adam Kilgarriff.

iii. A Corpus-based Bilingual Dictionary: Why and How? by Marie-Hélène Corréad.

c. No.1 – April 1995 (summarised). Topics include:
   i. Multilingual Natural Language Processing (MNLP) Project, by David Wigg.
   ii. Machine Translation – Ten Years On: Cranfield Conference Report, by Derek Lewis.
   iii. CAT2: A Unification-Based Machine Translation System, by Ruslan Mitkov.
   iv. Practical Aspects of the Use of METAL at Siemens Nixdorf, by Keith Roberts.
   v. Linguistic Resources on the Internet, by Roger Harris.

4. Machine translation resources on the Internet and elsewhere

a. A-Z of linguistic and MT items (contacts, street addresses, e-mail addresses, WWW sites (http), ftp, gopher)

b. Books about MT (periodicals, books, booksellers)

c. E-mail linguistic lists (receive and contribute items of linguistic interest)

d. Newspaper corpora (vast text resources in many languages)

e. Suppliers of MT systems

f. Translators (translation services, employment agencies seeking translators)

g. Usenet newsgroups (on-line discussion groups)

I should like to thank Pam Bolwell, BCS Net Editor, for her expertise and patience in editing and correcting my HTML coding. The BCS web-site (URL: http://www.bcs.org.uk/) contains details of all the BCS specialist groups, forthcoming meetings and other items of interest.

If, after inspecting the NLTSG web-site, you have any comments or suggestions, then I should be happy to hear from you. Please send an e-mail to

rwsh@dircon.co.uk,

or telephone +44 (0)181 800 2903.

Book Reviews

B.T.S. Atkins and A. Zampolli (1994) *Computational Approaches to the Lexicon*, Oxford: Clarendon Press. Hardback £55. 479 pages. ISBN 0-19-823979-3.

This publication contains a selection of sixteen papers from the fifth Pisa International Summer School on Computational Lexicology and Lexicography (1988). It attempts to cover the major scientific and technical aspects of computational lexicology and lexicography. After an initial introduction and overview, the volume is divided into three parts. The first of these, containing four papers, is concerned with the collection and processing of textual data. The second part includes seven papers and is concerned with the theoretical infrastructure of lexical analysis. The final part, containing three papers, is concerned with methodology and tools.

In their overview in Chapter Two, Atkins, Levin and Zampolli trace the evolution of computational lexicology and lexicography up to the start of the current decade. They begin with a short review of pertinent work in theoretical linguistics and observe how the emerging fields of computational and corpus linguistics sprung from corresponding progress in theoretical linguistics and the introduction of computational techniques into linguistic research. Next a summary of the developing role of the computer in lexicography proper is given. This is followed by a survey of lexical resources, their availability and reusability. They conclude by relating the various chapters of the book to present work in the field of computational approaches to the lexicon.

The section on the collection and processing of textual data is concerned with two problems that face today’s corpus-builders. On the one hand there are the criteria for the selection of texts for inclusion in a corpus. On the other hand, annotation and structuring have to be considered, as does the level of mark-up. This section begins with a chapter by Gunnel Engwall in which she discusses corpus design features and criteria relevant to the construction of a representative corpus. This chapter is illustrated by Gunnel’s own experience in designing the Swedish corpus of modern best-selling French novels. There follows a chapter by Stig Johansson which reflects his contribution to the Text Encoding Initiative. A summary of methodology of basic encoding of electronic textual resources is given. Johansson shows that transformation of texts into textual resources may be regarded as a process of interpretation. Thus compilers of on-line corpora possess a responsibility normally associated with the role of an editor. The stages of the conversion process are described and certain challenges are addressed. These include both familiar problems, such as the encoding of special characters, as well as more esoteric problems such as the encoding of editorial comments. This is followed by a chapter in which Donald Hindle discusses the problems of parsing a corpus tagged with Fidditch, a part-of-speech tagger designed to parse both written text and transcripts of spoken text. Whilst Hindle’s chapter is essentially concerned with syntagmatic relationships, the chapter by K. W. Church et al. discusses paradigmatic relationships. It gives a detailed description of the employment of statistical methods to elicit lexical data from text corpora. Examples from Associated Press newswire are used by way of illustration. A statistical measure of substitutability is proposed which is effectual in the selection of a set of words which are both syntactically and lexically ‘substitutable’: in other words, members stand in some statistical relationship to one another in connection with co-occurrence with some lexical item. Identifying potential sets of synonyms is an obvious use for this device.

