MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society
Issue No. 7
April 1998

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
School of Modern Languages
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

72 Brattle Wood
Sevenoaks
Kent, TN13 1QU

Tel: 01732 455446
Office: 0171 815 7472
Fax: 0171 815 7550
E-mail: wiggjd@sbu.ac.uk

2 June 1998

I regret I have to open this letter by announcing the recent death of the initiator and founder of our Specialist Group in 1976, Walter Goshawke. Walter was an enthusiastic and unswerving promoter of his vision of a machine translation system based on a numeric interlingua. He worked doggedly on this system for many years and attended Committee meetings to encourage us in our efforts to promote machine translation until increasing infirmity finally prevented him from going out. We shall miss his enthusiasm and determination.

The Group is now entering its fourth year of publication of the *Machine Translation Review* which is a very creditable performance due to Derek Lewis' stalwart efforts as Editor and the voluntary contributions of papers and information supplied by members.

However, as I have intimated previously, although we are generously supported by the British Computer Society, publishing twice a year is exhausting our reserves and we will have to consider our future options soon. There are also signs that we are also exhausting our sources of material for publication, so one way or another we may still have to reduce publication to annually.

We have considered upgrading the *Review* into a refereed journal, but although we could probably obtain an excellent refereeing panel we still cannot feel confident of attracting sufficient material for regular publication.

In the short run it has been suggested that we publish on our web pages at the BCS, which Roger Harris maintains so well for us, with the option of members being able to ask for printed copies at cost. In any event we would still welcome more articles, papers and reports on the subject of machine translation and related subjects such as computer assisted language teaching, computer based dictionaries and aspects of multilinguality in computing etc. We would welcome papers from academic staff and students in linguistics and related disciplines, from translators and any other users of MT software.

If you are sufficiently interested in machine (assisted) translation to read this publication you could well have some interesting knowledge or experiences to pass on to other members, so please do not be backward in coming forward with further contributions.

Perhaps I could remind members that they do not need to live near London to assist the Committee. We do not have sufficient funds to pay travel expenses for all Committee

members to attend meetings, but we still welcome Correspondent Members. Correspondent Committee Members are otherwise treated as full members of the Committee and kept advised of all committee business. Anyone interested in helping should contact me or any other Committee member.

The Proceedings of the International Machine Translation Conference at Cranfield in 1994 are now complete and negotiations are in hand to print them. We expect the cost to be about £25-£30. We are considering organising another Conference in the Autumn of 1999, probably at Exeter University. If you would like to take any part therein or if you have any comment to make about it, please contact me or any other Committee member.

Finally, I should mention our AGM, which took place at King's College, London, on 11 December 1997. Apart from the usual business, our speaker was Derek Lewis, who presented for demonstration and informal evaluation three MT systems: Globalink's Power Translator, Langenscheidt's T1 system, and the Easy Translator (an example of a Web Page translator). A fuller comparative account of these systems may be found in the article 'Machine translation today', in *The Linguist*, Vol. 37, No. 2, 1998, pp. 38-43.

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.

Website

The website address of the BCS-NLTSG is: http://www.bcs.org.uk/siggroup/sg37.htm

Translation Technology and the Translator

by

John Hutchins
University of East Anglia, Norwich

Introduction

Translators are perhaps the most critical audience for presentations about the automation of translation. Many of them will agree with comments made by J.E.Holmström in a report on scientific and technical dictionaries submitted to Unesco in 1949.

Having heard that some researchers were investigating the possibilities, he thought that 'the resulting literary style would be atrocious and fuller (sic) of 'howlers' and false values than the worst that any human translator produces'. The reason was that 'translation is an art; something which at every step involves personal choice between uncodifiable alternatives; not merely direct substitutions of equated sets of symbols but choices of values dependent for their soundness on the whole antecedent education and personality of the translator.' His comments preceded by five years the first tentative demonstration of a prototype system, and were based on pure speculation. Nevertheless, such comments have been repeated again and again by translators for nearly fifty years, and no doubt they shall be heard again in the next fifty.

However, we shall see that computer-based translation systems are not rivals to human translators, but they are aids to enable them to increase productivity in technical translation or they provide means of translating material which no human translator has ever attempted. In this context we must distinguish (1) machine translation (MT), which aims to undertake the whole translation process, but whose output must invariably be revised; (2) computer aids for translators (translation tools), which support the professional translator; and (3) translation systems for the 'occasional' non-translator user, which produce only rough versions to aid comprehension. These differences were not recognised until the late 1980s; the previous assumption had been that MT systems, whether running on a mainframe or a microcomputer, could serve all these functions with greater or less success. In part, this failure to identify different needs and to design systems specifically to meet them has contributed to misconceptions about what translation technology can do for the professional translator.

The First MT Systems

When MT was in its infancy, in the early 1950s, research was necessarily modest in its aims. It was constrained by the limitations of hardware, in particular by inadequate computer memories and slow access to storage, and by the unavailability of high-level programming languages. Even more crucially it could look to no assistance from the language experts. Syntax was a relatively neglected area of linguistic study and semantics was virtually ignored. The early researchers knew that whatever systems they could develop would produce poor quality results, and they assumed major involvement of human translators both in the pre-

editing of input texts and in the post-editing of the output. They proposed also the development of controlled languages and the restriction of systems to specific subject areas.

In this atmosphere the first demonstration systems were developed; notably, collaboration took place between IBM and the Georgetown University in 1954. Based on small vocabularies and carefully selected texts, the translations produced were impressively colloquial. Consequently, the general public and potential sponsors of MT research were led to believe that good quality output from automatic systems was achievable within a matter of a few years. The belief was strengthened by the emergence of greatly improved computer hardware, the first programming languages, and above all by developments in syntactic analysis based on research in formal grammars (e.g. by Chomsky and others.)

For the next decade MT research grew in ambition. It became widely assumed that the goal of MT must be the development of fully automatic systems producing high quality translations. The use of human assistance was regarded as an interim arrangement. The emphasis of research was therefore on the search for theories and methods for the achievement of 'perfect' translations. The current operational systems were regarded as temporary solutions to be superseded in the near future. There was virtually no serious consideration of how 'less than perfect' MT could be used effectively and economically in practice. Even more damaging was the almost total neglect of the expertise of professional translators, who naturally became anxious and antagonistic. They foresaw the loss of their jobs, since this is what many MT researchers themselves believed was inevitable.

Progress was much slower than expected, and the output of systems showed no sign of improvements. In these circumstances it was not surprising that in 1966 a committee set up by US sponsors of research — the Automatic Language Processing Advisory Committee (ALPAC) — found that MT had failed according to its own aims, since there were no fully automatic systems capable of good quality translation and there seemed little prospect of such systems in the near future.

While this ALPAC report brought to an end many MT projects, it did not banish the public perception of MT research as essentially the search for fully automatic solutions. The subsequent history of translation technology is in part the story of how this mistaken emphasis of the early years has had to be repaired and corrected. The neglect of the translation profession has been made good eventually by the provision of translation tools and translator workstations. MT research has itself turned increasingly to the development of realistic practical systems where the necessity for human involvement at different stages of the process is fully accepted as an integral component of their design architecture.

Since the early 1970s development has continued in three main strands: computer-based tools for translators; operational MT systems involving human assistance in various ways; and 'pure' theoretical research towards the improvement of MT methods.

MT in Operation

Until the late 1980s one paradigm dominated the utilisation of MT systems. It had been inherited from the very earliest days: the system produced large volumes of poorly translated texts, which were either used for the assimilation of information directly or submitted for extensive post-editing, with the aim of obtaining texts of publishable quality for dissemination. As a means of improving the quality many organisations introduced controls on the vocabulary, structure and style of texts before input to systems, and this has been how

Systran, Logos, METAL and similar mainframe systems have been used (and continue to be used) by multinational companies and other large organisations.

