This document is subject to the following notice:

NOTICE: This material may be protected by copyright law.

The copyright law of the United States (Title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material; the person using this equipment is liable for any infringement.
Benjamins Translation Library

The Benjamins Translation Library aims to stimulate research and training in translation and interpreting studies. The Library provides a forum for a variety of approaches (which may sometimes be conflicting) in a socio-cultural, historical, theoretical, applied and pedagogical context. The Library includes scholarly works, reference works, post-graduate text books and readers in the English language.

General editor
Gideon Toury
Tel Aviv University

Associate editor
Miriam Shlesinger
Bar Ilan University

Advisory board
Arlan Gaddis Rose
Binghamton University

vlves Gambier
عركس University

Daniel Gile
université Lumière Lyon 2 and ISIT Paris

Jrich Heid
University of Stuttgart

Jna Hung
Chinese University of Hong Kong

V. John Hutchins
University of East Anglia

Anna Jettenmarová
Charles University of Prague

Verner Koller
frogen University

Jnt Kruger
NISA

osé Lambert
atholic University of Leuven

Franz Pochhammer
University of Vienna

Rosa Rabadán
University of León

Roda Roberts
University of Ottawa

Juan C. Sager
UMIST Manchester

Mary Snell-Hornby
University of Vienna

Sonja Tirkkonen-Condit
University of Joensuu

Lawrence Venuti
Temple University

Wolfram Wilss
University of Saarbrücken

Judith Woodworth
Mt. Saint Vincent University Halifax

Sue Ellen Wright
Kent State University

Volume 35
Computers and Translation: A translator’s guide
Edited by Harold Somers

John Benjamins Publishing Company
Amsterdam/Philadelphia 2003
CHAPTER 1

Introduction*

Harold Somers
UMIST, Manchester, England

1. Preliminary remarks

This book is, broadly speaking, and as the title suggests, about computers and translators. It is not, however, a Computer Science book, nor does it have much to say about Translation Theory. Rather it is a book for translators and other professional linguists (technical writers, bilingual secretaries, language teachers even), which aims at clarifying, explaining and exemplifying the impact that computers have had and are having on their profession. It is about Machine Translation (MT), but it is also about Computer-Aided (or -Assisted) Translation (CAT), computer-based resources for translators, the past, present and future of translation and the computer.

Actually, there is a healthy discussion in the field just now about the appropriateness or otherwise of terms like the ones just used. The most widespread term, “Machine Translation”, is felt by many to be misleading (who calls a computer a “machine” these days?) and unhelpful. But no really good alternative has presented itself. Terms like “translation technology” or “translation software” are perhaps more helpful in indicating that we are talking about computers, the latter term emphasising that we are more interested in computer programs than computer hardware as such. Replacing the word “translation” by something like “translator’s” helps to take the focus away from translation as the end product and towards translation as a process carried out by a human (the translator) using various tools, among which we are interested in only those that have something to do with computers.

We hope that this book will show you how the computer can help you, and in doing so we hope to show also what the computer cannot do, and thereby reassure you that the computer, far from being a threat to your livelihood, can become an essential tool which will make your job easier and more satisfying.
1.1 Who are we?

This book has been put together by academics (teachers and researchers, employees of software companies, and — yes — even translators. All the contributors have an interest in the various aspects of translation and computers, and between them have several hundred years' worth of experience in the field. All are committed to telling a true story about computers and translation, what they can and cannot do, what they are good for, and what they are not. We are not trying to sell you some product. But what we are aiming to do is to dispel some of the myths and prejudices that we see and hear on translators' forums on the Internet, in the popular press, even in books about translation whose authors should know better!

1.2 Who are you?

We assume that you are someone who knows about and is interested in languages and translation. Perhaps you are a professional linguist, or would like to be. Or perhaps you are just a keen observer. In particular, you are interested in the topic of translation process, and are open to some new ideas.

You are probably not a computer buff: if you are looking for lots of stuff about bits and bytes, integer float memory and peripheral devices then this is not the book for you. On the other hand, you are probably a regular computer-user, perhaps at the level of word-processing and surfing the World Wide Web. You know, roughly, the difference between "software" and "hardware", you know about windows and desktops, files and folders. You may occasionally use the computer to play games, and you may even have used some software that involves a kind of programming or authoring. But by enlarge that's not really your area of expertise.