The section dealing with the theoretical infrastructure of lexical analysis is concerned with the representative selection of prevailing work on the theoretical basis of computational lexicology and lexicography. This gives rise to a diversity of chapters. Here the emphasis is mainly on the syntactic rather than the semantic component of the lexicon. The chapter by A. Zaenen and G. Engdahl investigates the requirement of various syntactic theories in a lexicon. They discuss ways in which the outcome of such work can aid lexicographers and compilers of lexical resources. This includes a description of the characteristic syntactic properties of a number of kinds of sentential-complement-taking verbs. Furthermore they investigate the requirements that these verbs impose on a lexicon with reference to two frameworks: Government-Binding and Lexical Functional Grammar. The authors show that the information required by these two models is similar. The conclusion drawn is that classes of lexical items are only partially understood. As a result, linguistic theory is currently of limited use to the designer of lexical resources.

The remaining modules in this section describe case studies. The first of these outlines aspects of work at the Laboratoire d'Automatique Documentaire et Linguistique at the Université de Paris VII. This project involves the construction of lexicon-grammar of French. A number of difficult issues that emanate from recording the properties of verbs are discussed.

The following chapter concerns types of grammatical data in a lexicon or more specifically the lexical facts that contribute to determining word order. E. Hajicova postulates that adequate linguistic description of a lexical item should include its meaning, grammateme, complement frame and other subcategorisation conditions. She chooses to focus on the complement frame.

The chapter by S. Dik discusses the requirements placed on the lexicon by a computational Functional Grammar (FG). A lexical component of a Prolog implementation of a Functional Grammar of English is described. This implementation involves a minimalist view of the lexicon in which the lexicon contains only non-derivable information and lexical entries include information about form, meaning and collocational properties.

J. Pustejovsky and B. Boguraev discuss another aspect of the semantic component concerned with knowledge representation and the theoretical issues which are entailed by the construction of this component in the lexicon. They argue against the notion that word meanings are fixed and inflexible, where lexical ambiguity must be treated by multiple entries in the lexicon. Instead they postulate that that this component should be 'generative' because the facets of word meaning that aid in deciding the types of productive relationships in which words may participate, are thus made explicit. In their system, word senses are related by logical operations defined by the well formedness rules of the semantics. As a result novel and ambiguous usages are more easily managed.

The chapter by S. Nirenburg is concerned with a lexical knowledge acquisition system for machine-translation. This is part of wider research programme at the Center for Machine Translation at Carnegie-Mellon University. He describes the structure of the lexicons in two application-oriented NLP systems and the tools which were used for knowledge acquisition in these systems.

The final chapter in this section, by C. J. Fillmore and Atkins, arises from the collaboration of a theoretical linguist and a lexicographer. Data extracted from a text corpus was employed in conjunction with a different framework within which to describe the meaning of a word. This leads to the elucidation of certain lexicographical problems and the more subtle and faithful description of interaction between semantics and syntax. The authors analyse the descriptions of the word ‘risk’ found in ten dictionaries and indicate a number of semantic and syntactic issues that have been the cause of confusion. They propose a different way of analysing this word which accounts for many of these issues. They conclude by arguing that the traditional printed dictionary format is too restricted to provide an adequate description of the usage of ‘risk’ and suggest that new electronic resources could be utilised in a new approach to lexicography.

Part Four is comprised of three chapters. The first of these, by M. Nagao, deals with the rationale and methodology applied in the creation of the *Iwanami Encyclopedic Dictionary of Computer Science,* a terminological dictionary, at Kyoto University. She describes how computational resources were fully utilised in this project for both lexicographical evidence and as tools in a machine-assisted process of compilation.

In the chapter which follows, E. Weiner describes his experience in the preparation of the second edition of the Oxford English Dictionary (twenty volumes). He discusses machine assisted compilation from the point of view of the editor of a large-scale and highly sophisticated scholarly dictionary. He looks into the future of the lexicography-computer relationship in general language lexicography. He postulates a computer workstation equipped with a wide range of facilities that will enable the dictionary editor to undertake his work with a minimum of unnecessary labour. Such facilities would include on-line access to existing dictionaries, highly sophisticated data entry and updating systems, generic editors, cross-reference followers, electronic quotation capture, on-line quotation files, ancillary research and revision files, and electronic mail for use between editors and correspondents.

