INTERNATIONAL JOURNAL OF TRANSLATION
Vol. 15, No. 1, Jan.-June 2003

Relevance of Parallel Corpora to the Latest Developments of Machine Translation and Computer-Assisted Translation

FEDERICO GASPARI
Centre for Computational Linguistics

1. INTRODUCTION

MACHINE TRANSLATION AND COMPUTER-ASSISTED TRANSLATION

At the outset of this discussion on the practical impact of parallel corpora on machine translation (MT) systems and computer-assisted translation (CAT) tools a clarification is necessary to avoid confusion or ambiguity: machine translation is solely understood here in the strict sense of the fully automatic process in which the computer actually does the translating, and at most humans are entrusted with the supervision of what it produces (e.g. giving feedback to designers and software engineers to enhance the robustness of the MT system, updating the internal dictionaries, lexicons and terminological components), or with minor interventions taking place before, during or after the automated translation task performed by the machine.

These procedures, such as pre-editing of the source text, interactive use of MT systems and post-editing of the raw output, are typically aimed at maximising the quality of the resulting text, so as to guarantee its readability in the target language, thus avoiding serious hindrances to general understanding. In summary, machine translation is here intended as a process or activity which is automated to such an extent that it almost completely relies on the computer's performance, and accordingly the role of humans is heavily dependent on the product of the system. In this case the autonomous activity of the computer affects to a predominant extent the final translation.

As a result, in the interest of clarity in this discussion the notion of machine translation will be kept clearly distinct from that of computer-assisted translation, since the latter includes a wide range of tools that offer support to human translators, who can avail themselves of these resources to work in a semi-automated environment. By integrating CAT tools into their working routine, humans nevertheless continue to play the leading role, while they are helped by the computer to avoid some of the

most repetitive and tedious tasks entailed by translating, which can be successfully managed if CAT software is there to assist the professional translator.

Bearing in mind that pure machine translation and computer-assisted translation are separate entities, and in principle they underlie very different approaches to the activity of translation, they will be explored alongside each other in this paper, by considering with particular attention the practical relevance of parallel corpora as an ideal link between them.

2. BACKGROUND
CORPUS-BASED APPROACH TO FULLY AUTOMATIC MACHINE TRANSLATION

The literature devoted to machine translation of the last couple of decades puts increasing emphasis on the growing influence of corpus linguistics on innovative approaches to machine translation[2]. One of the most recent paths to the implementation of fully automatic machine translation systems in fact relies on the exploitation and manipulation of large collections of bilingual texts, whose aligned and matched sentences are used to provide the machine-translated text in the target language, i.e. the output.

This strategy, which is called example-based machine translation (EBMT), represents a radical deviation from the conventional rule-based approach to machine translation (RBMT), and it was for the first time put forward in the early 1980's by the Japanese researcher Makoto Nagao. He outlined the basic sequence of steps taken by EBMT systems, namely matching strings of text in the source language against a collection of real examples in the target language, pairing the corresponding translation fragments, and finally rearranging them to give the output in the target language (Nagao 1984; cf. also Somers 1998:23).

When this corpus-based approach to machine translation systems was first proposed, a lot of emphasis was laid on the lack of explicit linguistic knowledge in their design, which was superseded by a parallel aligned corpus of translation examples representing their most significant component. Nagao's intuition relied on the principle that it is possible for computers to handle and re-combine large sets of real examples of language to come up with a translation, thus suggesting for the first time the idea of what is now known as example-based machine translation. A few years after Nagao's original proposal, many researchers realised the potential interest and feasibility of this approach, and in the 1990's there was quite a lot of intensive work going on to investigate its possible implementations and actual potential.

3. MAIN FEATURES OF EXAMPLE-BASED MACHINE TRANSLATION (EBMT)

Pure example-based machine translation systems, which are to a large extent developed for research purposes, completely rely on the examples stored in a database (provided by a parallel aligned corpus), and on the underlying algorithms; as a result, they lack any other kind of explicit encoded information, which is usually found in traditional rule-based MT systems (e.g. lexicons, grammar formalisms, syntax rules, etc.).