When the first PC versions of MT systems appeared it was widely assumed that they would be used in much the same way: to obtain 'rough gists' for information purposes or as 'draft translations' for later refinement. In both cases, it was also widely assumed that the principal users of MT systems would be translators or at least people with a good knowledge of both source and target languages; and, in the case of large organisations, it was expected that most MT users would be professionally trained translators.

However, during the late 1980s — and with increasing pace since the early 1990s — this paradigm and its assumptions have been broken by developments on a number of fronts. Firstly, there has been the commercial availability of translator workstations, designed specifically for the use of professional translators; these are essentially computer-based translation tools and not intended to produce even partial translations fully automatically. Secondly, the PC-based systems have been bought and used by an increasingly large number of people with no interest in translation as such; they are being used as 'aids for communication', where translation quality are of much less importance. Thirdly, there came the development of domain-specific systems by clients themselves: custom-built systems accepting input in a constrained vocabulary and integrated closely in documentation and publication systems. Fourthly, the growth of telecommunication networks with communication across many languages has led to a demand for translation devices to deal rapidly in real time with an immense and growing volume of electronic language. Finally, the wider availability of databases and information resources in many different languages has led to the need for multilingual search and access devices which incorporate translation modules.

All current commercial and operational systems produce output which must be edited (revised) if it is to attain publishable quality. Only if rough translations are acceptable for information analysis purposes can the output of MT systems be left unrevised. Commercial developers of MT systems now always stress to customers that MT does not and cannot produce translations acceptable without revision: they stress the imperfect nature of MT output. They recognise fully the obligation to provide sophisticated facilities for the formatting, input, revision and publication of texts within total documentation processing from initial authoring to final dissemination.

It is now widely accepted that MT works best in domain-specific and controlled environments. The first domain-specific success was Meteo, a system designed for translating weather forecasts from English into French and used continuously since 1977 by the Canadian broadcasting service. The use of controlled input was taken up in the late 1970s by Xerox for its implementation of the Systran system. Other applications of controlled input have followed in the 1980s and 1990s with other general-purpose systems, e.g. for the localisation of computer software for sale in many countries and languages.

However, rather than adapting general-purpose MT systems in this way, it is now recognised that it is better to design systems ab initio for use with controlled language. A number of independent companies outside the academic MT research community have been doing this in recent years (e.g. Volmac); the largest current development is the Caterpillar project based on the research at Carnegie Mellon University.

Tools for Translators

In general most commentators agree that MT (full automation) as such is quite inappropriate for professional translators. They do not want to be subservient to machines; few want to be revisers of poor quality MT output. What they have long been asking for are sophisticated translation tools. Since the early 1990s they can now have them in the shape of translation workstations. These offer translators the opportunity of making their work more productive without taking away the intellectual challenge of translation.

Translator workstations combine access to dictionaries and terminological databanks, multilingual word processing, the management of glossaries and terminology resources, appropriate facilities for the input and output of texts (e.g. OCR scanners, electronic transmission, high-class printing).

The development of translation tools became feasible, firstly with the availability of real-time interactive computer environments in the late 1960s, then with the appearance of word processing in the 1970s and of microcomputers in the 1980s and, subsequently, with intra-organisational networking and the development of larger computer storage capacities. Although workstations were developed outside the older MT research community, their appearance has led to a decline of the previous antagonism of translators to the MT community in general. They are seen to be the direct result of MT research. Indeed, the 'translation memory' facility, which enables the storage of and access to existing translations for later (partial) reuse or revision or as sources of example translations, does in fact derive directly from what was initially 'pure' MT research on bilingual text alignment within a statistics-based approach to automatic translation.

At the present time, the sales of translator workstations incorporating translation memories are increasing rapidly, particularly in Europe. Their success has built upon translators' experience with terminology management systems and upon the demonstrable improvements of productivity, terminological consistency and overall quality. The next stage of development will be the fuller integration of MT modules in order to provide automatic translation of sentences or text fragments when required, e.g. if the existing texts in a translation memory do not provide usable translation sources.

Research for Machine Translation

After ALPAC, research on MT has, of course, continued. However, the field has continued to attract perfectionists. Very often systems have been developed without any idea of how they might be used or who the users might be. MT has been seen as a testbed for exploring new linguistic and computational techniques. In nearly every case, it was found that the 'pure' adoption of a new theory was not as successful as initial trials on small samples appeared to demonstrate. The basic lesson is that MT demands an eclectic approach, the use of hybrid methods combining a variety of techniques; and, above all, no quick results can be expected with any new approach.

What was often forgotten is that MT is the application of computational and linguistic methods and techniques to a practical task; that translation is itself a means to an end — a task which has never been and cannot be 'perfect'; there are always other possible (often multiple) translations of the same text according to different circumstances and requirements. MT can be no different: there cannot be a 'perfect' automatic translation. The use of an MT system is contingent upon its cost effectiveness in practical situations.

Within the last ten years, research on spoken translation has developed into a major focus of MT activity. Research projects such as those at ATR in Japan, Carnegie-Mellon University in

the US and on the Verbmobil project in Germany are ambitious. But they do not make the mistake of attempting to build all-purpose systems: systems are constrained and limited to specific domains, sublanguages and categories of users. Nevertheless, there are obvious potential benefits even if success is only partial.

Research has begun also on systems for speakers or writers who are ignorant of the target language, an area neglected in the past. In these cases what is required is a means of conveying a message in an unknown language; it does not have to be a straight translation of any existing original. From interactive dialogue a translatable (MT-amenable) 'message' can be composed for automatic conversion into an idiomatic and correct message in the target language without further involvement of the originator.

As for translation for those wholly ignorant of the source language, this need has been met until recently by the use of unrevised outputs from older batch-processing systems, i.e. as by-products of systems primarily intended to produce translations for revision before publication. Within the last decade, however, cheap PC-based software has appeared on the market which can be (and undoubtedly is being) used by monolinguals who want only to grasp something of the gist of texts. They are not wholly satisfactory, of course, and the development of fully automatic systems specifically for this potentially huge market is a challenge for future MT research.

Translation and Networking

With the expansion of global telecommunications (the Internet and World Wide Web) has come the networking of translation services. Nearly all the larger MT software vendors now offer their systems as a service to individual or company customers. Texts can be sent on-line for immediate 'rough' translation with no post-editing, or for treatment in a more traditional manner with expert revision, editing and preparation for publication by the service. This form of networked MT is clearly a further development of familiar translation services, and one with considerable growth potential. It is assumed that in future there will emerge various forms of networked 'translation brokerage' services which will advise customers on the most appropriate MT service for their needs, e.g. in terms of costs, languages, speed, dictionary coverage, terminology control, overall translation quality, post-editing support, etc. Some of these 'translation brokers' may themselves be automated, and undertake searches of the Web for particular client needs. As a consequence, we may well see the emergence of more specialised MT systems (for particular domains and language pairs), some of which will thrive and others which will fail in the global competitive market.

Even more significant for the future, however, is the appearance of systems for on-line and real-time translation of electronic mail messages. In 1994 the CompuServe service introduced automatic translation from and to English and French, German or Spanish for messages on one of its forums. It became so popular that within the next two years the facility was extended to two other on-line services. Now thousands of messages a day are being translated. The software used was not, of course, designed originally to deal with the frequently ungrammatical conversational style and the sometimes idiosyncratic vocabulary of electronic mail. Hence much of the output is garbled and barely comprehensible; but a large number of users have found the results to be a valuable aid for comprehension.

Only a fully automatic system could operate in real-time on this scale. The potential market for network MT systems is enormous. At CompuServe alone there are more than 3,000 other on-line services where MT could be introduced; and other Internet services could easily

follow their lead. It has been estimated that there are currently over 40 million electronic mail messages a month. If only a small fraction of these were candidates for translation, the demand would be enormous.

