Teaching the automation of the translation process to future translators

Benoît Robichaud and Marie-Claude L'Homme

Département de linguistique et de traduction
Université de Montréal
C.P. 6128, Succ. Centre-ville
Montréal (Québec)
Canada H3C 3J7

benoit.robichaud@umontreal.ca
lhommem@ling.umontreal.ca

Abstract

This paper describes the approach used for introducing CAT tools and MT systems into a course offered in translation curricula at the Université de Montréal (Canada). It focuses on the automation of the translation process and presents various strategies that have been developed to help students progressively acquire the knowledge necessary to understand and undertake the tasks involved in the automation of translation. We begin with very basic principles and techniques, and move towards complex processes of advanced CAT and revision tools, including ultimately MT systems. As we will see, teaching concepts related to MT serves both as a wrap-up for the subjects dealt with during the semester and a way to highlight the tasks involved in the transfer phase of translation.

1. Introduction*

This paper describes several aspects of a course entitled "*Outils informatiques des langagiers*" (Engl. "Computer tools for language professionals") in which a wide range of computer-assisted translation (CAT) tools and other applications – including machine translation (MT) systems – must be covered. The course is offered in all undergraduate translation curricula of the Département de linguistique et de traduction at the Université de Montréal (Canada). The focus here will be on the automation of the translation process and the strategies developed to help students progressively acquire knowledge on translation tools. We begin with fairly simple concepts, principles and techniques of basic tools and move on to the understanding of more complex processes found in advanced CAT and revision tools, and ultimately in MT systems.

As we know, concepts related to MT are not readily accessible, especially for students who do not have a background in linguistics or computing. Nonetheless teaching those concepts may serve both as a wrap-up for the various subjects covered during the semester and as an excellent way to draw attention to the tasks involved in the transfer phase of translation, which is often accomplished intuitively.

We first present the general objectives of the course and discuss some relevant technical details, as well as the academic background of the students. In the remaining sections, we set out the course contents, including the different themes, techniques and tools that are reviewed and taught throughout the semester, giving short examples of exercises and assignments. We close these sections by focusing on the benefits of ending the semester with the presentation of the concept of "fully automatic MT." We conclude the article with a few remarks on the reactions of the students to the course and suggestions for improvements in the future.

2. Course objectives and settings

The main goal of the course is to present a broad range of computer applications that translators will be asked to work with on modern workstations. It focuses on giving trainee translators the means to master the computerized tools in a professional setting, either in translation firms or as freelance workers. It draws heavily on the automation continuum of the translation process from the human's

* We would like to thank Elizabeth Marshman for very helpful suggestions. Thanks also to three anonymous reviewers for useful comments on an earlier version of this paper.

point of view, as pioneered in Kay (1980) and others.[1] Ultimately the course should enable the learners to evaluate the possibilities, limitations and impacts of each tool against the gain it represents for various translation needs and in different settings.

Apart from Master's and Ph.D. levels,[2] the course is compulsory in all translation curricula leading to translation markets. Groups are composed of undergraduate students (certificates and Bachelor of Arts degrees) and graduate students (in a one-year specialization program). Most of the students work in English-French translation, but a growing number of alumni have other working languages (Spanish, for instance).

The course is given in 45 hours over fifteen weeks. Except for two classes devoted to exams, each class is divided equally between a lecture and a hands-on session in a computer laboratory. The laboratory is equipped with workstations on which the various applications are installed. The students are instructed to read some chapters of a reference book (L'Homme, 2000) to prepare for the lectures,[3] and to refer to a dedicated university website for supplementary notes and links (L'Homme and Robichaud, 1999). It is noteworthy that the website also contains practical instructions for the exercises students must complete at the laboratory sessions. This medium has proved to be a very valuable tool for both students and instructors. Students can refer to the website whenever they want and can do the exercises at their own pace. Finally, the evaluation is based on two exams on the material covered during the lectures in addition to five assignments consisting of selected exercises involving the use of specific computer applications in concrete translation situations.

It must be emphasized that "*Outils informatiques des langagiers*" is the only course entirely devoted to computer applications that is offered as part of the translation programs. Hence, it must cover several different topics in a relatively short period of time (see Figure 1). However, an increasing number of professors are becoming aware of the importance of including some applications, and ask students to use them.

Although all of the students are registered in a translation program, their levels of competence and knowledge of the professional world and of the different computer applications are somewhat heterogeneous. Therefore, special attention is paid to developing course content that suits the needs of all the students, whether they are well accustomed to working with computers but are less acquainted with translation practices as such, or have practical experience in translation and are taking the course to remain up-to-date. In all cases, students are warned at the very beginning that the course is neither an introduction to computer use nor to word processing, and they are advised to do extra work in order to catch up if necessary.