After an appropriate corpus has been selected to provide the main linguistic component of the system, example-based machine translation cannot be implemented until the corresponding segments of the original and translated texts have been correctly aligned, that is matched against each other, deciding which level of granularity is desired. The granularity defines the size and span of the paired fragments or segments in the corpus and it may vary, but in general the alignment operation of parallel corpora takes place at the sentence level.

There has recently been an increasing attention devoted by researchers (see for instance Gale & Church 1991, Macklovitch & Hannan 1998 and Simard & Plamondon 1998), as well as by the software industry, addressing the problem of aligning texts, which has many useful applications, but can also be extremely complicated, as in the cases in which the texts to be aligned are in languages with different writing systems and character sets, e.g. English with roman characters and Japanese with Kanji, Hebrew, Arabic or other major Asian languages which are very interesting for commercial purposes.

Some remarkable trends of present-day machine translation research and development, however, do not take the pure example-based approach as dogma. Slight changes to the basic model give rise to the so-called hybrid machine translation systems, so as to circumvent, the bottlenecks normally encountered by truly example-based MT systems, for example by incorporating a lexicon into the machine translation engine. This is of course a major deviation from the basic paradigm of EBMT, since introducing a lexicon clearly gives predominance to computational encoding of some sort of real-world and linguistic knowledge, or, in more general terms, explicit linguistic rules, which are foreign to the original entirely corpus-based approach of pure EBMT.

Nevertheless, this represents a choice that offers straightforward solutions to most of the problems encountered in the pure data-driven design, and as far as these approaches to machine translation are concerned, it seems appropriate to talk about various “flavours of EBMT”, as Somers (1998:28) does; in fact, many researchers and developers have implemented

the integration at varying degrees of the example-based and rule-based methods with experimental hybrid MT systems.

These are the so-called multi-engine systems, in which "EBMT operates in parallel with two other techniques: knowledge-based MT and a simpler lexical transfer engine" (Somers 1998:30). As can be easily understood, therefore, example-based MT does not exclusively prove interesting in itself, but it also offers a viable support to other more traditional rule-based approaches, and the resulting hybrids open up a whole range of exciting variegated scenarios.

4. EBMT, MINORITY LANGUAGES AND LOW-DENSITY LANGUAGES

One of the reasons why example-based MT received much attention over the last years is that it seemed to offer a reasonable and viable path towards the design and implementation of MT systems especially for the so-called minority languages, like those that receive little attention from a commercial point of view, or do not appeal to a host of potential users for lucrative purposes (cf. Somers 1997).

It is a well-known truth that the languages covered by fully-fledged working MT software (e.g. commercial PC-based or in-house proprietary systems) are those that are most likely to attract substantial funding, mainly because they play a key role in the communication workflow in arenas such as industry, trade and business[3], science, or are crucial for socio-political reasons (e.g. are widely used within supra-national institutions where multilingual policies are adopted[4] or in bi-/multi-lingual states and regions[5]).

Taking into consideration the machine translation systems that are currently available on the market or the on-line MT services that can be accessed through the Internet free of charge[6], it is easy to agree that "MT has not served minority languages well, the main reason being the commercial reality of insufficient sales to justify the massive cost of creating the machine translation software" (Gordon 1997).

It can be interesting to note in passing that offering minor languages that are not particularly well covered by the language service industry is one of the concerns of human professional translators. Whether they work on a freelance basis or are employed in a translation department or bureau, by and large professionals often try to provide expertise that is either highly specialised (proficiency in particular text-types or subject matters), or difficult to find on the market (e.g. offering the so-called exotic languages, whose needs are not adequately catered for and accordingly are in great demand).

Within the present discussion these crucial issues that apply to human translation have a clear parallel in the availability of extensive resources

that have already been developed and can be exploited to create working MT systems. This is particularly relevant for the so-called low-density languages, that is to say natural human languages for which the body of available computationally exploitable resources (e.g. computerised corpora in machine-readable form, implemented grammar formalisms, parsers, etc.) is poor or, in the worst cases, completely absent.

Similarly to what happens for minority languages, low-density languages cannot rely on significant resources and materials that could be exploited for practical MT-oriented applications. The following discussion should help to illustrate why and how the corpus-based approach could at least partially help to circumvent the need to resort to an extensive range of pre-existing resources in the creation and implementation of fully automatic MT systems, in particular for minority and low-density languages.

