Teaching commercial MT to translators: Bridging the gap between human and machine

Natalie Kübler

University Paris 7 Denis Diderot
kubler@ccr.jussieu.fr

Abstract

This paper presents the experiment that was led at the University Paris 7 to teach machine translation to translation students. Linking machine translation with translation-training in order to help students acquire the linguistic and technical skills increasingly required on the translation market has become part of a number of translation syllabuses in the last few years. Although, MT knowledge is more and more required for professional translators, most of them are still skeptical, not to say hostile to MT. Our objective here is then twofold: showing translation trainees why and how MT can be used for professional purposes, from the text to be translated to the final result.

1. Introduction

Nowadays, new needs are arising from the ongoing globalization process. The impact of globalization has been to augment cross-cultural communication, hence the need for intercultural language resource management, which is part of content management. Intercultural language resource management is the creation, processing, and management of multilingual and multicultural language resources. Therefore, new skills and competences are required from language professionals on the labour market, especially ICT, technology, and language resource management competences, as identified in the three following surveys: the EU-sponsored SPICE-PREP II report on content localisation (updated 2001), the LETRAC survey on the skill and competences needed by industry for translators, and the translators section of the LEIT Industry Needs Survey. Although the LETRAC project was criticised for its emphasis on technology, various authors have identified such a need more recently, and, what seems to be more and more essential, the necessity for "constant adjustments" (Torres del Rey 2000) to constantly developing technology and new business processes. Localization companies, such as Bowne Global Solutions have acquired machine translation (MT) systems. However, companies usually do not have qualified staff which can use or enhance MT systems. Some international organisations, such as the European Community for example, have been using MT for years in their translation tasks. Therefore, linking MT with translation training in order to help students acquire the linguistic and technical skills increasingly required on the translation market has become part of a number of translation syllabuses in the last few years, as shown for example on the LISA Education Initiative Taskforce website.[1] Although, MT knowledge is more and more required for professional translators, most of them are still skeptical, not to say hostile to MT.

Our objective here is then twofold: showing translation trainees why and how MT can be used for professional purposes, from the text to be translated to the final result.

This paper presents the two teaching scenarios that attempted to achieve those objectives and the results obtained. Section 2 deals with the academic context in which the experience was led. In section 3, the projects, data and tools used will be described. Section 4 explains how the projects were led, describing in details the different stages necessary to achieve a translation task, using MT, which will lead to the conclusion.

2. Academic Context

The experiment that was led involved post-graduate students following a one-year training program, which is called DESS ILTS[2] and divided into two options:

[1] http://www.lisa.org/leit/

[2] Diplôme d'Etudes Scientifiques Spécialisées en Industrie des Langues et Traduction Spécialisée, i.e. equivalent of the last year of a MA in language industry and specialised translation.

– *Specialised translation*, with more translation courses, and two languages other than French (English and either German, Spanish, or Portuguese); translation courses are more oriented towards business and economy, i.e. students have little knowledge in technical domains;

– *Language industry*; the languages are English and French; students have more technical courses, such as mark-up languages (XML, advanced HTML), SQL and PHP, Unix systems, linguistics (syntax and text linguistics) technical translation, technical writing, and localization (using tools such as Catalyst).

This translation training is semi-professional, since students spend every other week as interns with a private company, either a translation/localisation company or corporate companies which have a language management department.[3]

Students usually come from translation departments or translation schools; they must have a "Maîtrise" level (French four-year degree). Most candidates have basic computer skills (word processing and Internet knowledge), basic knowledge in linguistics, and no knowledge in natural language processing or MT. Around eighty candidates are selected according to tests and to their curriculum vitae; they have four months to find a position as interns with one of our industrial partners. The first forty students who find a position are accepted in the curriculum.

3. Projects, Data, and Tools

In order to show students how MT can be used in professional translation, they had to carry out projects through the whole workflow: from receiving the task to submitting the final result to the customer. Various language data and tools were used to achieve the projects.

3.1. Projects

As our syllabus aims at allowing students to work not only as translators, but also, more generally, as language professionals around translation, the projects included different tasks, and not only translation. The syllabus is divided into two options that have slightly different orientations, as explained in section 2. For this reason, the projects the two options had to realise were slightly different, according to the different skills the students had to acquire.

The specialised translation group (TS) had to translate the Linux HOWTOs as yet untranslated into French. The HOWTOs are the user manuals of the Linux operating system, and are thus highly technical texts. They have been translated into different languages, and the French Linux community is very active in translating those.