3. Contents of the course

Figure 1 gives a general overview of the material covered during the course. The important topics are discussed in detail in the following sections.

- Historical overview of translation and computing with key terminological distinctions such as MT, MAT, MAHT, HAMT, FAHQMT, etc.
- Translation situations, translators' needs and the automation of the translation process
- Computer networking, Internet and telecommunications (basics)
- Word processing ("advanced" functions)
- Translation problems and terminology management (database management systems and terminology management software)
- Text corpora (indexing and searching monolingual concordances)
- Paired text corpora (aligning and searching bilingual concordances)
- Recall tools (translation memories)
- Proofreading tools (spelling and grammar checkers)
- Machine translation

Figure 1: Overview of the contents of *"Outils informatiques des langagiers."*

[1] See discussion in Somers (1999).

[2] Master's and Ph.D. students in translation can take an advanced course entitled "*Traductique*" where the point of view on CAT tools shifts from use to design. Let us mention that advanced computational linguistics courses are offered in graduate linguistics and computer science curricula, where functioning and development aspects of current MT systems are covered in more depth. This division is similar to the one mentioned in Kenny & Way (2002).

[3] Other references of recent introductory textbooks such as Austermühl (2001) and Bowker (2002) are also given.

The first lectures are dedicated to a brief introduction to the historical context of MT and machine-aided translation (MAT) and show how this distinction developed. Several factors that have led to the computerization of the translation profession are then presented and important terminological distinctions are covered to show how computers can be included in the translation process and how they change the way translation is carried out nowadays.

The remainder of the course is devoted to the specific applications with special attention to the order in which they are introduced and practiced. We try to situate the uses of each application in the context of basic but realistic translation tasks (see next sections). The general idea is to begin with relatively simple processing strategies and techniques and move towards more complex ones. The objective is to facilitate the understanding of the overall functioning of a given application and to show how other applications may go further towards fulfilling specific translation needs by automating particular tasks.

For example, during the class on word processing, a few explanations are given on the word count and the search and replace functions. This allows the instructors to introduce basics of character string manipulation and to detail various problems related to word recognition, such as part-of-speech and semantic ambiguities. Once the concept of "ambiguity" is mastered in the context of a search and replace task where the "replace all" function cannot be used satisfactorily, students can easily understand why it becomes a real problem when using a grammar checker or MT software. Another example is the order in which monolingual corpora, aligned corpora and translation memories are introduced. Students learn first how to build text concordances and execute search queries in indexed monolingual corpora. They are taught next how source texts and their translations can be aligned, then also indexed and searched. Last, they discover how sentences can be automatically located and retrieved by translation memory software from its database of aligned texts. Several other examples are given below.

Finally, when teaching the various computer applications, we try to focus on the processing techniques involved and not on specific off-the-shelf software. Students should normally be able to transpose their knowledge to other commercial utilities when they begin working as professional translators.

4. Starting with the translation process

During one of the first classes, we introduce a simplified, seven-stage translation process (see Figure 2). We illustrate how each phase may be automated in some way, along with the diverse tasks that the translator will have to accomplish individually or with coworkers. Throughout the course, this model is augmented and adapted to cover different translation situations.

- Reception of the source language text
- First reading
- Documentary and terminological searches
- Translation *per se*
- Proofreading and correction
- Revision
- Delivery of the target language text

Figure 2: A simplified translation process.

The presentation of this simplified translation process leads us in the following classes to the introduction of the specific software applications and in parallel to the various degrees of processing difficulties that translation automation faces.

5. Basic tools

The first two applications that are dealt with are Internet browsers and e-mail tools. We do not spend much time on these topics in lectures, but introducing them during the first session at the laboratory ensures that all the students will master document downloading and exchanging. This is essential since all assignments are to be handed in as e-mail attachments. This first practical session also covers logins (on computer networks as well as e-mail and course website accounts) and a brief review of the OS environment, file system storage and software accessories such as archiving tools, PDF readers, etc.

The major and most important application that all translators must master at various stages of the translation process presented above is evidently the word processor. It might seem surprising that word

processing is still taught in a course at the university level, but experience has shown us that many of the functions we review are unknown to a majority of the students.

We quickly review basic functions such as word count, search and replace, tables and sorting, etc. These functions are excellent starting points for introducing linguistic and computational concepts such as the internal representations of characters, words and sentences for computer software, related word and sentence segmentation problems, string comparison, and the like. All of these notions will be crucial when we bring in advanced CAT and revision tools. We then move toward functions we view as "advanced" and classify them according to specific tasks. These include format conversions, document management and comparison, hyperlink insertion and management, annotation and tracking changes, macros, autocorrection lists, spell checking and personal dictionary management.[4] Exercises with the word processors are developed and related to the different parts of the translating process involving word processing. A few examples of exercises are given in Figure 3.