In addition to electronic messages, the amount of information available in text form on Web pages can now counted in their hundreds of millions, and it is growing exponentially at a high rate (10% between 1995 and 1996). The non-English content is estimated as 80% of the total, and there is no doubt that readers everywhere prefer to have text in their own language, no matter how flawed and error-ridden it may be, rather than to struggle to understand a foreign language text. The Japanese software companies have already recognised the huge potential market and there are a number of English-Japanese translation modules available for integration with Web software. Similar Web translation software is being developed and sold for other languages, both by existing vendors of MT systems and by new companies.

A further factor will be the growth of multilingual access to information sources. Increasingly, the expectation of users is that on-line databases should be searchable in their own language, that the information should be translated and summarised into their own language. The European Union is placing considerable emphasis on the development of tools for information access for all members of the community. Translation components are obviously essential components of such tools; they will be developed not as independent stand-alone modules, but fully integrated with the access software for the specific domains of databases. The use of MT in this wider context is clearly due for rapid development in the near future.

There is no gainsaying the enormous potential for the translation of electronic messages. Only a fully automatic process capable of handling large volumes with close to real-time turnaround can provide the translation capacity required by on-line markets. In addition the on-line ‘culture’ favours rapid and ‘shallow’ assimilation of information; for these reasons MT is the obvious future. It is now evident that the true niche market for MT is in ‘cyberspace’. While poor quality output is not acceptable to translators, it is acceptable to most of the rest of the population. How long it will be acceptable is an open question; inevitably there will be expectations of improvement, and a challenge for the MT community must be the development of translation systems designed specifically for the needs of the Internet.

Implications for Professional Translation

Where do these developments leave the professional translator? It is plausible to divide the demand for translation into three main groups. The first group is the traditional demand for translations of publishable quality: translation for dissemination. The second, emerging with the information explosion of the twentieth century, is the demand for translations of short-lived documents for information gathering and analysis which can be provided in unedited forms: translation for assimilation. The third group is the demand for on-the-spot translation — the traditional role of the interpreter — which has taken a new form with electronic telecommunications: translation for interaction.

Translation for dissemination has been satisfied with mixed successes and frequent failures by the large-scale MT systems which are most familiar to translators. Cost-effective use of relatively poor quality output, which has to be revised by human translators, is difficult to achieve without some control of the language of input texts (at least for terminology consistency). It has been an option for only the largest multinational companies with large

volumes of documentation, which cannot be dealt with except by automating parts of their total documentation processes. In recent years translation workstations have offered a feasible and probably more attractive route for professional translators: translations of publishable quality can be made at higher productivity levels while maintaining translators' traditional working methods. In the future we can expect the majority of professional translators to be using such tools — not just from commercial expediency but from personal job satisfaction.

Translation for assimilation has not traditionally been undertaken by professional translators. The work has been done in organisations often by secretaries or other clerical staff with some knowledge of languages as an occasional service and usually under time pressures. Those performing the work have naturally been dissatisfied with the results, since they are not professionally trained. In this function MT has filled a gap since the first systems were available in the early 1960s. The use of Systran at the European Commission illustrates the value of such 'rough' translation facilities. This use exceeds by far its use for the production of translations for dissemination. It is believed that most of the use for the cheaper PC-based translation software is translation for information assimilation, mainly for personal use but sometimes within an organisation. Rarely, if ever, do professional translators see this output. Undoubtedly, there will continue to be a large and growing demand for this type of translation need — one which the translation profession as such has not been able to meet in the past.

Translation for interaction covers the role of translation in face-to-face communication (dialogue, conversation) and in correspondence, whether traditional mail or the newer electronic, more immediate, form. Translators have often been employed occasionally by their organisations in these areas, e.g. as interpreters for foreign visitors and as mediators in company correspondence, and they will continue to do so. But for the real-time translation of electronic messages it is not possible to envisage any role for the translator; for this, the only possibility is the use of fully automatic MT systems.

However, the very familiarity of MT systems will alert a much wider public to translation as a major and crucial feature of global communication, and probably to a degree never before experienced.

Inevitably, translation will itself receive a much higher profile than in the past. People using the crude output of MT systems will come to realise the added value (i.e. higher quality) of professionally produced translations. As a result, the demand for human produced translation will rise, and the translation profession will be busier than ever. Fortunately, professional translators will have the support of a wide range of computer-based translation tools, enabling them to increase productivity and to improve consistency and quality. In brief, automation and MT will not be a threat to the livelihood of the translator, but will be the source of even greater business and will be the means of achieving considerably improved working conditions.

References

For the history of machine translation see: W.J.Hutchins: *Machine Translation: Past, Present, Future*, Chichester (UK): Ellis Horwood (1986)

For a survey of current use of MT systems see: C.Brace, M.Vasconcellos and L.C.Miller: 'MT users and usage: Europe and the Americas', in *MT News International* No.12 (October 1995): 14-19

For a review of MT research see: W.J.Hutchins: 'Research methods and system designs in machine translation: a ten-year review, 1984-1994', in *Machine Translation Ten Years On*, International Conference, 12-14 November 1994, Cranfield University (in press)

For further details see: M.Flanagan: 'Two years online: experiences, challenges and trends', in *Expanding MT Horizons,* Proceedings of the Second Conference of the Association for Machine Translation in the Americas, 2-5 October 1996, Montreal, Quebec, Canada: 192-197

Machine-Aided Translation Tools for Slavonic Languages

by

Michael S. Blekhman, Andrei Kursin, Igor Fagradiants

Linguistica 93 Co., ETS Publishers Ltd

Introduction

You may already know how the automatic translation systems by Lingvistica '93 Co. work. They have been described elsewhere, including *Machine Translation Review* and *Language Today*. The present paper introduces you to another development area of our team, the POLYGLOSSUM-PARS family of machine-aided (computer-assisted human) translation systems. It is our goal to develop user-friendly dictionary look-up programs; to compile representative bi-directional dictionaries for such language pairs as Russian to and from English, German, French, Swedish, Finnish as well as Ukrainian to and from English; and to make these translation tools available to both professional translators and language students all over the world, either separately or together with the PARS automatic translation systems.

Within this framework the Moscow-based ETS Publishers developed the Polyglossum software and a family of general usage and specialist dictionaries. ETS invites leading Russian lexicographers to supply their most-up-to-date dictionaries in electronic form, after which they are converted into the Polyglossum format. ETS supplies these professional dictionaries on CD-ROM. The largest of them is the Russian-English-Russian Polytechnical Dictionary comprising over 300,000 translations in each part. Its printed-on-paper Russian-English analogue will consist of four volumes, three of which have already been issued by ETS.

Some dictionaries, such as those on medicine and computers, are converted from the PARS communication format.

Other dictionaries in this family are English-Russian-English dictionaries on mathematics (80,000 translations), ecology (40,000), oil and gas (75,000), a series of German-Russian, Swedish-Russian, and Finnish-Russian bi-directional dictionaries. The Polyglossum dictionaries comprise millions of translations in total.

Polyglossum 3.0, which was released in late 1997, operates in stand-alone and network modes under Windows 3.1, Windows 95, and Windows NT. The Top Key Cyrillic font support is supplied with the system. Polyglossum 3.0 supports the French-Russian technical dictionary of 60,000 translations in each part as well as two Russian dictionaries: one is the classical explanatory dictionary of the Russian language by Vladimir Dal, first published in Russia in the nineteenth century; the other contains Russian proverbs and sayings.

This paper describes the latest version of the Polyglossum dictionary support software — PG-PARS (Polyglossum-PARS). It was designed as a Windows 95 application to support English-Russian-English Polyglossum dictionaries as well as English-Ukrainian-English dictionaries. A conversion program is being developed to transform the Polyglossum dictionaries into the new format.

One of the main peculiarities of PG-PARS is the 'smart search' mode which is based on morphological analysis of Slavonic words found in the dictionaries. This is especially important if the user is not a native speaker of Russian or Ukrainian.

Another important feature is the Selection option which allows the user to mark a portion of the dictionary entry and paste it into the text.

Dictionary Structure

PG-PARS dictionaries are organized much like traditional printed-on-paper dictionaries.