5. EXAMPLE-BASED MT VS. RULE-BASED MT

Traditional rule-based machine translation (RBMT) systems need extensive resources, funding, and qualified personnel with a specific computational and linguistic expertise to be designed and implemented; they tend to be difficult to maintain, and are prone to inconsistency due to the labour-intensive and time-consuming encoding of formalisms and input of data, which can also be divided up among various people in case of multilingual systems. Moreover, rule-based machine translation systems are in general difficult to scale up or revise and correct when inconsistencies are discovered or bugs need to be fixed.

Example-based machine translation systems, on the other hand, are much easier to build, provided that there is the availability of the necessary bilingual corpora, whose alignment is a mundane task, but at present there are several technologies that can help to carry it out quite effectively. Furthermore, they require no in-depth computational expertise, due to the simple underlying algorithms, especially if compared with those of rule-based machine translation.

There are currently large multilingual corpora and a reasonably wide range of resources available for a number of language combinations that can be applied to the implementation of EBMT systems, so that several language pairs could be theoretically covered. Even with small-scale efforts, this picture could easily accommodate some of the low-density languages, e.g. by investing limited amounts of time in the creation of computerised corpora and their alignment, which could be effectively carried out by personnel that is not specialised or highly trained.

In summary, this discussion suggests that EBMT may represent a particularly suitable approach for low-density and minority languages, since e.g. in some respects the conventional rule-based strategies to design and implement fully automatic machine translation systems are not feasible for languages that can only rely on very limited resources. In some cases, when for instance there is a need to develop from scratch a (perhaps toy) MT system for whatever reason including in its coverage a largely neglected language, it may well be the case that this cannot rely on previously developed resources.

In such circumstances, then, EBMT may be an adequately flexible candidate to investigate and carefully consider, possibly integrating it with the rule-based approach into hybrid engines, and not necessarily restricting the choice to a mutually exclusive alternative option, which would be a misleading simplification.

6. EBMT at Work: Consequences of Adopting a Corpus-Based Approach to MT

Regardless of the source and target languages involved in the translation process, in pure EBMT systems all the linguistic knowledge is implicitly contained in the aligned parallel corpus. As a result, in case of internal problems, debugging and troubleshooting take place by spotting the examples which cause problems, for instance with a simple keyword search procedure through the corpus.

The deceptive fragments included in the data can be accordingly removed from the collection, or improved upon by eliminating ambiguities or confusing examples (e.g. those containing metaphors or representing translations with instances of compensation, etc., which are all cases and strategies that professional translators are very familiar with). Another option consists in removing from the database redundant examples that instead of increasing the effectiveness of the system form a useless burden.

An extensive parallel corpus of normal texts will contain overlapping instances of language use and translation phenomena that can be mainly grouped into two categories: some examples reinforce themselves mutually, because they contain the same language patterns, or they are virtually identical; otherwise, some examples may be in conflict, giving rise to different inconsistent translations due to their misleading effects. As a result, when they are used as models to produce the output in the target language, these overlapping examples may represent an irrelevant or even potentially damaging redundancy.

When the examples reinforce each other, it should be decided in each case whether they should be partially removed or not, but this is absolutely necessary when contrasting examples present the system with difficult choices that it is not able to handle:

> *Some systems involve a similarity metric [...] which is sensitive to frequency, so that a large number of similar examples will increase the score given to certain matches. But if no such weighting is used, then multiple similar or identical examples are just extra baggage, and in the worst case may present the system with a choice - a kind of "ambiguity" — which is simply not relevant (Somers 1998:24)*

Furthermore, introducing new examples may help to extend the possible applications of example-based machine translation systems, since this in practice enriches their linguistic knowledge, and accordingly the domains and possibly, even if to a lesser extent, the text-types they can successfully deal with. At the same time, it is also possible to integrate additional modules (explicit rules, partial formalisms and lexicons) aimed at improving the performance and reliability of the system by favouring hybrid approaches.

Considering how heavily EBMT relies on the role of the examples stored in the database to provide the output, one crucial issue affecting this approach to machine translation concerns the selection of the bilingual parallel corpora that provide the aligned examples. It should be borne in mind that they tend to impose a restriction on the subject matter of the texts and accordingly on the text-types that can be machine-translated.