In the final chapter, Calzolari and E. Picchi give a detailed description of the design and function of a system built at the ILC at Pisa. This system provides resources needed by both lexicographers and scholars using on-line resources. They give an account of a workstation that links textual data to a structured machine-readable dictionary database and thereby enhances it. More specifically the system provides access to textual corpora, creation of a lexical database, access to the lexical database, and integrated access to texts and dictionaries.

I would happily recommend this volume, which discusses a variety of current issues in the emerging new discipline of computational lexicography. The material contained in the book is diverse, providing insights into how certain individuals solved their particular problems. The book includes an index, something which is not always included in collections of papers. I would, however, have preferred to see the references collated at the end of the volume rather than at the end of each chapter.

Jon Mills

Maria Theresia Rolland (1994) *Sprachverarbeitung durch Logotechnik: Sprachtheorie. Methodik. Anwendungen* Bonn: Ferdinand Dümmler Verlag. Paperback, xxv + 597 pp. ISBN 3-427-83741-6.

Maria Theresia Rolland presents *Logotechnik* as a new method for the analysis and computer processing of natural language, in this case German. *Logotechnik* means 'manipulation of the word'. The term reflects the method's theoretical basis in Leo Weisgerber's *inhaltbezogene Grammatik* (content-orientated grammar), a model which attracted some attention in Germany during the early 1960s. At the heart of the grammar is the view that the word and its semantic content (as determined by its relationships with other words) is the fundamental unit of analysis. In this volume Rolland attempts to revive interest in Weisgerber's approach, which fell into obscurity as contemporary linguists focused on the syntax-oriented Chomskyean grammars that emerged at the time. She concedes that the model is difficult to understand but explains how it might be implemented as a basis for natural language processing.

Although the reader's understanding of what constitutes a 'word' is not helped by some rather convoluted and decidedly circular formulations (most of which boil down to statements which are best summarised as 'a word is a word'; see, for example, p. 53), the model in practice adopts and greatly refines traditional word classes. A word is seen as a semantic totality which may extend over separate phonetic and syntactic units. Thus there is no single verb *waschen* ('to wash') with a reflexive variant *sich waschen* ('to wash oneself', 'have a wash'). Although both forms share the notion of 'washing', they are considered to be separate verbs: the fact that *waschen* has an active and a passive form while *sich waschen* occurs only in the active contributes to the distinctive 'total content' (*Gesamtinhalt*) for each form (p. 65). This approach inevitably leads to a proliferation of word classes and subclasses as determined by their perceived content. Thus the traditional classification of personal pronouns into six (singular and plural) persons is rejected in favour of an 'eight subject-category' scheme which distinguishes the *er, sie, es* pronouns in German that may refer to a person (e.g. a man (masculine), a woman (feminine), and a child (neuter)) from the phonetically identical forms denoting non-human subjects (p. 67). The model also allows for some highly abstract (and not always transparent) classes of general word-types. For instance, the inflected forms of nouns (*des Tishes, diese Tische* = 'of the table', 'these tables') are seen as a manifestation of the general category of *Verlaufswort* ('a progressive or continuous word'), which also includes the inflected forms of verbs, present participles, and conjunctions such as *und* ('and'; see p. 118). Words and phrases combine to give more general syntagmatic structures at clause and sentence level. The book is largely concerned with identifying and labelling such structures.

Rolland claims that the model contains all that is required to produce a parsing algorithm and a knowledge database for German. The method is worked through on paper step-by-step. Morphological variations of words appear to be stored in a lexicon as separate items in lists, as are permissible phrasal structures and frames for sentence-types. These lists form the basis for identifying word and phrase classes; there are no generalised parsing rules. The first step in sentence analysis is to draw up a list of high frequency items in the sentence which are not part of the main verb (such as articles, pronouns, adverbs, prepositions, conjunctions, and the basic forms of auxiliary verbs); ambiguity of grammatical function is resolved during sentence processing. For sentence analysis a left-to-right look-up procedure identifies word classes and establishes the central verb predicate structure together with its associated case elements. The output of the analysis is stored as an 'analogue sentence': this is a flat list (not a tree) of labelled word and phrase categories; the labels reflect the content-based classes and subclasses described above. Pairs of stored input and output (analogue) sentences may be built up automatically and stored as the ‘knowledge base’ of a linguistic database. This may be queried using stereotyped questions of the form ‘who does what to whom, when and where’; the elements of the query are simply mapped onto the elements of the analogue sentence.