- Display and correct spelling errors
- Carry out a word count on a text and find out if the count is similar to that carried out by a translator
- Build a simple macro to manage word count adequately
- Create and sort tables to organize terminological data
- Standardize terminological choices both in texts and terminological tables with the search and replace function
- Automate the creation of a terminological table with a macro
- Compare two versions of a document
- Save a file in different exchange formats
- Organize the desktop to work efficiently with many open files and/or applications

Figure 3: Exercises with word processors.

The next important task of the translation process that we cover is terminology management, for which we use database management systems.[5] Of course, many translators may be able to afford and/or will be required to work with dedicated terminology-management systems (TMS). Nevertheless, we feel that it is important to renew with basic concepts such as records, fields and data types, relating them to a central concept, namely the "query." We use a generic database management system for this purpose and show students how to build terminology databases from scratch. The links established with the query permit us to introduce many new concepts.

First, we cover basic elements of set theory, such as union, intersection and complement. Students are then expected to be able to handle Boolean operators in various types of queries. Secondly, we present the notion of character masks such as those used in regular expressions and practice the use of wildcards to perform stemming and other kinds of searches. Third, based on the previous elements, we introduce the concept of adjacency (relative proximity and order) of words and practice the extraction of collocations from certain text fields. The mastering of those concepts will be very important for later classes devoted to text concordances.

Lastly, we briefly review record structures and basic retrieval functionalities[6] that are commonly offered by commercial TMS and compare them (price, learning curve, flexibility of importing and merging, etc.) with generic database management system that are now universally distributed with office suites. Different exchange formats that are presently available or emerging are also discussed at this point.

Exercises are related to the design, construction and querying of small terminological databases. Capturing, importing and exchanging terminological data are also thoroughly practiced. Examples of exercises students are asked to do at this time are given in Figure 4.

[4] Grammar checking is postponed until later on as we use stand-alone software and present the functionality in the context of proofreading and revision. During the class on word processing, we show students how spelling errors can be displayed automatically on screen or browsed one by one after launching the "Spelling and Grammar" function.

[5] Public term banks such as *TERMIUM®*, *EURODICAUTOM* and the *Grand dictionnaire terminologique* are covered in an earlier course exclusively devoted to documentation and terminology searches.

[6] Terminology extraction and pre-translation functionalities are also discussed later with advanced CAT tools.

- Design and build databases of varying complexity from scratch
- Build appropriate front-ends to display, capture and manage terminological data
- Import lists (created with the word processor or found on the Internet) automatically
- Perform different types of queries in databases
- Link different databases
- Organize and export selected contents of databases

Figure 4: Exercises with database management systems.

One of the important abilities for translators these days is undoubtedly to know how to build and query text concordances locally for terminology or documentation purposes. We introduce the distinctions between text collections and corpora (Bowker, 2000) along with the functioning and use of different types of monolingual concordance software, and we again address the concepts of indexation and searching. During the laboratory session, students learn how to search for strings, words and combinations of words in technical text corpora. The set notions, as well as the wildcard and word adjacency concepts discussed above are of primary importance here since the results of simple queries may be awfully noisy or on the contrary extremely slim.

- Build concordances of monolingual text corpora
- Perform various types of queries to find words, parts of words, word combinations and collocations
- Find ways to locate information that may help solve specific translation problems
- See and understand how textual data provides information that is not normally compiled in standard reference works such as dictionaries or term banks
- Place appropriate information retrieved in terminological databases

Figure 5: Exercises with concordancers.

The results of the various searches carried out as exercises are placed in new, manually created databases specifically designed for the different kinds of terminological problems that are being addressed. A few examples of tasks students must accomplish are given in Figure 5.

We next turn to bilingual concordances and explain techniques used to build bilingual corpora from source texts and their translations stored on a local workstation. Students must first create very short aligned corpora with manual and semi-automatic alignment methods using hyperlinks and a word processor macro. Then they learn to import and exploit the resulting corpora within the generic database management system. They are next taught to use and evaluate automatic alignment methods such as the ones found in Trados' *WinAlign* and Terminotix' *Logiterm* commercial applications.[7] We finish by introducing the RALI large bilingual databases and review the functionalities implemented therein to mask inflection and search for ellipses.[8] Examples of exercises are given bellow.