A dictionary entry consists of the head word as well as its translations, phrases which comprise the head word, their translations, comments and examples. The head words are presented in alphabetical order. The word entry structure may vary slightly from dictionary to dictionary owing to the preferences of its author.

PG-PARS dictionaries are all bi-directional, i.e. they include, for example, both the English-Russian and Russian-English parts. Each part has its alphabetical index of entries displayed on a separate tab in the PG-PARS main window. The word entries are displayed for both parts of the dictionary, and translations of translations can be found easily, which is sometimes very important for a professional translator.

A conventional search method can be used. This method (so-called 'simple search') consists in looking up a single word or a key word of a phrase in the index and then examining the dictionary entry for it. Alternatively, the 'smart search' mode can be used (see below). The user selects the key word in the index box, presses Enter, and double-clicks the key word in the index box. When the text is typed in the Keyword box, the current position in the index is automatically adjusted.

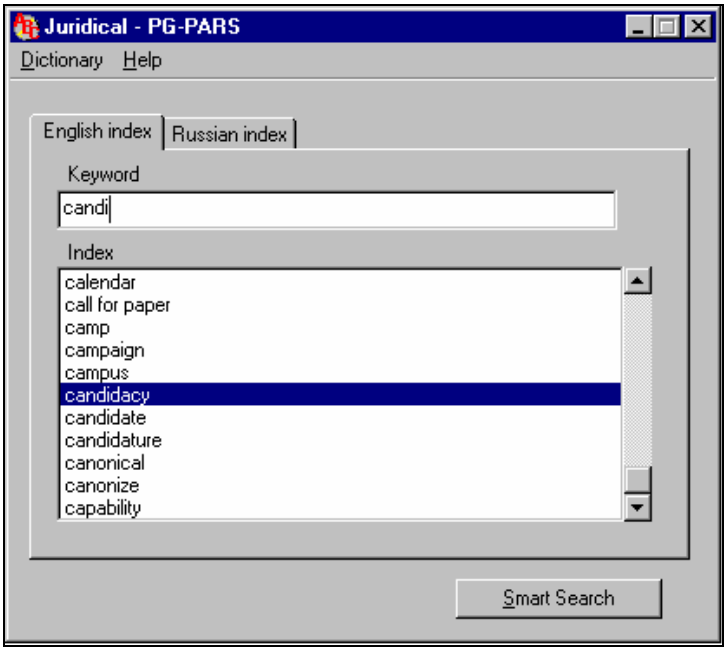

Fig. 1: English index

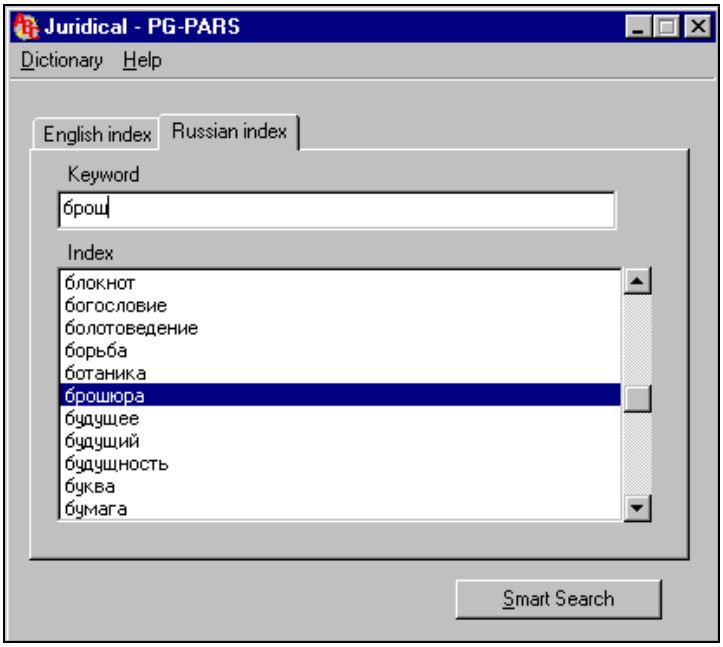

Fig. 2: Russian index

Smart Search

The smart search mode is one of the most important features of PG-PARS. In this mode, a word or phrase is searched regardless of the inflection form in which the query is specified. It allows the user to start the search operation by simply dragging or copying/pasting a word from the source text to the dictionary window.

If the 'smart search' operation succeeds, the corresponding entry is displayed in the Dictionary Entry window. Otherwise, simple search on the query is performed, i.e. the text of the query is placed in the Keyword box and the index position is adjusted accordingly. The 'smart search' mode may fail because of the absence of the word or phrase in the dictionary or due to certain homonymy of word forms or word endings in the case of single word search. However, our experiments show that search precision in the smart search mode is rather high: 95% for Russian and 94% for Ukrainian. This means that purely linguistic mistakes are rare.

To initiate 'smart search', the user drags the text to be found from another application and drops it anywhere in the PG-PARS window outside the Keyword box. Alternatively, when in the PG-PARS main window, he switches to the Index box and pastes the text to be found from the clipboard (Shift+Ins). He can also type the text in the Keyword box and click the Smart Search button.

Searching for Phrases

This can be done either in the smart search mode or by examining the dictionary entry for the key word of the phrase.

To open the entry, the user either types the key word or its beginning in the Keyword box of the PG-PARS main window or double clicks the corresponding element in the Index box if it has already been selected. The dictionary entry will be displayed.

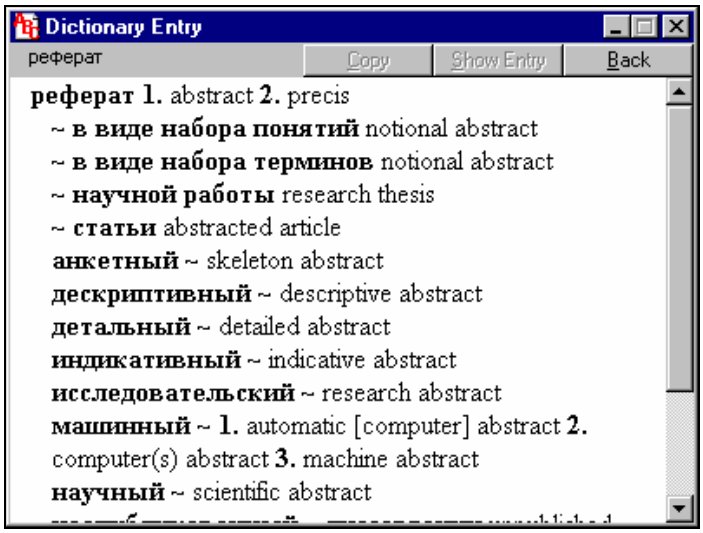

Fig. 3: Dictionary entry (Russian-English)

Instead of typing the keyword the user can paste it from the clipboard or drag it from the source text and drop it onto the Keyword box.

Selecting Text in the Dictionary Entry.

Another peculiarity of PG-PARS is the option of flexible selection of elements in a dictionary entry (such as translation, source phrase, etc.) by double-clicking it. Pressing Tab or Shift+Tab moves the selection to the next or previously visible selectable element of the entry.

An element can have variants given in square or curly brackets. To select the variant including the text in brackets, the user double-clicks inside the brackets.

Here is an example:

- bring to true [uniform] surface
  double-click on *bring* selects *bring to true surface*;
  double-click on *uniform* selects *bring to uniform surface*.
- tow(ing) rope
  double-click on *tow* selects *tow rope*;
  double-click on *ing* selects *towing rope*.

It is also possible to select a portion of an entry element by pressing the left mouse button and then dragging the mouse. In this mode, the user can select italic text.

Selecting an element is intended for copying the translation to another application. That is, when typing the text, the user selects a (the) translation of the unknown word and pastes it into the text.

When in the Dictionary Entry window, the user selects an element of the entry and clicks the Copy button to put the selected text into the clipboard. He can also send the selection to the target application.