On the other hand, you do know about language. We don't need to tell you about how different languages say things differently, about how words don't always neatly correspond in meaning and use, and how there's almost never an easy answer to the question "How do you say X in language Y?" (though we may remind you from time to time). We assume that you are familiar with traditional grammatical terminology (noun, verb, gender, tense, etc.) though you may not have studied linguistics as such. Above all, we don't need to remind you that translation is an art, not a science, that there's no such thing as a single "correct" translation, that a translator's work is often under-valued, that translation is a human skill — one of the oldest known to humankind — not a mechanical one. Something else you already know is that almost no one earns their living translating literary works and poetry: translation is mostly technical, often nonetheless demanding, but just as often routine and sometimes — dare we admit it? — banal and boring. Whatever the case, the computer has a role to play in your work.

1.3 Conventions in this book

This is a technical book, and as such will, we hope, open avenues of interest for the reader. For that reason, we give references to the literature to support our arguments, in the usual academic fashion. Where specific points are made, we use footnotes so as to avoid cluttering the text with unwieldy references. We also want to direct the reader to further sources of information, which are gathered together at the end of each chapter. Technical terms are introduced in bold font. Software product names are given in italics, and are thus distinguished typographically from the (often identical) names of the company which produce them.

Often it is necessary to give language examples to illustrate the point being made. We follow the convention of linguistics books as follows: cited forms are always given in italics, regardless of language. Meanings or glosses are given in single quotes. Cited forms in languages other than English are always accompanied by a literal gloss and/or a translation, as appropriate, unless the meaning is obvious from the text. Thus, we might write that key-ring is rendered in Portuguese as porta-chave lit. 'carry-key', or that in German the plural of Hund 'dog' is Hunde. Longer examples (phrases and sentences) are usually separated from the text and referred to by a number in brackets, as in (1). Foreign-language examples are accompanied by an aligned literal gloss as well as a translation (2a), though either may be omitted if the English follows the structure of the original closely enough (2b).

(1) This is an example of an English sentence.

(2) a. Ein Lehrbuchbeispiel in deutscher Sprache ist auch zu geben. a text-book-example in German language is also to give
   'A German-language example from a text-book can also be given.'
   b. Voici une phrase en français. this-is a sentence in French

We follow the usual convention from linguistics of indicating with an asterisk that a sentence or phrase is ungrammatical or otherwise anomalous (3a), and a question-mark if the sentence is dubious (3b).
(3) a. *This sentence are wrong.
b. *Up with this we will not put.

2. Historical sketch

A mechanical translation tool has been the stuff of dreams for many years. Often found in modern science fiction (the universal decoder in Star Trek, for example), the idea predates the invention of computers by a few centuries. Translation has been a suggested use of computers ever since they were invented (and even before, curiously). Universal languages in the form of numerical codes were proposed by several philosophers in the 17th Century, most notably Leibniz, Descartes and John Wilkins.

In 1933 two patents had been independently issued for “translation machines”, one to Georges Arsrouti in France, and the other to Petr Petrovich Smirnov-Troyanskii in the Soviet Union. However, the history of MT is usually said to date from a period just after the Second World War during which computers had been used for code-breaking. The idea that translation might be in some sense similar at least from the point of view of computation is attributed to Warren Weaver, at that time vice-president of the Rockefeller Foundation. Between 1947 and 1949, Weaver made contact with a number of colleagues in the USA and abroad, trying to raise interest in the question of using the new digital computers (or “electronic brains” as they were popularly known) for translation; Weaver particularly made a link between translation and cryptography, though from the early days most researchers recognised that it was a more difficult problem.

2.1 Early research

There was a mixed reaction to Weaver’s ideas, and significantly MIT decided to appoint Yehoshua Bar-Hillel to a full-time research post in 1951. A year later MIT hosted a conference on MT, attended by 18 individuals interested in the subject. Over the next ten to fifteen years, MT research groups started work in a number of countries: notably in the USA, where increasingly large grants from government, military and private sources were awarded, but also in the USSR, Great Britain, Canada, and elsewhere. In the USA alone at least $12 million and perhaps as much as $20 million was invested in MT research.

In 1964, the US government decided to see if its money had been well spent, and set up the Automated Language Processing Advisory Committee (ALPAC). Their report, published in 1966, was highly negative about MT with very damaging consequences. Focussing on Russian–English MT in the USA, it concluded that MT was slower, less accurate and twice as expensive as human translation, for which there was in any case not a huge demand. It concluded, infamously, that there was “no immediate or predictable prospect of useful machine translation”. In fact, the ALPAC report went on to propose instead fundamental research in computational linguistics, and suggested that machine-aided translation may be feasible. The damage was done however, and MT research declined quickly, not only in the USA but elsewhere.