This is similar to what happens with sublanguage-based MT systems (the most famous and successful case of this kind is the Canadian Meteo system, see e.g. Thouin 1982 and Kittredge 1987). As far as example-based machine translation systems are concerned, accurate choices should be made to pursue the most appropriate selection of corpora and the most effective degree of hybridisation to deal with specific text-types or domains. In this respect, Somers (1998) convincingly comments as follows:

> *EBMT is closely allied to sublanguage translation, not least because of EBMT's reliance on a real corpus of real examples: at least implicitly, a corpus can go a long way towards defining a sublanguage. On the other hand, nearly all research nowadays in MT is focused on a specific domain or task, so perhaps all MT is sublanguage MT. (Somers 1998:28)*

7. TRANSLATION MEMORY CAT TOOL

Now that some of the most noticeable features of example-based machine translation have been outlined and discussed, the remaining part of the paper will concentrate on the translation memory (TM) tool, one of the best-known members in the family of CAT (computer-assisted translation) software applications. It should be noted that the term *translator's workstation* is sometimes used in a fairly wide sense to encompass not only translation memory software, which is nonetheless one of its basic components, but also a number of other tools that make up the PC-based working environment of professional translators (including for instance terminology management packages, dictionary look-up facilities, etc.).

The technology behind translation memory software is sometimes still wrongly confused with example-based machine translation, because in both cases the translation strategy hinges on the retrieval and manipulation of previously stored examples, which are then used as components or models to produce the target-language text. However, this is a very misleading imprecision, which the following exposition will try to clarify.

Translation memories provide options available to human users (most often professional translators), who are then free to reject the proposed matches (significantly called candidates), edit them, change the similarity parameters which establish the threshold of least correspondence, etc. The proper memory consists of a database of aligned translation units stored and matched against each other, and the software displays the retrieved fragments taken from the collection of paired translation correspondences as models.

In order to recycle the proposed models, the translators select and edit the most appropriate candidates, among those corresponding with varying degrees of similarity (fuzzy matches) to the translation passage under consideration. In the case of perfect matches (when there is a 100% correspondence between the model and the passage of text to be translated) it may be the case that no editing at all is required, since the candidate fragment found in the memory might fit the rest of the text perfectly.

More commonly when using a translation memory tool, though, the selected candidate is in general added to the target text (by pasting it into the file or document of the translation at the right place), and manipulated to make sure that this insertion does not lower the quality of the target text, e.g. providing for agreements of gender and number among the words in the sentence, grammatical correctness, syntactic fluency, etc. The translation memory is accordingly updated and augmented, so as to include new information which might be relevant to the human translator in the near or even distant future, thus recycling previous work and avoiding the need to

translate the same passage from scratch again for the language pair concerned.

8. BENEFITS OF USING TRANSLATION MEMORY SOFTWARE FOR PROFESSIONAL TRANSLATORS

Especially when dealing with repetitive text-types and technical documents, one of the greatest advantages of using translation memories is that the requirements of terminological consistency and standardised phraseology are safeguarded. The appropriate and homogeneous use of terms to designate the corresponding concept or object, which is in particular of great relevance in technical texts such as manuals, should in fact be guaranteed not only within one single document or translation project, but also across similar documents that for whatever reason share some textual requirements, or belong to the same family of texts.

Common experience shows that these requirements of consistency are extremely difficult to manage manually without the assistance of CAT software, especially if the translations are performed by more than one translator. As a matter of fact, using translation memory software can also be very beneficial when a family of similar documents is translated by various people over a long period of time (possibly for the same client or if the documentation refers to the same line of products): no matter whether a single translator or a team of colleagues is involved in the task, consistency across texts is guaranteed.

At the same time, existing multilingual (e.g. already translated) material should be leveraged to the greatest possible extent to reduce turnaround time and enhance throughput, favouring also homogeneity in the company's preferred language style and complying with their policy of document production (see e.g. Heyn 1998 and O'Brien 1998). An effective organisation and management of the textual databases that make up the translation memories help professional translators to successfully cope with such complex translation projects, that would be much more labour-intensive, time-consuming and inevitably error-prone if performed manually in the traditional fashion, with the translator working with paper-based dictionaries and printed glossaries.