If a tilde is present in the selection, it will be automatically replaced with the word it substitutes. Brackets, semicolons in the ends of translations, etc., are not copied.

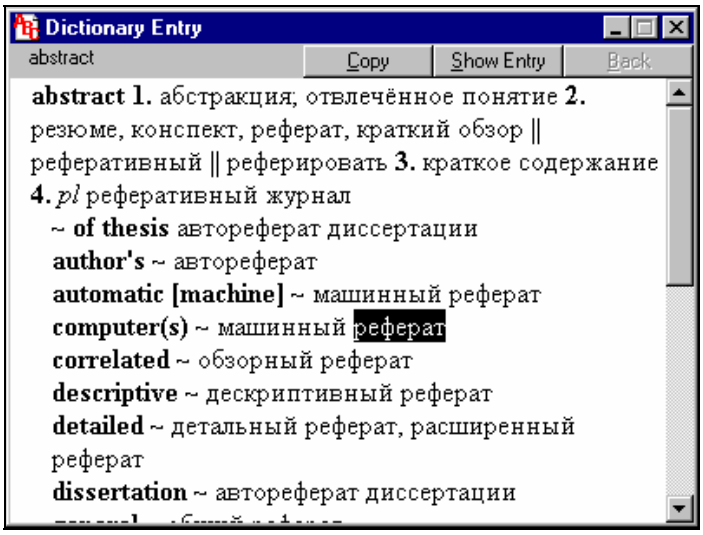

Fig. 4: Selecting text in a dictionary entry (Russian part)

Many professionals find it useful to examine back translations of translation variants before choosing one of them. When in the Dictionary Entry window, the user selects the translation and clicks the Show Entry button, after which the entry for the text selected is displayed. The user can also return to the previously displayed entry by clicking the Back button.

From the technical point of view, it should be pointed out that PG-PARS supports the Drag&Drop and Plug&Play protocols. The user can drag text to be found in the dictionary to the PG-PARS main window and drop it

- onto the Keyword box to replace text in it and adjust the index position accordingly
- or anywhere else within the window to start 'smart search'.

It is also possible to drag the text selected in the Dictionary Entry window to another application.

The dictionaries of the Polyglossum family are supplied in Russia by ETS Publishers. In North America, they are distributed by the Virginia-based POLYGLOSSUM, Inc.

Our next steps in the machine-aided translation area will be to develop Polyglossum versions of English-Russian and English-Ukrainian dictionaries for all subject areas (science, technology, business, general-usage words,) and to introduce translation memory facilities into PG-PARS. In the long run, PARS and Polyglossum will be transformed into a hybrid memory-based MT/MAT system equipped with a special user-oriented text editor, the latter having specific editing functions (such as 'Insert article', 'Transpose words', etc). We will be happy to inform the readers of our new developments!

Understanding Commercial Machine Translation Systems for Evaluation, Teaching, and Reverse Engineering: the Treatment of Noun Phrases in Power Translator Deluxe

by

Mario J. Mira i Giménez and Mikel L. Forcada

Department de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

1 Introduction

In the last decade, a number of low-prices PC-based machine translation (MT) systems have been released to the market. In most personal computing magazines, these low-end commercial systems are reviewed in terms of ease of use, user-friendliness and integration with other software (e.g. wordprocessing packages). Examples of this approach are Myers' reviews of Japanese-English translation programs (Myers 1996a,b). However, the strengths and weaknesses of such systems in terms of translation quality are often given only in terms of percentage accuracy or by showing some token examples of incorrect translation; there is not usually discussion of the translation results in connection with what would be expected from a particular MT strategy. This last approach would be intimately related to *diagnostic evaluation*, of which Hutchins (1996) says 'is the concern mainly of researchers and developers'[1], but could also, in our opinion, be of great utility to people outside the research and development community, because it may help to identify patterns and consistencies in observed problems[2], and, therefore, be an invaluable companion to corpus-based or test-suite-based evaluation (Lewis 1997).

It is also the case that software companies producing these systems seldom give any details of the proprietary MT strategies used in their programs. However, it is our belief that even a partial understanding of the translation strategies used by commercially available MT programs may be very important in the following three areas:

- When evaluating a MT system prior to adoption in a business environment. A good understanding of the MT strategies used by a given system may be used to define, for example, what to avoid in source documents (pre-edition) in order to minimize the revision (post-edition) of machine-translated documents. In a 'machine translation for dissemination' setting, this could even lead to the adoption of a controlled source language suited to the particular MT system (for more on controlled languages in MT, the reader is referred to Newton (1992), Arnold et al. (1994), Cole (1996), and Huijsen (1998)).

[1] Hutchins (1996) distinguishes three main types of evaluation: *adequacy*, *performance*, and *diagnostic* evaluation.

[2] Some companies specialize in selling detailed reports on the evaluation of machine translation systems; a quick internet search reveals two examples: Data Research DPU ab of Lidingö , Sweden (www.dpu.se), sells a report on translation technology products for U.S. $2,200; Multilingual Technology Ltd., of Harpenden, U.K. (www5.red.net/homepages/mtl) offers similar language-engineering consulting services.

- When teaching a course on machine translation, a partial understanding of the MT strategies used by commercial programs may be used to illustrate them in the laboratory using low-cost, off-the-shelf software, and to teach students to evaluate MT systems in a more systematic way, by meaningfully using hypotheses on their internal workings.[3] The issue of MT evaluation in an educational setting has been recently discussed by Lewis (1997).
- Obviously, when designing a new MT system, the understanding of the MT strategy used in existing MT systems may be part of a reverse-engineering or design recovery effort which may be aimed at discovering problems and flaws that may be solved in a new MT product.

The knowledge about the translation strategies used by a particular product has to be obtained by means of what may be called 'black box' modelling: analyzing inputs and outputs, hypothesizing the nature of the translation strategy used, and inferring the particular rules applied to each sentence.

We will focus on a low-priced PC-based MT system that was apparently released in 1995, Power Translator Deluxe 1.0 for Windows (PTDS) by Globalink Inc. (URL: www.glabalink.com). The English-Spanish version of this program was available as a CD-ROM in newsstands and bookstores in Spain in 1996 for less than U.S. $40 (4,995 Spanish pesetas).

In particular, we will study the treatment of noun phrases (NPs) when translating from English to Spanish. Noun phrases are especially interesting because of the radical differences in word order existing between English and Spanish. In particular, we will only study NPs containing adjectives (*a*), nouns (*n*), and the particle (''s') indicating the possessive case (*s*). Our analysis will not try to study the rules used to add definite articles in the Spanish translation of the NP, and will avoid words that may belong to more than one morphological category; that is, we will use words that are only nouns or adjectives in the dictionary of PT.

Comment [d1]:

2 *Black-Box Modelling of Noun-Phrase Translation by Power Translator*

This section gives details of the black-box modelling approach used to obtain a hypothesis for the strategy used by PT when translating noun phrases from English to Spanish and to infer the collection of rules used. The incremental approach used is informally described, but enough detail is given so that it may be reproduced.

2.1 *Determining the Translation Strategy*

A preliminary study of some NPs, especially long NPs, revealed that PT was unable to identify them and manipulate them as a whole and often seemed to reorder and translate correctly parts of the NP that could be interpreted as smaller NPs, leaving other portions incorrectly translated and disconnected. This clearly suggested that we should discard an analysis based on a syntactical transfer strategy (Hutchins and Somers 1991), which, in principle, would be able to identify and manipulate NPs and their constituents regardless of their length. Therefore, we hypothesize that only a shallower (i.e. morphological) level of analysis occurs: this is indeed the main hypothesis of our work. The system indeed appears to use a very simple strategy to translate NPs, which may be classified as a morphological

[3] Our experience in a Computers for Translation course confirms the interest of this approach.

transfer or so-called 'direct' MT strategy (Hutchins and Somers 1991) with local morphology-driven word reorderings (reorderings are necessary because of the differences in the syntax of NPs in English and Spanish).