Actually, the conclusions of the ALPAC report should not have been a great surprise. The early efforts at getting computers to translate were hampered by primitive technology, and a basic under-estimation of the difficulty of the problem on the part of the researchers, who were mostly mathematicians and electrical engineers, rather than linguists. Indeed, theoretical (formal) linguistics was in its infancy at this time: Chomsky’s revolutionary ideas were only just gaining widespread acceptance. That MT was difficult was recognised by the likes of Bar-Hillel who wrote about the “semantic barrier” to translation several years before the ALPAC committee began its deliberations, and proposals for a more sophisticated approach to MT can be found in publications dating from the mid- to late-1950s.

2.2 “Blind idiots”, and other myths

It is at about this time too that much repeated (though almost certainly apocryphal) stories about bad computer-generated translations became widespread. Reports of systems translating out of sight, out of mind into the Russian equivalent of blind idiot, or The spirit is willing but the flesh is weak into The vodka is good but the meat is rotten can be found in articles about MT in the late 1950s; looking at the systems that were around at this period one has difficulty in imagining any of them able to make this kind of quite sophisticated mistranslation, and some commentators (the present author included) have suggested that similar stories have been told about incompetent human translators.

2.3 The “second generation” of MT systems

The 1970s and early 1980s saw MT research taking place largely outside the USA and USSR: in Canada, western Europe and Japan, political and cultural needs were quite different. Canada’s bilingual policy led to the establishment of a significant research group at the University of Montreal. In Europe groups in France, Germany and Italy worked on MT, and the decision of the Commission of the European Communities in Luxembourg to experiment with the Systran system (an American system which had survived the ALPAC purge thanks to private funding) was highly significant. In Japan, some success with getting computers to handle the complex writing system of Japanese had encouraged university and industrial
research groups to investigate Japanese–English translation.

Systems developed during this period largely share a common design basis, incorporating ideas from structural linguistics and computer science. As will be described in later chapters, system design divided the translation problem into manageable sub-problems — analysing the input text into a linguistic representation, adapting the source-language representation to the target language, then generating the target-language text. The software for each of these steps would be separated and modularised, and would consist of grammars developed by linguists using formalisms from theoretical linguists rather than low-level computer programs. The lexical data (dictionaries) likewise were coded separately in a transparent manner, so that ordinary linguists and translators could work on the projects, not needing to know too much about how the computer programs actually worked.

2.4 Practical MT systems

By the mid 1980s, it was generally recognised that fully automatic high-quality translation of unrestricted texts (FAHQT) was not a goal that was going to be readily achievable in the near future. Researchers in MT started to look at ways in which usable and useful MT systems could be developed even if they fell short of this goal. Many commentators now distinguish between the use of MT for assimilation, where the user is a reader of a text written in an unfamiliar language, and dissemination, where the user is the author of a text to be published in one or more languages. In particular, the idea that MT could work if the input text was somehow restricted gained currency. This view developed as the sublanguage approach, where MT systems would be developed with some specific application in mind, in which the language used would be a subset of the “full” language, hence “sublanguage” (see Chapter 15). This approach is especially seen in the highly successful Météo system, developed at Montreal, which was able to translate weather bulletins from English into French, a task which human translators obviously found very tedious. Closely related to the sublanguage approach is the idea of using controlled language, as seen in technical authoring (see Chapter 14).

The other major development, also in response to the difficulty of FAHQT, was the concept of computer-based tools for translators, in the form of the translator’s workstation (see Chapter 2). This idea was further supported by the emergence of small-scale inexpensive computer hardware (“microcomputers”, later more usually known as personal computers, PCs). Here, the translator would be provided with software and other computer-based facilities to assist in the task of translation, which remained under the control of the human: Computer-Aided (or -Assisted) Translation, or CAT. These tools would range in sophistication, from the (nowadays almost ubiquitous) multilingual word-processing, with spell checkers, synonym lists (“thesauri”) and so on, via on-line dictionaries (mono- and multilingual) and other reference sources, to machine-aided translation systems which might perform a partial draft translation for the translator to tidy up or post-edit. As computers have become more sophisticated, other tools have been developed, most notably the translation memory tool which will be familiar to many readers (see Chapter 3).