Translation memory software represents a very versatile CAT tool, in that the aligned parallel corpora containing the repository of examples can be constructed either incrementally (which means while translating, even if starting from scratch with empty databases) or using existing translated texts, after an appropriate alignment phase. The preparation of the textual database with the matched translation units is worth undertaking

if the available parallel corpora contain texts that are somehow similar to those of future translations (either for the style used or the contents, more in general) for a given language pair, irrespective of the direction of the translation (i.e. swapping between source and target language does not make any difference in this respect).

The employment of the translation memories is bound to speed up the whole process of translation and maximise the productivity of freelance professionals and in-house staff working in translation departments alike. These CAT tools can in fact be used and shared by a team of translators working connected to a local area network (LAN), so as to guarantee that terminological consistency and homogeneous phraseology are safeguarded and maintained throughout a whole translation project, when for instance a team is simultaneously working on different parts or chapters of a long manual or technical text to be translated.

This could be the case when the source documentation amounts to several hundred pages, and even though translators work independently, the target text increases its consistency while it is being translated, since the database of the translation memory is updated and shared by the translators at any moment while the work of the team is in progress. The translators access the same aligned parallel corpus and add new translation units that are made available on the spot to the other colleagues when they come across similar passages or repeated phrases. This environment of shared resources therefore enhances and maximises the benefit of teamwork among colleagues who master the same TM technology.

Using translation memories to manage and retrieve passages and units taken from previous jobs and projects, and possibly to reuse them, is clearly subject to the degree of comprehensiveness and accuracy of the parallel corpora containing the multilingual data. In this perspective, feeding the textual databases represents an investment, in that it is most likely to yield productivity gains in future translations.

For this reason, the aligned segments stored in the linguistic database as translation units represent a valuable asset that can have a crucial impact on repetitive or similar translation projects. Professional translators who rely on translation memory software capitalise on their present job in the interest of the productivity of their future work, especially in terms of streamlined workflow and higher productivity (increased speed, reduced turnaround time, maximised throughput and enhanced consistency).

9. Differences Between Translation Memory Software and Example-Based Machine Translation

Aligned parallel corpora of real texts account for the basic component in both example-based machine translation systems and translation memory tools. In spite of this, the basic difference between them should be emphasised: unlike example-based machine translation systems, translation memory software presupposes and takes for granted the operational intervention of the human user (in fact, a professional translator) during the translation phase, to choose and re-elaborate the candidate fragments that are inserted into the target document and become part of the translated text.

On the contrary, as has been illustrated above, example-based MT systems provide an autonomous analogical recombination of the textual fragments stored in the database, which is directly aimed at producing a translation in the target language, without being necessarily subject at all to a significant intervention of human users in the intermediate steps of the process. As a result, at this stage of the discussion it should be clear that example-based machine translation pushes automation much further than the technology governing translation memory software.

Mentioning a specific bottleneck typically encountered by EBMT can clearly emphasise the different potential and strengths of example-based machine translation systems on the one hand and translation memory tools on the other. One of the most difficult ways to use EBMT systems consists in translating from gender-neutral to gender-explicit (or gender-marking) highly inflected languages.

EBMT systems in general fail in this respect; human post-editing of the raw machine-translated output, even to obtain an understandable version for skimming purposes aimed at the basic gist of the text (general information gathering), would necessarily require an incredible amount of minor corrections (e.g. to guarantee gender and number agreement for nouns, adjectives, pronouns, articles, etc.), that ultimately make the adoption of the example-based approach a counter-productive strategy.

In this case, the labour for a human translation from scratch could possibly take less time, providing of course a better end product at reduced costs. As a result, when considering and assessing the real usefulness of machine translation systems in general, the degree of human involvement (e.g. in terms of the necessary time and cost) is to be taken into account:

> *The key issue is how much of the total effort can be handled by a computer and how much must still be done by human labor. Text*

> *input, pre-editing, and post-editing can take as much human time and effort as complete human translation (Henisz-Dostert et al. 1979:81)*

A considerable need for post-editing done by human revisers would of course make the use of machine translation useless and not desirable, or rather, even less desirable than it usually is. On the other hand, when using the translation memory CAT tool, issues related to grammatical gender agreements are solved by straightforward strategies, basically consisting in setting the appropriate similarity threshold for matches and adequate acceptability parameters, followed by simple step-by-step editing of the previous translations offered as candidates by the translation memory software.