The aims were to teach students how to:

- Use an MT system and improve translation results by adding appropriate bilingual dictionaries;
- Use immediately available resources, such as web-based bilingual glossaries, self-made or web-based corpora, and term extraction software that didn't require specific computing skills;
- Carry out a translation project with an MT system, as is done in large governmental organizations, and submit it to the "customer" (the French Linux community).

The language industry group had to translate dictionary entries from the Free On-Line Dictionary of Computing,[4] while a second part of the project consisted in analyzing Systran's translation problems and sending the Systran's linguistics team their reports. Students were divided into small groups, and each group had to present their result on a web site.[5]

The objectives for this group were slightly different, as they had to learn how to:

- Use and customise an MT system to improve translation results
- Use available resources and tools to create bilingual term bases

[3] Some examples: AFNOR, Aventis, Bowne Global Solutions, IBM, France Telecom, General Electrics Medical Systems, Institut National de la Recherche en Agronomie, SYSTRAN, Messier Dowty, Union Internationale des Chemins de Fer, Thales Communications, Epson, Alcatel CIT, Nexans, Sun Microsystems, Renault, Thompson, Aérospatiale.

[4] FOLDOC, http://www.foldoc.org, http://wall.jussieu.fr/foldoc.

[5] It will be accessible on http://wall.jussieu.fr.

- Carry out a language industry project and submit it to the "customer" (Systran's linguistics team).

To achieve the projects, students had access to language data, such as monolingual and bilingual corpora and on-line term bases; they also had to use term extraction tool and MT.

3.2. Language Data

The projects implied specialised dictionaries creation, that were to be used by the MT system. Corpus linguistics has shown for the last decades, that linguistic information extracted from corpora could be reliable enough. Zanettin (1998), Yuste (2001), or Kübler (2002) for example, have shown the benefits of using corpora in translation teaching in general. This is even more true in the case of creating machine customised dictionaries to be compiled into an MT system. Only information extracted from corpora can be reliable enough to yield the best MT results possible. As shown in Senellart (2001), MT language resources can be augmented by extracting information from parallel specialised corpora. Systran have been augmenting their language resources on a large scale by using corpora for several years.

Languages for specific purposes, such as computer science contain numerous specialised terms that cannot be found in general dictionaries. On the other hand, specialised dictionaries are often incomplete and obsolete. In quickly evolving fields, such as computer science, new terms are created everyday, and specialised dictionaries cannot be updated quickly enough. Furthermore, specialised dictionaries usually contain almost only nouns, as a term, according to the canonical definition, that were in use until recently, could only be a noun. Various authors, such as L'Homme (1998) for example, have shown that verbs, adjectives, or adverbs could be considered as terms as well. Building customised MT dictionaries also implies finding the translations of the terms, and finding phraseological information on predicate nouns, adjectives, and verbs.

This kind of information can be extracted from corpora. Students had therefore access to various corpora to help them find the necessary information. The corpora are available on a Web-based interface, that was developed at the University of Paris 13 to teach computer science English to French students (Foucou and Kübler 1998), and that is accessible at the University Paris 7 (http://wall.jussieu.fr). Available corpora for the computer science are the following:

- The *Linux HOWTOs*: A parallel corpus that contains the English HOWTOs that have already been translated into French, aligned at the section level with their French translation. Each language contains ca. 500'000 words.
- The *Internet RFCs*: A monolingual English corpus consisting of the Internet Request For Comments, a highly technical collection of texts containing ca. 8,5 million words.
- *FOLDOC*: The Free On-Line Dictionary of Computing is also accessible as a corpus, and can be queried in the same way as the other corpora.
- A series of smaller corpora, that were collected by students, (between several tens to several hundreds of words) on various sub fields of computer science, such as economic intelligence, digital camera, games, peripherals, etc.

Students also have access to small "general language" corpora to check the differences between the use of a term in LSPs and in general language.[6]

- *The Times* : 3.5 million words
- *The Herald Tribune*: 1.5 million words
- *Le Monde*: 1 million words

Students also had access to freely accessible on-line dictionaries, such as *FOLDOC, Le grand Dictionnaire Terminologique*,[7] the French *HOWTO* community bilingual dictionary,[8] the French Committee for Terminology in Computer Science glossary,[9] etc.

---

[6] These corpora are not big enough to be representative of the language in general, as are the BNC or the Bank of English.