2.5 Latest research

Coming into the 1990s and the present day, we see MT and CAT products being marketed and used (and, regrettably sometimes misused) both by language professionals and by amateurs, the latter for translating e-mails and World Wide Web pages. This use will of course be the subject of much of the rest of this book. Meanwhile, MT researchers continue to set themselves ambitious goals.

Spoken-language translation (SLT) is one of these goals. SLT combines two extremely difficult computational tasks: speech understanding, and translation. The first task involves extracting from an acoustic signal the relevant bits of sound that can be interpreted as speech (that is, ignoring background noise as well as vocalisations that are not speech as such), correctly identifying the individual speech sounds (phonemes) and the words that they comprise and then filtering out distractions such as hesitations, repetitions, false starts, incomplete sentences and so on, to give a coherent text message. All this then has to be translated, a task quite different from that of translating written text, since often it is the content rather than the form of the message that is paramount. Furthermore, the constraints of real-time processing are a considerable additional burden. Try this experiment next time you are in a conversation: count to 5 under your breath before replying to any remark, even the most simple or banal, Good morning, or whatever. Your conversation partner will soon suspect something is wrong, especially if you try this over the telephone! But given the current state of the art in SLT, a system that could process the input and give a reasonable translation — in synthesised speech of course — within 5 seconds would be considered rather good.

Turning back to MT for written text, another concern is coverage of a wider variety of languages. So far, it is mainly the commercially important languages of western Europe and the Far East that have received the attention of developers. It is recognised that there are thousands more languages in the world for which MT or at least CAT software should be developed, but each new language pair poses huge problems, even if work done on a similar or related language can be used as a starting point. If a classical (linguistic) approach to MT is taken, then grammar rules for the new language must be written, “transfer” rules between the new and old languages and, the biggest bottle-neck of all, dictionary entries for thousands of words must be written. Even though the dictionaries used by MT systems contain information that
3. Overview of this book

The first seven chapters look at various uses to which a translator might put the computer while the second half of the book focuses more on MT. In Chapter 2 we describe the development of the ideas behind the translator's workstation, and look at some of the computer-based tools that can be made easily available to translators, with a special focus in Chapter 3 on one of these tools, the translation memory. Chapter 4 concerns the special place of terminology in the CAT scenario. Translators have always been aware of the need to access technical vocabulary and be sure that the terms chosen are correct and appropriate. As Lynne Bowker describes, computers can play a particularly useful role in this question, as term banks and other sources of terminology are available in various forms, both on-line and in the form of machine-readable dictionaries and thesauri.

A relatively new translation activity that has emerged in recent years goes under the name of software localization. In the early days of computers, most software (and hardware) that was produced was biased towards (American) English-speaking users. It has now been recognised that products aimed at a global market must be customized for local aspects of that global market. Software localization involves translating documentation, including on-line help files, but also often involves customizing the software itself, inasmuch as it contains language (for example, how to translate Press Y for Yes into French). In Chapter 5, Bert Esselink condenses some of the ideas from his comprehensive book on the subject, to give an indication of the problems involved, and some of the tools available to assist the translator in this specific task.

In today's commercially-oriented world, much translation work is motivated by commercial considerations. Socio-economic factors thus influence the development of MT and CAT systems, and it is the major European languages (English, French, Spanish, German, Italian, Portuguese, Russian) plus Japanese, Chinese, Korean and to a certain extent Arabic that have received attention from the developers. Bad luck if you work into (or out of) any of the several thousand other languages of the world. In Chapter 6 we look at the case of CAT and minority languages — an ironic term when one considers that the list of under-resourced languages includes several of the world's top 20 most spoken languages (Hindi, Bengali, Malay/Indonesian, Urdu, Punjabi, Telegu, Tamil, Marathi, Cantonese). We will consider what the prospects are for translators working in these languages (and other languages more reasonably described as "minority") and what kinds of computer-based tools and resources could be made available.

Our next chapter looks at the place of computers in the academic world of translator training. Sara Laviosa considers the use of the computer in Translation Studies; in particular, this chapter looks at how computer-based corpora — collections of translated text — can be used to study trends in translation practice.