The word reordering strategy is based on a collection of rules that recognize a given pattern of morphological categories and reorder it in order to render a correct Spanish translation. For example, one of the rules reads

$$a\ n \rightarrow n\ a$$

which may be interpreted as follows: if an English NP made of an adjective and a noun is detected, then its translation will be a Spanish NP having the corresponding noun and the corresponding adjective, in reverse order. The rule may only be applied once the source sentence has been morphologically analysed so that all of its words are assigned a given category. (When a word belongs to more than one category as in the case of *truck*, which may be a noun or a verb, the system appears to use a set of rules in which the immediate morphological context of the word is used to determine the category that will be subsequently used.) Also, we hypothesize that, once a word has been involved in a reordering, it cannot be involved in another reordering; that is, reordering patterns do not overlap. An open question remains: what happens when two different NP patterns match a given word sequence? Which one is used for reordering? Our initial hypothesis is that the NP pattern matching the *longest* word sequence in the sentence, starting from its left, will be used for translation. This hypothesis is found to hold in all but a few exception cases.

Gender agreement in the resulting Spanish output seems to occur after reordering and assigns the gender of a noun to all adjectives immediately to the right of it, regardless of whether or not they were involved in the reordering(s).

In view of the recursive nature of language, NPs may in principle be of any length, because a complete NP may be part of an NP: thus, for example, the NP 'tall computer scientist's desk' includes the NP 'computer scientist's desk'. That is, a finite collection of rules like the one given above is unable to capture all possible NPs. As a consequence, such a translation strategy will fail for NP patterns which have not been considered.

In this paper, we will use the above hypotheses to try to determine the collection of rules used by PT for noun phrases and the way they are applied.

Other PC-based MT systems that, in preliminary analyses by the authors, appear to use morphological transfer strategies with local reordering of noun clauses are Transparent Language's Transcend,[4] and Softkey's Translator Pro.[5] Other MT systems such as those in MicroTAC's Language Assistant Series[6] (for example, Spanish Assistant) seem to use a very different strategy which may be classified as a kind of syntactical transfer.

2.2 Inferring the Rules

[4] Transparent Language is in Hollis, NH (www.transparent.com).
[5] This software seems to have been discontinued and it is severely affected by translation errors. Softkey International is in Marietta, GA.
[6] Currently distributed by Globalink, Inc.

To define the collection of active reordering rules, we picked a vocabulary of words that could only act as adjectives or nouns (no categorical lexical ambiguity in PT's dictionary) in order to construct a set comprising all possible English NP patterns up to a certain number of words. Then, their translations to Spanish were analysed for reorderings and gender agreement, and word reordering rules were proposed to account for the results observed. New reordering rules were postulated only when the result or a given NP could not be explained with rules found for shorter NPs.

We could not find any active reordering rule for source word patterns longer than six words (note than when counting words, the particle used for the possessive case in English is taken as a separate word (category s)). When a reordering rule does not cover the complete NP, the part not involved in the reordering either stays the same or is reordered by a different rule.

The rules for deciding which rule is applied when more than one matches the input pattern were inferred from a systematic set of longer NPs designed specially for this purpose.

2.3 Results

When translating from English to Spanish, PT appears to use the following noun-phrase reordering rules, in order of increasing length of the source word pattern subject to reordering:

1. $a\ n \rightarrow n\ a$
2. $n_1\ n_2 \rightarrow n_2$ 'de' n_1
3. $a_1\ a_2\ n \rightarrow n\ a_2\ a_1$
4. $a\ n_1\ n_2 \rightarrow n_2\ a$ *'de'* n_1
5. $n_1\ a\ n_2 \rightarrow n_2\ a$ 'de' n_1
6. $n_1\ n_2\ n_1 \rightarrow n_1$ 'de' n_2 'de' n_1
7. $n_1\ s\ n_2 \rightarrow n_2$ 'de' n_2 'de' n_1
8. $\mathrm{a}_1\ a_2\ a_3\ n \rightarrow n\ a_3\ a_2\ a_1$
9. $\mathrm{a}_1\ a_2\ n_1\ n_2 \rightarrow n_2\ a_1$ 'de' $n_1\ a_2$
10. $n_1\ n_2\ n_3\ n_4 \rightarrow n_4$ 'de' n_3 'de' n_2 'de' n_1
11. $n_1\ s\ a\ n_2 \rightarrow n_2\ a$ 'de' n_1
12. $\mathrm{n}_1\ s\ n_2 \rightarrow n_3$ 'de' n_2 'de' n_1
13. $a_1\ a_2\ a_3\ n_1\ n_2 \rightarrow n_2$ 'de' $n_1\ a_3\ a_2\ a_1$
14. $n_1\ n_2\ n_3\ n_4\ n_5 \rightarrow n_5$ 'de' n_4 'de' n_3 'de' n_2 'de' n_1
15. $a_1\ a_2\ a_3\ a_4\ a_5\ \ n \rightarrow n\ a_5\ a_4\ a_3\ a_2\ a_1$

The rules are formulated without taking into account the rule-based addition of definite articles in the Spanish translation. It is quite interesting to note certain inconsistencies in the pattern set, such as the absence of a pattern for four adjectives preceding a noun when patterns exist for three and five adjectives preceding a noun (patterns 8 and 15). We have not found any case in which the behaviour of PT on a NP could not be explained in terms of the above patterns.

In our experiments we have found that, in general, the rule reordering the longest pattern matching the current NP (starting on the left of the pattern) is used for translation; the remainder of the NP may be reordered by other rules (reordered areas do not overlap). However, we have found a few cases in which the general rule (longest pattern from the left) is not applied. Here are two examples:

- The case of rules 6 and 10. The general rule in view of a four-noun pattern *nnnn* would be to use rule 10. However, when the sequence is preceded by an adjective (*annnn*), the system seems to prefer rule 6, which it applies to the first three nouns, leaving the adjective and the last noun in their original places. We cannot easily explain this case without invoking some effect of context around the reordered pattern.
- The case of pattern *annsan*. Sequences following this pattern, such as 'senior computer expert's large desk', are translated by applying first rule 1 to the left part of the NP and then rule 11 on the right part instead of using for example, rule 4 (which matches a longer pattern) on the left and then perhaps rule 1 on the right part.

The fact that we have not found an explanation for some cases where the rule of the longest pattern matching the sentence from left to right is violated does not completely invalidate the collection of reordering rules found to be used by PT: we find the rules can still be used to analyse the incorrect results produced, even in the mentioned conflicting cases, simply by relaxing the rule of the longest pattern.

3 Checking the Black-Box Model against PT's Internals

After doing the black-box modelling described in the previous section, we set out to find whether we could find some representation of the reordering patterns inside the program or its auxiliary files. We were rather surprised to find that word reordering rules are stored in a text (ASCII) file, revealingly named **engspan.pat** in directory 'dicts' (the corresponding file for Spanish to English is **spaneng.pat**). The **engspan.pat** file is identical in both versions of PT.

A careful analysis of the contents of **engspan.pat** (which would have surely been much more difficult without the knowledge built during the black-box modelling) shows that the file is an ordered list of reordering rules. Each word-reordering rule takes two lines in the file. The first line starts with the number of words in the pattern, followed by a list of one- or two-letter codes representing the sequence of word categories in the pattern (**A** for adjective, **N** for noun, etc.; we have not yet been able to decipher all codes). The second line contains a sequence of integer numbers representing, for each position of the reordered pattern, the relative position of the word in the original pattern: **O** means the same position, **-2** means two positions to the left, and **+1** means two positions to the right (a code of **99** means that the position in the reordered pattern is left blank). If the integer is appended a **D** this means that the Spanish preposition should be prepended to the word in the corresponding position of the *original* pattern, *before* reordering. This format was determined by comparing the file to the set of 15 rules proposed in section 2.3.