[7] http://www.granddictionnaire.com

[8] http://launay.org/HOWTO/Dico.html

[9] http://www-rocq.inria.fr/qui/Philippe.Deschamp/RETIF/19990316.html

3.3. Tools

Students had to learn to use several tools that are quite user-friendly, and do not require specific computing skill.

Terminology Extractor (*TE*)[10]

TE is a commercialised term extraction tool that works for English and French. It allows the user to extract the following information: A list of all the words that have been recognised by the embedded dictionaries or by the dictionaries the user can integrate into the system, all the "non-words", i.e. words that are not in the dictionaries, and collocations. Collocations are defined in the system as being all sequences of two to ten words (excluding stop-words) that are found at least twice in the text. The tools include a concordancer that gives concordances for the words, non-words, and collocations that have been detected by the system. Figures 1 and 2 show examples of non-words and collocations that are extracted by *Terminology Extractor*. In Figure 1, apart from *Dennis* and *accelleratw*, all the words are terms or product names in the computer science area.

Debian	Netscape	accellerate
Permedia	Dennis	XFCE
RedHat	Dialogs	Corel
RgbPath		FAQs
ServerFlags	Howto	Microdoft
ServerLayour	README	Linux
XkbLayout	XkbModel	RealAudio
Solaris		ISA
UI	KDE	GUI
USB	LeftOf	IRQs
WindowMaker	ModulePath	NFS

Figure 1: Result of the non-words extraction from a HOWTO document.

In-house concordancer

The corpora students have access to can be queried using a Web-based concordancer using Perl-like regular expressions. With the parallel *HOWTO* corpus, concordances give access to the aligned English/French sections from which the string (word or regular expression) has been extracted, allowing thus users to find the French equivalent of an English term that cannot be found in specialised dictionaries or terms bases,

Internet Gateway 3	{ Looking look } at the Network 3
IP aliasing 3	**name server** 4
ISA { card cards } 3	**Network { Device devices** } 4
latest version 3	**Linux computer** 3
DHCP Server 15	IP **{ addresses address** } 16
Linux gateway 3	**Linux box** 16
modules file 3	card on the Linux box 4
scripts / ifcfg 3	**DNS { Server servers** } 17
server will start 3	**interface configuration file** 3
{ **Network networking** } **{ Card Cards** } 12	

Figure 2: Results of a collocations extraction from a HOWTO document. The words in boldface are actual terms.

as in example (1).

(1) buffer overflow = débordement de mémoire tampon

MT

A customised version of Systran's Systranet is used with the students. This specific interface offers pedagogical features that do not appear on the usual Systranet site. The usual features that are offered on Systranet contain the dictionary manager that allows the user to customise Systran's MT system by creating specific dictionaries containing linguistic information: During the translation process, these dictionaries are applied prior to Systran's. Our Systranet access allows the teacher to create groups of students, to have access to all the students' dictionaries, and to check the use of the system. Systranet offers the features of aligning the source and target texts, highlighting in green the words that have been found in the customised dictionaries, and in red the unknown words, i.e. not in any – Systran's or user's – dictionary.

The customisable dictionaries allow the users to introduce linguistic information, as shown below:

Part-of-speech information: basic part-of-speech information can be attached to the entries, such as verb, noun, proper noun, adjective, and "sentence", which deals with

[10] http://www.chamblon.com

adverbs, adverbial phrases, or whole idioms, such as *your mileage may vary*.

Syntactic information, such as the governed prepositions for nouns, verbs, and adjectives, or direct objects for verbs. A verb which governs a preposition is shown in example (2).

(2) access (verb) (noprep)=accéder (verb)(prep:à)

Semantic information, such as the conceptual class of the possible direct object of a verb, as shown in example (2). In this example, the coding for the verb *run* indicates that the direct object must belong to the semantic class [OS], which means all terms sorted under the "operating system" class. Below the verb, the noun *Unix* is marked as belonging to the [OS] class. This means *Unix* can be the direct object of *run*.

(3) to run (verb)(context:OS)
Unix (noun) (SEMCAT:OS)

Morphological information, such as the plural form of a noun in any language, the gender of a noun in French, or forcing the number in the target or source language. Example (4) shows how the gender of *cache* can be forced to masculine. In general French, the noun *cache* ('hiding place') is feminine, whereas in computer science French, it is masculine, and means 'cache'.

(4) cache(noun) = cache (noun) (masculine)

The acronym *RFC* builds a plural in *-s* in English, i.e. *RFCs*, whereas in French, it is invariable; this type of information can be coded in the dictionary, as is shown in example (5).