The remaining chapters focus more closely on MT. In Chapter 8, Doug Arnold explains why translation is hard for a computer. Readers of this book will have views on what aspects of translation are hard for humans, but Arnold points out that some aspects of language understanding in the first place, and then the rendering of what has been understood in a foreign language in the second place, present difficulties to computers which, after all, are basically sophisticated adding machines. At least some of the difficulty is down to the nature of language itself, and in Chapter 9, Paul Bennett describes how the scientific study of language — linguistics — can help to provide some solutions to the problems.

The next three chapters focus on MT from the commercial point of view. In Chapter 10, John Hutchins, MT's unofficial historian and archivist-in-chief, details the current state of the art in commercially available MT and CAT software. Chapter 11 presents the developer's point of view. Co-authored by Laurie Gerber, formerly one of Systran's senior linguists, and Scott Bennett who, at the time of writing was a senior member of Logos's development team, and before that had helped oversee the productization of the Metal MT system by Siemens. In Chapter 12, Jin Yang and Elke Lange report on Systran's intriguing experiment in which they have made their MT system freely available on the World Wide Web. This contribution explains why the company is happy to see their product freely used, and reports on a period of close monitoring of the web-site, and users' feedback and opinions.

John White's chapter on how to evaluate MT will be essential reading for anyone thinking of MT or CAT tools as a solution to their translation needs, whether they be an individual freelancer, a small translation company or part of the translation department of a large company. White gives a practical and historical overview of what to evaluate, how to evaluate it, and, above all, some of the pitfalls to avoid.
The next three chapters address aspects of the practical use of MT. In Chapters 14 and 15 we look at two strategies for getting the best out of MT: Eric Nyberg, Terako Mitamura and Wolf Huijsen describe their approach to controlled language, explaining the basic idea behind the concept, and how it can be implemented within a translation scenario. An important feature of the controlled-language approach is the need to gain acceptance of the underlying idea from the authors of the texts to be translated, and to overcome the negative preconceptions of loss of author’s creativity and the resulting sterility of textual form that the term controlled language inevitably invokes. While the controlled-language approach restricts the syntax and vocabulary of the texts to be translated in a pre- and proscriptive way, the sublanguage approach takes advantage of naturally occurring restrictions and preferences in the style and vocabulary of the texts. In Chapter 15 we look at the classical example of a successful sublanguage MT system — the Canadian Météo system — and consider whether this was a “one-hit wonder” or whether the success of this experiment points the way for future MT success stories.

In Chapter 16, Jeffrey Allen looks at the question of revising MT output, usually termed post-editing to distinguish it from the parallel task of revision often performed on human translations. Allen brings out some of the differences in these two tasks, and outlines some strategies and techniques to make the task easier and more efficient.

In the final chapter we consider the use of MT and CAT tools in the teaching of translation, both to trainee translators, and to language students in general.

Further reading

The history of MT up to the mid 1980s is thoroughly documented in Hutchins (1986). Some “classical” papers have been collected in Nirenburg et al. (2003). Good, though now somewhat out-of-date, introductions to MT are given by Hutchins and Somers (1992) and Arnold et al. (1994). For more recent developments, see Trujillo (1999), which is also recommended for readers wishing to go into more technical detail about MT. Kay et al. (1994) is a good introduction to the problems of spoken language translation. References for other more specific aspects are given in the relevant chapters.

For on-going information, there are a number of good sources. The journal Machine Translation (published by Kluwer) is probably the premier academic journal in the field. A more informal source of information is the International Association for Machine Translation (IAMT) and its monthly newsletter MT News International. Membership of the IAMT is via one of its three regional chapters, AMTA in the Americas, EAMT in Europe, and AAMT for Asia and Australasia. All three maintain a web presence.

Notes

1. It is an interesting quirk of the English language that the vocabulary in general fails to distinguish between a process and its result: cf. painting, building, decoration, cooking, teaching, etc.

2. Some years ago, my professor — a pioneer in the field of computers and translation — was planning a prestigious talk he had been invited to give, and wanted my opinion of it. His opening sentence was “Translation is the world’s oldest profession.” I was just a young researcher, and wasn’t sure how to put it to him. “Er, Professor, I think that phrase is usually used about a slightly different line of work.” “Ah yes, of course”, he replied. I was eager to discover, a month later, what he would say instead. “Ladies and gentlemen,” he began, “translation is the world’s second oldest profession.”

3. “Sub-” in the mathematical sense, not “inferior”.

4. We refrain here from giving URLs since these are prone to rapid change. We are sure however that most search engines will easily find appropriate web pages using the search term “Machine Translation” or one of the acronyms listed in the text.

References