These actions mainly take place at the sentence level and are carried out by the human user, the translator, who is in charge of adapting the proposed models according to the structure and peculiarities of the passage, sentence, document and text-type under consideration. In this case, co-operation between the translators and the software actually establishes a productive symbiosis that is the rationale behind the successful employment of the translation memory CAT tool.

10 CONCLUSION: SOME FOOD FOR THOUGHT

This discussion should have shed some light on some of the reasons why both machine translation and computer-assisted translation tools, in spite of their differences, influence the activity of translators and affect translation itself in various ways. Particular emphasis was placed on the interesting contribution provided by corpus linguistics and parallel corpora to the recent progress in MT system design and in the use of computer-based tools aimed at professional translators, such as translation memory software. The overall picture shows that parallel corpora exploited in an automated or semi-automated environment currently have a strong impact on the translation activity, outlining a variety of scenarios that deserve careful consideration. Along these lines, a conclusion aimed at stimulating further reflection and thought on these issues, whose importance is certainly bound to rise in the near future, can be offered by a short yet convincing passage taken from an article devoted to discussing some of the possible trends and directions for the evolution of computer-assisted translation tools: “Use computers for what they are good at (e.g. fast searching among large quantities of data), and let humans take the baton when it comes to the *"génie de la langue"*” (Langé et al. 1997:42).

REFERENCES

Gale, W. & K. W. Church 1991. A Program for Aligning Sentences in Bilingual Corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, California. 177-183.

Gordon, I. 1997. The TM Revolution - What does it really mean? Translating and the Computer 19 - Papers from the Aslib conference held on 13 & 14 November 1997. London: Aslib.

Henisz-Dostert, B., R.R. Macdonald, & M. Zarechnak. 1979. *Machine Translation.* The Hague: Mouton Publishers.

Heyn, M. 1998. Translation Memories: Insights and Prospects. *Unity in Diversity? Current Trends in Translation Studies.* Eds., L. Bowker, M. Cronin, D. Kenny & J. Pearson. Manchester: St. Jerome Publishing. 123-136.

Hutchins, J. & W. Hartmann 2002. Compendium of Translation Software. Commercial machine translation systems and computer-aided translation support tools. Version 1.5 (Revision, October 2002). European Association for Machine Translation. [Can be downloaded at the URL http://www.eamt.org/compendium.html - Accessed 10 September 2002].

Kittredge, R.I. 1987. The significance of sublanguage for automatic translation. *Machine Translation: Theoretical and Methodological Issues.* Ed., S.Nirenburg,. Cambridge: Cambridge University Press. 59-67.

Langé, J.-M., É. Gaussier, & B. Daille, 1997. Bricks and Skeletons: Some Ideas for the Near Future of MAHT. *Machine Translation* 12, 1-2: 39-51.

Macklovitch, E. & M.-L. Hannan. 1998. Line 'Em Up: Advances in Alignment Technology and their Impact on Translation Support Tools. *Machine Translation* 13, 1:41-57.

McEnery, T. 1997. Multilingual Corpora - Current Practice and Future Trends. Translating and the Computer 19 - Papers from the Aslib conference held on 13 & 14 November 1997. London: Aslib.

McTait, K. 2001. Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns. Proceedings of the Workshop on Example-Based Machine Translation. Machine Translation Summit VIII. Santiago de Compostela, Spain - 18 September 2001. 23-34.

Nagao, M. 1984. A Framework of Mechanical Translation between Japanese and English by Analogy Principle. *Artificial and Human Intelligence.* Eds., A. Elithorn & R. Banerji. Amsterdam: North-Holland. 173-180.

Nirenburg. S., S. Beale & C. Domashnev 1994. Full-Text Experiment in Example-Based Machine Translation. Proceedings of the International Conference on New Methods in Language Processing (NeMLaP). Manchester. 78-87.

O'Brien, S. 1998. Practical Experience of Computer-Aided Translation Tools in the Software Localization Industry. *Unity in Diversity? Current Trends in Translation Studies.* Eds.. L. Bowker, M. Cronin, D. Kenny & J. Pearson, Manchester: St. Jerome Publishing. 115-122.