(5) RFC (noun) (plural:RFCs) = RFC (noun) (plural:RFC)

Translational information, such as "DNT", which means that the string must not be translated and stay as it is in the translation process. This feature is quite useful in computer science, as there are command names, for example, that are never translated, such as the Unix command *cd*, or *mkd*.

Figure 3 shows a dictionary sample, in which various types of coding are presented.

MT was introduced in the context of an introductory course to corpus linguistics and its application to translation and terminology, which is compulsory in both the options. In addition, students also had an introduction to MT, with specific focus on the latest release of Systran, in order to help them understand how the system works. This was essential for the analysis phase of the project, as students had to make hypotheses on the system's "mistake" and suggest possible solutions to improve translation results.

"AT&T" (company name)
byte-ordering (noun)=ordre des octets (noun)
cache (noun)=cache (noun) (masculine)
based (adjective)(noprep)=architecturé (adjective)(prep:autour)
basic language constructs (noun) (plural)=base de construction du langage (noun) (singular)
keyable (adjective) (prep:to)=sensible (adjective) (prep:à)
to log in (verb)=se loger (verb)
to introduce (verb) (context:extensions)=introduire
to carry (verb)(context:digital data)=transmettre (verb)
to prevent (verb)(context:fighting)=interdire(verb)(context :le combat)

Figure 3: Dictionary sample; for the sake of clarity, English terms are in boldface.

4. Project Workflow

Students had to achieve their projects working in groups of three people. Each group had to translate 3,000 words. The Specialised Translation groups had to translate some as yet not translated *HOWTOs* divided among the groups. The Language Industry groups had to translate several entries of the *FOLDOC* dictionary. For those groups, the project results had to be presented on a website. The project had to be achieved trough several stages. First, creating and testing customised dictionaries, then linguistic analysis of translation outcome, finally, post-editing MT results, and project presentation on a website.

4.1. Creating Customised Dictionaries

Customised dictionary creation and testing must be done in two steps.

Step one: The first step consisted in extracting term candidates from the documents to be translated, using Terminology Extractor. Then, students had to decide on the terms out of the list of term candidates, and to find the French equivalents of the terms, making use of all the

linguistic resources available, i.e. corpora, dictionaries. Finally, they had to find the linguistic information in the various corpora, in order to complete the customisable dictionary. This first-step dictionary was then compiled into the system to proceed to the first translation tests.

Step two: During this test phase, students had to analyse translation results using Systranet alignment feature. This allowed them to complete or modify the linguistic information included in their customised dictionaries. Once changes were done, translation results would be analysed again, and dictionaries would be modified untils no more change could improve translation results.

Figure 4 shows the difference between MT without and with customised dictionaries.

4.2. Linguistic Analysis

Students also had to analyse and describe the system's "errors", using and completing an error typology that was given them. The linguistic analysis consisted in examining the translation results, according to the following very basic error typology, which must be understood as decribing wrong or missing linguistic information:

- **Lexicon**: the word is a term and has therefore a different translation in French for example; the word is not in the dictionary.
- **Lexicon-grammar**: verb subcategorisation, or prepositional information that must be attached to nouns or adjectives.
- **Morphology**: wrong morphological information, such as grammatical gender in French,or morphosyntactic problems, such as subject–verb agreement or adjective–noun agreement in French.
- **Syntax**: POS ambiguïty, NP: determiners, NP coordination, transformations / ellipsis / cleft sentences / PP attachment
- **Semantics**: Polysemy, semantic classes in verb arguments position.
- **Pragmatics**: book titles, play titles.
- **Text structure**: problems related to text genres; a dictionary has specific structures that include sentences without verbs.
- **Format**: This category is specific to the MT system, which has problems dealing with parentheses, square brackets, etc.

Source text	This page contains a simple cookbook for setting up Red Hat 6.X as an internet gateway for a home network or small office network.
Without customized dictionary	Cette page contient un cookbook simple pour le chapeau rouge 6X d'établissement en tant que Gateway d'Internet pour un réseau à la maison ou le petit réseau de bureau.
With customized dictionary	Cette page contient un **livre de recettes** simple pour l'établissement ***Red Hat 6.***X en tant que ***passerelle Internet*** pour un ***réseau domestique*** ou un petit ***réseau de bureau***

Figure 4: Comparing translation results with and without customised dictionaries

- **Odds and ends**: This category deals with the system bugs, which can for example repeat a letter at the beginning of a word.
- **Style**: A problem that MT cannot solve at this point.