Rolland states that an operational system could be implemented from her model within ten years and used as a component in a machine translation (MT) system. In what may be possibly criticised as a rather simplified view of MT, she sees the process as a straightforward mapping between labelled parallel structures in source and target languages, with some re-ordering of constituents required; her model does not resolve the questions of the depth of semantic representation and the problems of transfer that have occupied researchers for many years. Although the model claims to depart radically from contemporary approaches to language processing, it operates with syntactico-semantic categories drawn from a carefully refined but essentially traditional grammatical framework. Whether enough information is stored on analogue sentences to facilitate mapping between different languages is debatable. Acknowledged problems in parsing, such as pronominal reference and anaphora resolution, are ignored. Perhaps the ten years envisaged for implementation would address such issues. It is uncertain whether this model could underpin a viable processing system for German, but the work represents a remarkable achievement for its study of word classes and its extensive lists of permissible structures.

Derek Lewis

(This review is reprinted and adapted by kind permission of the Germanic editors of the *Modern Language Review.*)

Conferences and Workshops

The following is a list of recent (i.e. since the last edition of the MTR) and forthcoming conferences and workshops. Telephone numbers and e-mail addresses are given where known (please check area telephone codes).

17–18 May 1996
Conference on Empirical Methods in Natural Language Processing
University of Pennsylvania, USA
Tel: +1 215 898 6564

4–6 June 1996
TAL+AI International Conference on Natural Language Processing and Industrial Applications
96 Moncton, New-Brunswick, Canada
Tel: +1 506 858 4521, fax: +1 506 858 4541, e-mail: nlp-ia@umoncton.ca

4–7 June 1996
ICCC96 International Conference on Chinese Computing 96
Institute of Systems Science, National University of Singapore
Tel: +65 772 3107, fax: +65 774 4998, e-mail: ICCC96@iss.nus.sg

9–12 August 1996
TALC96 Teaching and Language Corpora
Lancaster University, UK
Fax: +44 1524 843085, e-mail: mcenery@computing.lancaster.ac.uk
http://www.comp.lancs.ac.uk/computing/research/ucrel/talc/

13 August 1996
ECAI96 Workshop on Dialogue Processing in Spoken Language Systems
Budapest, Hungary
E-mail: maier@dfki.uni-sb.de, mast@heidelbg.ibm.com, susann@azrael.mitre.org

22 August 1996
SIGIR96 Workshop on Cross-linguistic Information Retrieval
E-mail: sigir96@ubilab.ubs.ch.
http://www.ubilab.ubs.ch/sigir96/welcome.html.

27 August 1996
Workshop on Future Issues for Multilingual Text Processing
Cairns, Australia
Tel: +61 3 9344 4227, fax: +61 3 9349 4326
E-mail: Dominique.Estival@linguistics.unimelb.edu.au

29–31 August 1996
ROCLING IX: Research on Computational Linguistics
Tseng-Wen Reservoir Youth Activity Center, Tainan County, Taiwan, R.O.C.
Fax: +886 6 2747076, e-mail: chwu@server2.iie.ncku.edu.tw

16–18 September 1996
NeMLaP-2 Conference, Bilkent University, Ankara, Turkey
Tel: +90 312 266 4126, fax: NeMLaP-2, registration c/o Kemal Oflazer

23–27 September 1996
Herausforderungen an die Computerlinguistik: Multilingualität, Multimedialität, Multidisziplinarität
University of Magdeburg, Germany
Tel: +49 391 6718718, fax: +49 391 672018, e-mail: herbstschule@iik.cs.uni-magdeburg.de
http://www-ai.cs.uni-magdeburg.de/herbstschule96.html

7–9 October 1996
KONVENS'96: Verarbeitung natürlicher Sprachen
University of Bielefeld, Germany
Tel: +49 521 1063679, fax: +49 521 1062996, e-mail: lobin@lili.uni-bielefeld.de
http://coral.lili.uni-bielefeld.de/konvens96/

5–8 November 1996
PAP'97: 5th International Conference on Principles of Knowledge Representation and Reasoning
Cambridge, Massachusetts, U.S.A.
Tel: +1 415 328 3123, fax: +1 415 321 4457, e-mail: kr@aaai.org
http://kr.org/kr/