- Build a bilingual corpus using different methods (first manually with hyperlinks, second with a word processor macro, and third automatically with alignment tools)
- Import a bilingual aligned corpus into a generic database management system and perform queries
- Perform queries within the interface of alignment tools that support this
- Find texts on the Internet that could lend themselves to automatic alignment
- Perform queries in large aligned corpora

Figure 6: Exercises with aligned corpora.

6. Advanced tools

Once all sorts of concordance application software along with relevant functionalities have been thoroughly studied, we introduce recall tools as the first type of advanced CAT tools. Those involve higher-level automation of the translation process in some well defined settings such as updated versions of previously translated texts, highly repetitive technical texts, etc. We show students how these new tools are designed to automatically reuse (or recycle) segments of previously translated text stored in their databases.

First, the translation memory (TM) is introduced as a set of textual databases (viewed as a sort of aligned bilingual corpus) supplemented by a pilot (control) program. Detailed explanations are given on the functioning of the pilot that automatically

[7] See www.trados.com and www.terminotix.com.
[8] See www.tsrali.com.

divides the source text into segments (usually sentences), then retrieves and displays the "best translation solutions" from its databases. A relation is established with aligned corpora that have been broadly dealt with in previous classes. Further explanations are given on the specific contexts of use and functionalities that are commonly offered (pre-analysis of the source texts to evaluate their level of repetition and the database's applicability, integration into word processors, manual look-ups and updating of the databases, etc). More importantly, we review the different strategies the software uses for establishing exact and fuzzy matches along with methods for fine-tuning the best type of match according to translation situations.

Secondly, we present a functionality usually found in TM or TMS, which we refer to as a "vocabulary translator." The tool is described as a computerized bilingual dictionary that is also enhanced by a pilot control program. This time, the pilot is capable of searching for and replacing words, terms or idioms automatically with their lexical or terminological equivalent in the text to be translated.[9] The students are warned that the result is a partially translated text (as opposed to aligned ones seen previously), which has to be lexically verified and rephrased at the syntactic level in a post-editing phase. A links are established with terminological databases and dictionary look-ups that have also been dealt with in previous classes. Explanations are again given on the specific contexts of use as well as on the regular functionalities that are expected from this second kind of recall tool. Examples of exercises given to the students are shown in Figure 7.

- Translate an updated version of a text using a (demo version) translation memory interfaced with the word processor
- Search and modify the content of the translation memory database manually
- Run a (demo version) "vocabulary translator" on a source text and rewrite the resulting text into an acceptable translation

Figure 7: Exercises with translation memory and vocabulary translator applications.

[9] This processing is sometimes referred as "active terminology recognition" or "pre-translation."

If time allows, we turn next to terminology extraction.[10] We present software applications that extract and produce lists of lexical or terminological items found in source texts. Context of use and expected functionalities (sort and synthesis options) are presented along with a discussion on the combination of various strategies implemented in TMS that perform this task automatically (segment repetition, syntactic patterns, statistical collocations, etc.). Links are made with word segmentation and recognition, dictionary look-ups and corpus indexing techniques that were dealt with in previous classes. Examples of exercises are shown in Figure 8.

- Extract terminological content from a technical text and present different lists using the sort and synthesis options
- From the same source text, evaluate the appropriateness and rank of the various items found in the lists
- Place genuine terminological items in previously created terminological databases
- Explain why some terminological items of the source text have not been placed or have been misplaced in the lists

Figure 8: Exercises with terminology extraction software applications.

As seen in the translation process presented above, proofreading is still a very important step in producing a high-quality translation. While translators are skilled writers, for reasons that we will not review here, they may sometimes miss errors and typos made during the translation *per se*. Moreover, other applications involved at previous stages of the translation process may also have introduced errors.

Next to be introduced are tools such as spelling and grammar checkers. We briefly review the historical context and the technical changes that have been made from spelling correction to full grammatical checking. We next present how such tools can carry out "parsing," described as a sort of symbolic calculus on syntactic structures and

[10] As mentioned above, many concepts, techniques and tools have to be covered during the course. From time to time, choices have been made by instructors to summarize this part in the context of terminology management in the interests of developing a deeper understanding of the other topics.

relying on grammatical and lexical knowledge. At this time we bring in structural ambiguities (related to the part-of-speech and semantic ambiguities discussed earlier) and how they can also lead to erroneous results. We then introduce notions such as "noise" and "silence" (in this context, overcorrection and undercorrection) to cover the different types of difficulties and limitations involved in using the tools. Finally, other technical topics such as integration into word processors and generally expected functionalities are covered. Examples of exercises students must complete for this class are shown in Figure 9.