Linguistic analysis is very efficient in terms of pedagogical approach. It required from students to actually understand how the system works and how to name the problems from a linguistics point of view. As most students had very little background in linguistics and NLP, this phase helped them acquire linguistic knowledge from a practical point of view.

4.3. Post-editing

Once customised dictionaries and linguistic analysis were over, students had to post-edit Systran's final translation results. As the system always has the same problems, some errors could be corrected most easily, as they were repeated all along the target text. The objective of post-editing was to obtain a correct text, without modifying too much the MT results, which means that the style could need some improvement.

Figure 6 shows an example of a section of text that demanded little post-editing. The example shown in Figure 7 has undergone more corrections during the post-editing phase.

Source text	MT result	Post-edited text
The sender uses a one-way hash function to generate a hash-code of about 32 bits from the message data	L'expéditeur emploie une fonction de hachage à sens unique pour produire **d'un** code de hachage d'environ 32 bits **des** données du message.	L'expéditeur emploie une fonction de hachage à sens unique pour produire **un** code de hachage d'environ 32 bits **à partir des** données du message.

Figure 6: The modified parts are in boldface.

Cryptography The practice and study of encryption and decryption – encoding data so that it can only be decoded by specific individuals.	Cryptographie La pratique et l'étude de chiffrement et de déchiffrement – codage de données de sorte **qu'elle puisse** seulement être décodée par les personnes autorisées.	Cryptographie Techniques de chiffrement et de déchiffrement, ou codage de données de sorte que seules les personnes autorisées **puissent** décoder ces données.

Figure 7: Important post–editing of MT results.

5. Conclusion

These experiences enormously helped students to grasp the advantages of MT AND of human translators. They completely changed their point of view towards MT, understanding how MT works, and why the results are not the same as translation done by human translators. The language industry group was also able to make correct assumptions about the system's "errors", and how to correct them.The specialized translation group, which was at the beginning very skeptical, discovered how customised MT could be used in industrial translation. They also managed to grasp why "general purpose" MT was very difficult to achieve.[11]

At the end of the project, students had also mastered the use of various tools for professional translation, developed cross-linguistic awareness, and acquired a good knowledge in specialised lexico-grammar and syntactic structures.

References

Foucou, P.-Y. and N. Kübler (2000): "A Web-based Environment for Teaching Technical English", in L. Burnard and T. McEnery (eds) *Rethinking Language Pedagogy: papers from the third international conference on language and teaching.* Frankfurt am Main: Peter Lang GmbH. 65-73.

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

Johns, T. (1988), "Whence and wither classroom concordancing?", in T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker (eds) *Computer Applications in Language Learning*. Dordrecht: Foris. 9-27.

Johns, T. (1993), "Data-driven learning: An update"', *TELL & CALL* 3.

Kübler, N. (2002): "Creating a Term Base to Customise an MT System: Reusability of Resources and Tools from the Translator's Point of View". In *Proceedings of the Language Resources for Translation Work and Research. Workshop of the LREC Conference*, Las Palmas de Gran Canarias, 44-48.

Kübler, N. (forthcoming). "How Can Corpora Be Integrated Into Translation Courses?" in Zanettin, F., S. Bernardini and D. Stewart (eds), *Proceedings of CULT2 (Corpus Use and Learning to Translate). Corpora in translator education*. Manchester: St Jerome.

LEIT: The LISA Education Initiative Taskforce, http://lisa.org/leit/pubs/industrysurvey.htm

LETRAC: Language Engineering for Translators Curricula, http://www.iai.uni-sb.de/LETRAC/home.html

L'Homme, M.-Cl. (1998): "Définition du statut du verbe en langue de spécialité et sa description

[11] They stopped laughing at MT results, without knowing what it meant.

lexicographique", *Cahiers de lexicologie* 73 (2), 61-84.
Pearson, J. (1998) *Terms in Context*. Amsterdam: John Benjamins.
Senellart, J., P. Dienès and T. Varadi (2001): "New Generation Systran Translation System", in *Proceedings of the MT Summit VII,* Santiago de Compostela.
SPICE-PREP II: Export potential and linguistic customisation of digital products and services: www.hltcentral.org/usr_docs/spice/ spice_final_report.pdf
Yuste Rodrigo, E. (2001): "Making MT Commonplace in Translation Training Curricula – Too Many Misconceptions, So much Potential", in *Proceedings of the MT Summit VII,* Santiago de Compostela.
Zanettin, F. (1998): "Bilingual Comparable Corpora and the Training of Translators", *Meta*, 43(4), 616-630.