Rico Pérez, C. & A. Martin de Santa Olalla Sánchez 1997. New Trends in Machine Translation. *Meta* 42, 4: 605-615.
Simard, M. & P. Plamondon 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation* 13, 1: 59-80.
Skousen, R. 1989. *Analogical Modeling of Language.* Dordrecht: Kluwer.
Somers, H.L. 1997. Machine Translation and Minority Languages. Translating and the Computer 19 - Papers from the Aslib conference held on 13 & 14 November 1997. London: Aslib.
Somers, H.L. 1998. "New paradigms" in MT: the state of play now that the dust has settled. Proceedings of the 10th European Summer School in Logic, Language and Information, Workshop on Machine Translation. Saarbrucken. 22-33.
Somers, H. 1999. Review Article: Example-based Machine Translation. *Machine Translation* 14, 2: 113-157.
Somers, H. 2001. EBMT Seen as Case-based Reasoning. Proceedings of the Workshop on Example-Based Machine Translation. Machine Translation Summit VIII. Santiago de Compostela, Spain - 18 September 2001. 56-65.
Sumita, E. & H. Iida 1991. Experiments and Prospects of Example-Based Machine Translation. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, California. 185-192.
Sumita, E. & H. Iida 1995. Heterogeneous Computing for Example-Based Translation of Spoken Language. Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation. Leuven, Belgium. 273-286.
Thouin, B. 1982. The Meteo System. Practical Experience of Machine Translation. Proceedings of a Conference, London, 5-6 November 1981, V. Lawson, ed. Amsterdam: North-Holland Publishing Company. 39-44.
Turcato, D. & F. Popowich 2001. What is Example-Based Machine Translation?. Proceedings of the Workshop on Example-Based Machine Translation. Machine Translation Summit VIII. Santiago de Compostela, Spain - 18 September 2001.43-48.
Way. A. 2001. Translating with Examples. Proceedings of the Workshop on Example-Based Machine Translation. Machine Translation Summit VIII. Santiago de Compostela, Spain - 18 September 2001. 66-80.

NOTES

1 This paper is based on a talk given by the author at the International Conference CULT 2K (Corpus Use and Learning to Translate 2000), held in Bertinoro (Italy) in November 2000. The paper presented here has been considerably extended and new more up-to-date references have been added where necessary.

2 There is not enough space here to review individually and discuss in detail the most influential contributions in this field. However, the references

mentioned in this footnote are for the convenience of readers who wish to look at background literature and up-to-date research papers of particular interest: Nagao 1984, Skousen 1989, Sumita & Iida 1991, Nirenburg et al. 1994, Sumita & Iida 1995, McEnery 1997, Rico Pérez & Martin de Santa Olalla Sánchez 1997, Somers 1998, Somers 1999, McTait 2001, Somers 2001, Turcato & Popowich 2001 and Way 2001.

3 Large volumes of the technical and specialist documentation that are available today throughout the world are originally produced or need to circulate in English, which is by far the most widely used lingua franca at present. There are few other languages with significant volumes of circulating documentation. For instance, assuming that a technical text (say, a user manual for a domestic appliance) is already available in English, a common list of prioritised target languages for its translation and distribution in Europe would be those indicated by the acronym FIGS, which stands for French, Italian, German and Spanish.

4 In this respect, the European Union and the Pan American Health Organization provide two convincing examples.

5 One may think for instance of Canada, traditionally associated with the success story of the Meteo sublanguage-based MT system to translate between English and French; without referring to MT in particular, but expanding the focus to linguistic resources as well, the situation remains similar, since for instance in Canada the Hansard corpus was created, i.e. a massive bilingual parallel corpus of parliamentary proceedings in English and French, due to the bilingual status of that country.

6 Hutchins & Hartmann (2002) provide a comprehensive and up-to-date listing of MT software, including on-line MT systems that translate Internet and Web content. Most of these MT products and web-based services cover a limited number of language combinations or only very few specific language pairs.

FEDERICO GASPARI
Centre for Computational Linguistics
UMIST
PO Box 88
Manchester M60 1QD
United Kingdom
E-mail: F.Gaspari@postgrad.umist.ac.uk