For example, the format used in the file for rule 4 (found in lines 171 and 172 of **engspan.pat**) is:

3	**A**	**N**	**N**	**X**
	+2	**-1D**	**-1**	**X**

where the integers in the second line have the following meaning: **+2** means that the first word in the reordered pattern is taken from a position in the original pattern two positions to the right of it (the second **N**); **-1D** means that the second word in the reordered pattern is taken from the first position in the original pattern (the adjective), the **D** meaning that, prior to reordering, the preposition *de* is prepended to the second word in the original pattern; finally, **-1** means that the third word in the reordered pattern is taken from the second position in the original pattern (the first **N**). The **X** is used as an end-of-pattern marker.

We found the 15 rules hypothesized in section 2.3 to be present in file **engspan.pat**, together with rules for more general NPs (containing adverbs, numerals, conjunction, etc.) and for other constructions (about 100). One difference between the hypothetical rules and the ones actually used is that nouns in the possessive case are treated as a single word coded **NG** (with **G** presumably standing for *genitive*) and not two separate words, as described in section 2.3. This varies the length of some patterns and solves the apparent inconsistency found for sentences such as 'senior computer expert's large desk' in section 2.3 (**NG** is not the same as **N** in a pattern). Also, the problem observed for sequences of the form *annnn* is easily explained now. The program has a rule of the form

$$a\ n_1\ n_2\ n_3 \rightarrow a\ n_3\ \text{'de'}\ n_1$$

which we did not consider during black-box modelling because we believed that the adjective was not involved in the reordering (translations produced by applying this rule sound strange in most cases in Spanish because of the initial position of the adjective).

The rules in the file are given in order of decreasing length, which is compatible with the rule of the longest pattern matching the sentence from left to right.

The analysis of the rules present in the file confirms the main hypotheses about the translation strategy obtained after the black-box modelling; roughly, PT seems to work as follows:

1. Morphological analysis takes place and a word category is assigned to each word in the sentence (this may be done in advance for the whole sentence or word by word, during the execution of the following steps). This corresponds to the *analysis* part of a transfer system. The *transfer* part follows:
2. The list of word categories is processed from left to right so that the rule matching the longest pattern present in the sentence is chosen.
3. The sentence is reordered by the rule chosen.
4. Processing continues immediately to the right of the reordered sequence, in step 2.
5. After the whole sentence has been reordered, Spanish words are substituted for each English word in the sentence — this corresponds to the *generation* in a transfer system. The word substitution may also occur in parallel to the transfer steps.

3.1 Modifying the Behaviour of Power Translator

In view of the fact that reordering rules are stored in a text file which is presumably read only once when the program is activated, we decided to see if we could add new patterns to Power Translator to correct its behaviour for some NPs.

As was mentioned in section 2.3, a class of NPs that were not correctly translated follows the pattern *annsan* or **A N NG A N**, due to the lack of a rule matching this pattern. We decided to add the following two lines to **engspan.pat** just after all of the 5-word patterns:

5 A N NG A N X
+4 +2D OD -3 -3 X

After this modification, PT translated 'senior computer expert's large desk' as 'el escritorio grande de experto major de computadora', which is more acceptable than the result 'la computadora major escritorio grande de experto'. We have checked that PT does not drop the last rule if a new rule is inserted in the list (it apparently reads the whole **engspan.pat** file).

Adding reordering patterns to PT's files to correct some inadequacies in translation may be a very interesting exercise in an advanced machine translation course. It may also improve the performance of the software when a word pattern not considered appears frequently in the texts that have to be translated in a given installation.

4 Concluding Remarks

We have described a black-box modelling process in which, first, a hypothesis regarding the translation strategy of a system may be made and then this hypothesis may be used to infer the collection of translation rules used by the system. In particular, we have determined the set of rules used by Globalink's Power Translator Deluxe 1.0 and Power Translator Professional 5.0 for Windows (English/Spanish), two low-priced PC-based machine translation systems, to translate noun phrases (NPs) containing nouns, adjectives and the possessive case particle ('s') by hypothesizing that the system uses a strategy of morphological transfer supplemented by a morphology-driven, local word reordering. The black-box modelling suggests that Power Translator (PT) uses a general rule (the rule of the longest pattern from left to right) to select which reordering is applied when more than one reordering is possible, a rule which appears to hold in all but a few exceptional cases.

As in any black-box analysis, one cannot claim that the strategy and rules inferred are actually used in the product we have analysed; however, we may state that they are consistent with the observed translation results, and, therefore, may be used as a tool to analyse the behaviour of the MT system. The black-box modelling process described in this paper may be applied to constructions other than NPs and to other PC-based MT systems which seem to apply a very similar strategy.

The validity of the hypotheses on the nature of the MT strategy and the particular rules inferred during the black-box modelling process have been confirmed, with minor changes in the rules, after examining one of the auxiliary files used by PT, which happens to be easily readable and interpretable.

Having such a detailed hypothesis on the internal workings of a PC-based MT system may be of great interest in three different application fields: evaluation of a MT system prior to adoption, illustration of translation strategies in a MT course, and reverse engineering of existing MT systems when designing a new one.

Acknowledgements:

The financial support of the Spanish *Comisión Interministerial de Ciencia y Tecnología* (CICyT) through project TIC97-0941 is gratefully acknowledged. We also thank J. D. Ortiz-Fuentes and R.C. Carrasco for suggestions.

References

Arnold D., Balkan, L., Meijer, S., Humphreys, R. L., Sadler L. (1994) *Machine Translation: an Introductory Guide*, London: Blackwells-NCC

Cole, R.A. (ed.) (1996) *Survey of the State of the Art in Human Language Technology*, URL:www.cse.ogi.edu/CSLU/HLTsurvey/

Huijsen, W. O. (1998) 'Controlled Languages Homepage',

URL:wwwuilots.let.ruu.nl/Controlled-languages/

Hutchins, W. J. (1996) 'Evaluation of Machine Translation and Translation Tools', in Cole, R.A.

Hutchins, W. J., and Somers, H. L. (1991) *An Introduction to Machine Translation,* Academic Press: London

Lewis, D. (1997) 'MT Evaluation: Science or Art' in *Machine Translation Review*, No. 6: 25-36

Myers, S. (1996a) 'Can Computers Translate?', in *Computing Japan*, April 1996. URL:http://www.japanese.com/library/publications/mt.html

Myers, S. (1996b) 'English-to-Japanese Translation Programs: Can They Do the Job?', in *Computing Japan*, August 1996. URL:http://www.japanese.com/Publications/e2j_mt.html

Newton, J. (1992) 'The Perkins Experience', in Newton, J., (ed.), *Computers in Translation: A Practical Appraisal,* London: Routledge

Wojcik, R. H., Hoard, J. E. (1996) 'Controlled Languages in Industry', in Cole, R.A.

Book Review

D. Jones and H.L. Somers (eds) (1997) *New Methods in Language Processing*, *Studies in Computational Linguistics*, London: UCL Press. Hardback xiii + 369p. £45. ISBN 1-85728-711-8.

This book is a selection of papers given at the conference of the same name that took place at UMIST, Manchester, in 1994. The papers describe the application of different methods to several areas of natural language processing, reflecting the influence that methods which have originally been developed in other fields have on those areas. With language being very much in the centre of most activities that involve human interaction or information processing it is not surprising that people from many different disciplines end up trying to solve language processing problems; and this brings with it a refreshingly varied collection of 'new' methods with their own advantages and disadvantages over established ones.

Having to decide on a primary criterion for arranging this colourful mixture of papers, the editors chose to group them according to method. This results in seven parts, ranging from analogy-based methods to methodological issues.

In the first chapter, three papers on analogy-based methods describe how machine learning techniques are applied to different problems in order to learn regularities from a set of given examples and then to generalise from those for dealing with new data. The general conclusion is that good performance has to be paid for by higher computational costs than other machine-learning algorithms, and that the computer groups linguistic entities according to other features than humans would do. It is probably not very surprising that, for a computer, surface characteristics such as character sequences seem more interesting than the implicit morphological features that are realised with those characters. However, since no claim is made on the cognitive adequacy of these methods, this does not lessen the good classification results achieved.