- Correct a list of incorrect sentences with the grammar checking functionality integrated in a word processor
- Correct the same list of sentences with a stand-alone grammar checker
- Compare corrections of both grammar checkers by highlighting "noise" and "silences" for each of them.
- Identify the problems involved when the corrections are erroneous (part-of-speech and structural ambiguities, etc.)

Figure 9: Exercises with spelling and grammar checkers.

The semester normally ends with a class on MT systems, which enables us to come back to important distinctions between MAT and MT. The ideal contexts in which MT systems may be used and the impact of these uses on the translation profession are also briefly discussed.

We next describe basic concepts of well known analysis-transfer-generation approaches of the kind introduced by Vauquois (1968) as well as related principles of new approaches such as example-based MT. Links are established with several of the concepts and much of the information introduced in connection with previous processes, and the focus is then narrowed to the difficulties associated with the transfer phase in particular. At this stage of the course, students have acquired the necessary background to understand the most important steps of the automation of the translation process and the difficulties and limitations that fully automatic MT systems may face. They also learn why those errors and limitations may compound each other, leading in a domino effect to gross translation mistakes. Examples of exercises are shown in Figure 10.

- Translate a short text with some MT systems (one on the workstations, the others available on the Internet)
- Compare machine outputs and for each classify translating errors into known categories (part-of-speech and structural parsing ambiguities, semantic and structural transfer ambiguities, word order and agreement errors, etc.)
- Choose the best output and rewrite it into an acceptable translation

Figure 10: Exercises with MT systems.

7. Concluding remarks

This course has been given for nearly a decade and has evolved considerably since the first time it was given. Certainly it is important to take into account advances in the field of translation computerization and remain up to date. However, the major motivation for the changes the course has undergone was to build a better understanding of the automation process itself. Before redesigning the course we used to teach the various applications in a modular fashion and failed to establish links between them. Students were simply learning to manipulate separate functions and did not have a general perspective on the field.

Using the translation process and basic concepts related to natural language processing as the backbone of the course appears to be the best solution to organize the material and to help students acquire the fundamentals in the field. A simple example will serve as an illustration of this last observation. At the end of the semester, during the sessions on proofreading and MT, most students are impressed by everything that is involved in producing the output and stop focusing exclusively on errors. They can also develop a critical evaluation and pinpoint specific problems instead of simply stating that the output results are "bad."

We think it is worthwhile to keep moving in this direction and it seems possible to continue incorporating easily new CAT applications as well as MT systems in this overall organization.

8. References

Austermühl, F. (2001) *Electronic Tools for Translators*, Manchester: St. Jerome Publishing.

Bowker, L. (2000) 'Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources.' *International Journal of Corpus Linguistics*, 5, pp. 17-52.

Bowker, L. (2002) *Computer-aided Translation Technology. A Practical Introduction*, Ottawa (Canada): University of Ottawa Press.

EURODICAUTOM (2003) *The European Commission's multilingual term bank.* European Commissions. (http://www.europa.eu.int/eurodicautom)

Le Grand dictionnaire terminologique (2003) *La base de données terminologiques de l'Office québécois de la langue française.* Gouvernement du Québec (http://www.granddictionnaire.com)

Kay, M. (1980) 'The Proper Place of Men and Machines in Language Translation.' Report CSL-80-11 Xerox Palo Alto Research Center, Palo Alto, CA. Reprinted in [1997] *Machine Translation*, 12, pp. 3-23.

Kenny, D. and A. Way (2001) 'Teaching Machine Translation & Translation Technology: A Contrastive Study.' *MT Summit VIII Workshop on Teaching Machine Translation*, Santiago de Compostela, pp. 13–17.

L'Homme, M.-C. (2000) *Initiation à la traductique*, Brossard (Canada): Linguatech.

L'Homme, M.-C. and B. Robichaud (1999) "*Outils informatiques des langagiers*," Course website, Université de Montréal.

Somers, H.L. (1999) 'Review-Article: Example-based Machine Translation.' *Machine Translation*, 14(2), pp. 113-157.

Somers, H.L. (2001) 'Three Perspectives on MT in the Classroom', *MT Summit VIII Workshop on Teaching Machine Translation*, Santiago de Compostela, pp. 25–29.

TERMIUM® (2003) *La base de données terminologiques et linguistiques du gouvernement du Canada*. Gouvernement du Canada. (http://www.termiumplus.com)

TS-RALI (2003) *TransSearch search tool on the Canadian Hansard bilingual corpus.* (http://www.tsrali.com)

Vauquois, B. (1968) 'A Survey of Formal Grammars and Algorithms for Recognition and Transformation Machine Translation.' *IFIP Congress-68*, Edinburgh, pp. 254-260.