14–15 November 1996
Translating and the Computer 18
ASLIB: The Association for Information Management
1, Great George Street, Westminster, London
Tel: +44 171 253 4488, fax: +44 171 430 0514, e-mail: pdg@aslib.co.uk
http://www.aslib.co.uk

16–18 December 1996
International Conference on Knowledge-based Computer Systems
Bombay, India
Tel : +91 22 620 1606, fax: +91 22 621 0139, e-mail: kbcs@konark.ncst.ernet.in
http://konark.ncst.ernet.in/~kbcs/kbcs96.html

8–10 January 1997
IWCSII: 2nd International Workshop on Computational Semantics
Tilburg, The Netherlands
Tel: +31 13 466, fax: +31 13 466 311030, e-mail: Computational.Semantics@kub.nl60
http://tkiwww.kub.nl:2080/tki/Docs/IWCS/iwcs.html

3–4 April 1997

TIA'97: Terminologie et Intelligence Artificielle
University of Toulouse-le Mirail, France
Tel: +33 61 503608, fax: +33 61 504677, e-mail: pery@cict.fr

11 April 1997
2nd Aston Corpus Linguistics Seminar: Register and Corpus Dynamics"
Aston University, UK
Tel: +44 121 359 3611, e-mail: c.j.gledhill@aston.ac.uk

20–23 April 1997
SDAIR97 6th Annual Symposium on Document Analysis and Information Retrieval
Alexis Park Resort, Las Vegas, Nevada, USA
E-mail: pedersen@parc.xerox.com

21–28 April 1997
TPAP97: 5th International Conference on the Practical Application of PROLOG
London, UK
Tel: +44 1253 358081, fax: +44 1253 353811
http://www.demon.co.uk/ar/TPAC

3–7 June 1997
ACH-ALLC97 Association of Computing in the Humanities, Association of Literary and Linguistic Computing
Queen's University, Kingston, Ontario, Canada
Tel: +1 613 545 2083, fax: +1 613 5456522
http://www.qucis.queensu.ca/achallc97

14–16 July 1997
1st International Workshop on Human-Computer Conversation
Grand Hotel Villa Serbelloni, Bellagio, Italy
http://www.dcs.shef.ac.uk/research/ilash/Meetings/Bellagio/

24–28 July 1997
TMI'97: 7th International Conference on Theoretical and Methodological Issues in Machine Translation, St John's College, Santa Fe, New Mexico, USA
http://crl.nmsu.edu/Events/TMI

27–31 July 1997
SIGIR'97: Research and Development in Information Retrieval
DoubleTree Hotel, Philadelphia, PA, USA
Tel: +1 301 975-3761, fax: +1 301 840-1357, e-mail: ellen@potomac.ncsl.nist.gov

1–2 August 1997
EMNLP-2: 2nd Conference on Empirical Methods in Natural Language Processing
ACL Special Interest Group SIGDAT
Brown University, Providence, Rhode Island, USA
Claire Cardie, e-mail: cardie@cs.cornell.edu, tel: +1 607 255 9206

11–22 August 1997

ESSLLI'97: 9th European Summer School in Logic, Language, and Information
Literary Faculty, University of Aix-Marseille I, Aix-en-Provence, France
Fax: +33 442 595096, e-mail: esslli97@lpl.univ-aix.fr
URL: http://www.lpl.univ-aix.fr/~esslli97

22–24 August 1997
ROCLING X: Research on Computational Linguistics
Academia Sinica, Taipei, Taiwan
Tel: +1 908 582 5296, fax: +1 908 582 3306
E-mail: rocling@hp.iis.sinica.edu.tw; also: rocling@research.bell-labs.com

22–25 September 1997
EUROSPEECH97: 5th European Conference on Speech Communication and Technology
Rhodes, Greece
Tel: +30 61 270388, fax: +30 61 223 335, e-mail: cpr@pat.forthnet.gr
http://ophale.icp.grenet.fr/esca/esca.html

MEMBERSHIP: CHANGE OF ADDRESS

If you change your address, please advise us on this form, or a copy, and send it to the following
(this form can also be used to join the Group):

Mr. J.D.Wigg
BCS-NLTSG
72 Brattle Wood
Sevenoaks, Kent TN13 1QU
U.K.

Date: ....../....../......

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

Note for non-members of the BCS: your name and address will be recorded on the central computer records of the British Computer Society.

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/was 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.4, Oct. ‘96) 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.