The main problem of connectionist NLP is how to map a (usually long) linear sequence of symbols onto a few input nodes, i.e. the transition from symbolic to subsymbolic processing. The two main solutions chosen in the articles in the respective section are the use of symbolic pre-processing modules to extract a set of features from the input data, using recurrent networks. Feature extraction reduces the data to a small set of parameters relevant to the investigation. This can, for example, involve pragmatic and syntactic features used in a hybrid abstract generation system. While most neural networks require the whole input pattern to be present, recurrent networks allow a sequential processing mode, where a part of the pattern is presented at each step, and while the pattern is shifted at the next time slice, the previous pattern is fed back into the system through an internal feedback-loop. This way a trace of past steps is still present, and temporal events can be captured.

While neural networks are mainly used in pattern recognition or classification tasks, the papers in this section describe attempts to utilise them for linguistic tasks like parsing, or recognition of higher level features such as semantic relationships.

The next chapter on corpus-based methods, as with the following one on machine translation, does not quite fit into the editors' grouping. Corpora are the objects to which new

methods are applied; they do not specify the methods themselves. These methods are quite varied, ranging from machine learning and parts-of-speech tagging, to terminology extraction and the automatic identification of sublanguages. Again, as in the other chapters of this book, one can benefit from the description of the new methodology, which can be applied to a lot of other areas. Corpus-based methods are especially well suited for NLP, as they are based on real data. In a way it would be appropriate to label corpus-based language processing 'applied NLP'.

After the failure of the rather ambitious early attempts of machine translation, a fully automatic translation has been abandoned as the main goal of work in that field. Instead, smaller targets have been set which can be dealt with quite successfully in sub-areas like example-based machine translation. Many of these methods are based on parallel corpora, where automated methods are used in order to align expressions of varying length (like compounds which are written as one word in German, but two (or more) in English), or to learn transfer functions. Additionally, sub-symbolic approaches to MT are tried out, like the use of neural networks. This works quite well, but unfortunately only with very restricted vocabulary and a fixed sentence structure.

The statistical approaches would also fit into the corpus-section; they are mainly concerned with parsing and tagging. Two of the text analysis problems being investigated are PP-attachment (or rather the evaluation of it) and anaphora resolution, and there also is an interesting paper on applying techniques of information retrieval to the analysis of Japanese discourse structure.

The section on hybrid approaches features more parsing, both object oriented and stochastic, as well as evolutionary methods applied to optimising a dialogue system. The main problem here — as it is with the connectionist systems — is the interfacing of symbolic processing to non-symbolic methods.

The final chapter on methodological issues features just one paper. This focuses on the design of NLP systems, especially on software re-use, which could save a lot of development effort in NLP projects, but does not really receive the amount of attention that it deserves.

Altogether, this is a book well worth reading. It describes a variety of different methods which can be applied to NLP, and at the same time provides an overview of current problems in the area. It demonstrates how valuable fresh ideas are for a discipline, especially one that overlaps with as many others as natural language processing.

Oliver Mason, Corpus Research, Department of English, School of Humanities, The University of Birmingham, Edgbaston, Birmingham B15 2TT, O.Mason@bham.ac.uk

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).

2–3 April 1998
EAMT: European Association of Machine Translation Workshop
Geneva, Switzerland
Tel: +41 22 791 2317, fax: +41 22 791 3995, e-mail: pasteuro@who.ch
http://www.lim.nl/eamt

21–22 May 1998
CLAW98: The Second International Workshop On Controlled Language Applications
Pittsburgh, PA. 15213 USA
http://www.lti.cs.cmu.edu/CLAW98/

26–30 May 1998
First International Conference on Language Resources and Evaluation
Granada, Spain
Tel: +34 58 24 41 00, fax: +34 58 24 41 04, e-mail: reli98@goliat.ugr.es

2 June 1998
EMNLP3: 3rd Conference on Empirical Methods in Natural Language Processing
Granada, Spain
Tel: +1 914 437 5988, fax: +1 914 437 7498, e-mail: ide@cs.vassar.edu
http://www.cs.vassar.edu/~ide/emnlp3.html
http://www.cs.jhu.edu/~yarowsky/sigdat.html

29 June – 1 July 1998
FSMNLP'98: International Workshop on Finite State Methods in Natural Language Processing
Ankara, Turkey

13–24 July 1998
ELSNET's 6th European Summer School on Language and Speech Communication.
Robustness: Real Life Applications in Language and Speech
Barcelona, Spain
Fax: +34 3401 6447, e-mail: summer98@gps.tsc.upc.es
http://gps-tsc.upc.es/veu/ess98/

19–23 July 1998
TESS: The Text Encoding Summer School
Oxford University, Oxford, UK
E-mail: hcu@oucs.ox.ac.uk
http://users.ox.ac.uk/~tess/

24–27 July 1998
TALC98: Teaching and Language Corpora 1998
Keble College, Oxford, UK
E-mail: talc98@oucs.ox.ac.uk
http://users.ox.ac.uk/~talc98/

1–4 August 1998
DAARRC2: Colloquium on Discourse, Anaphora and Reference Resolution
Lancaster University, UK
Fax: +44 1524 843 085, e-mail: eiaamme@msmail.lancaster.ac.uk

5–7 August 1998
9th International Workshop on Natural Language Generation
Niagara-on-the-Lake, Ontario, Canada
Tel: +1 519 888 4443, e-mail: cdimarco@logos.uwaterloo.ca

10–14 August 1998
COLING-ACL98: 17th International Conference on Computational Linguistics
36th Annual Meeting of the Association for Computational Linguistics
University of Montreal, Canada
E-mail: coling-acl98@iro.umontreal.ca
http://coling-acl98.iro.umontreal.ca

15–16 August 1998
6th Workshop on Very Large Corpora
University of Montreal, Quebec, Canada
e-mail: ec@cs.brown.edu
http://coling-acl'98.iro.umontreal.ca

16 August 1998
ACL/COLING98: Translingual Information Management. Current Levels and Future Abilities
University of Montreal, Quebec, Canada
http://www.cs.vassar.edu/~ide/translingual.html

16 August 1998
ACL/COLING98: Partially Automated Techniques for Transcribing Naturally Occurring, Continuous Speech
University of Montreal, Quebec, Canada
E-mail: trans98@cs.concordia.ca

17–21 August 1998
ESSLLI98: European Summer School in Logic, Language and Information Workshop on Recent Advances in Corpus Annotation
http://www.dcs.warwick.ac.uk/~esslli98/workshops.html

23–28 August 1998
ECAI98: 13th Biennial European Conference on Artificial Intelligence

University of Brighton, UK
E-mail: Lynne.Cahill@cogs.susx.ac.uk
http://www.cogs.susx.ac.uk/ecai99/

24–28 August 1998
ESSLLI98: Workshop on Machine Translation
Saarbrücken, Germany
http://www.dcs.warwick.ac.uk/~esslli98/workshops.html

7–9 September 1998
Literature, Philology and Computers
Tel. 44+131-6503646 Fax: 44+131-6506536
E-mail: itadfp@srv0.arts.ed.ac.uk
http://www.ed.ac.uk/~esit04/italian.htm

23–26 September 1998
TSD98: Workshop on Text, Speech and Dialog
Brno, Czech Republic
http://www.fi.muni.cz/tsd98/

5–7 October 1998
KONVENS98: Computers, Linguistics, and Phonetics between Language and Speech, 4th Conference on Natural Language Processing
University of Bonn, Germany
http://www.ikp.uni-bonn.de/Konvens98

30 November–4 December 1998
ICSLP98: 5th International Conference on Spoken Language Processing
Sydney Convention Centre, Sydney, Australia
E-mail: icslp98@tourhosts.com.au
http://cslab.anu.edu.au/icslp98

August 1999
ESSLLI-99: Eleventh European Summer School in Logic, Language and Information
Utrecht, The Netherlands

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.7, April '98) 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.