Creating a Term Base to Customise an MT System: Reusability of Resources and Tools from the Translator's Point of View

Natalie Kübler

Intercultural Centre for Studies in Lexicology
University Paris 7
2, Place Jussieu, 75251 Paris Cédex 05, France
kubler@ccr.jussieu.fr

Abstract

This paper addresses the issue of combining existing tools and resources to customise dictionaries used for machine translation (MT) with a view to providing technical translators with an effective time-saving tool. It is based on the hypothesis that customising MT systems can be achieved using unsophisticated tools, so that the system can produce output of sufficient quality for post-translation proofreading. Corpora collected for a different purpose, together with existing on-line glossaries, can be reused or reapplied to build a bigger term base. The Systran customisable on-line MT system (Systranet) is tested on technical documents (the Linux operating system HOWTOs), without any specialised dictionary. Customised dictionaries, existing glossaries completed by adding corpus-based information using terminology extraction tools, are then incorporated into the system and an improved translation is produced. The dictionary will be augmented and corrected as long as modifications generate significant results. This process will be described in detail. The resulting translation is good enough to warrant proofreading in the normal way. This last point is important because MT results require specialised editing procedures. Compared with the time taken to produce a translation manually, this methodology should prove useful for professional translators.

1. Introduction

The growth in the volume of documentation for translation and the constant enhancement of tools have brought about great changes in the world of translation. Corpus linguistics has opened up new perspectives for both translation studies and the process of translating. As Baker (1993) pointed out as early as 1993, corpora can offer new insights into the theoretical and practical aspects of translation. The different stages in which various types of corpora can help in the translation process have been investigated by Aston (2000), while Varantola (2000) evaluates the use of dictionaries and specialised corpora, and other researchers investigate issues in the area of translator training, which is currently undergoing deep changes. The use of corpora and MT in the translation classroom has become a subject in its own right (Zanettin 1998; Yuste 2001, and Kübler forthcoming).

The translator is no longer seen as an isolated individual, working with a paper dictionary. A range of new resources are available for translators, particularly for translating technical documents[1]. However, there is a fear that machines, especially MT, will eventually replace translators[2]. MT has already changed the way professional translators work, but will not replace human beings. Today, it can be used as a tool to provide translators with quick on-the-fly versions that need thorough proofreading. The experiment described in this paper deals with the **next step**: Customising MT systems to provide translators with **a time-saving tool producing good quality results**.

We shall show how MT systems can be customised using existing resources, such as on-line glossaries and existing or self-made corpora, initially collected for a different purpose. A combination of resources, such as terminology extraction and conventional corpus linguistics tools, can be applied in the building of complete dictionaries containing sophisticated linguistic information. The recycled resources will be described, together with the tools used. The Systran user-customisable on-line MT system is then presented, with the linguistic features that can be integrated. The methodology applied in the creation of new dictionaries is detailed, and samples of improved translations are provided. A time-based evaluation of manual and MT outcome is included. The conclusion points to some work that remains to be done.

2. Resources

The project was carried out by **recycling** existing language resources, and using on-line Web-based resources. The tools that were used are simple to implement and do not require specific programming knowledge. The language resources that are readily available for assembling dictionaries can be divided into three categories:

- on-line bilingual technical glossaries;
- monolingual and parallel technical **corpora**;
- the Web as a corpus[3].

In this computer-science-based project, all three types of language resource were used .

2.1. Bilingual glossaries

On-line Web-based bilingual glossaries generally propose aligned lists of English terms and equivalents in French. These dictionaries are normally small, containing a few hundred headwords, usually with few verbs, adjectives or multiword units. They do provide useful lists of bilingual entries in the specialised area of computing,

[1] Translation memory, term extraction tools, term base management software can all help when translating Languages for Specific Purposes (LSP), including Web sites, user manuals, help files, and financial documents.

[2] *Ouaibe et traduction: que craindre du Systran?* http://www.geocities.com/aaeesit/art21.html

[3] i.e. making linguistic queries with search engines, and search tools like WebCorp (see section 2.3. below).

though they partly have the same headwords. Three glossaries were selected initially, because they contain terms that do not cross LSPs because they are domain-specific. They were downloaded, corrected, and formatted, to be compiled as customised dictionaries in Systranet. Here is the list of selected glossaries and the number of headwords for each:

- The HOWTO translation project glossary[4]: a small glossary of 200 words discussed and agreed upon in the project discussion list .
- Netglos Internet Glossary[5]: a multilingual glossary of Internet terminology compiled in a voluntary, collaborative project, containing 282 terms.
- The RETIF[6] site glossary. This short glossary contains 73 terms approved of by the French Governmental Terminology Commission for Computing and the Internet.

2.2. Corpora

Corpora make up the core resource exploited by the Systran team. Smaller corpora, exploited with simple tools, produce interesting results on a more individual scale. The smaller corpora used in the experiment had been collected to teach computer science English to French-speakers (Foucou & Kübler 2000). The texts used are highly technical and freely available on the Web:

- Internet RFC[7]: 8.5 million words: monolingual English corpus. This corpus consists of the Internet Request For Comments available on the RFC documentation site.
- Linux HOWTOs: English to French aligned corpus, ca. 500 000 words. The English HOWTOs and their translations in several languages are available on the Linux documentation site[8].

The above-mentioned corpora are embedded in a Web-based environment that can be accessed on our Wall[9] site.

2.3. The Web

The Internet has become a necessary resource for linguists, lexicographers, translators, and other language researchers, providing them with on-line dictionaries, reference documents, newsgroups. The Web can also be considered as an open-ended, unstructured corpus which can be queried using search engines, though these are not tailored for linguistic search. A specific linguistic search tool is Webcorp[10] (Kehoe & Renouf, forthcoming), which provides users with concordances, collocates, and lists of words found on Web pages; we have used this for a variety of purposes. A Web-based search strategy should be used in conjunction with the off-line, finite, corpus-based approach, since they yield complementary information.

2.4. Tools

The first tool used is an on-line concordancer featuring perl-like[11] regular expressions, which gives access to aligned paragraphs of French and English texts from which a concordance has been extracted. Another on-line tool is a tokeniser, which allows the user to sort the words of a text in alphabetical order, or by frequency.

As the general philosophy of this experiment was to use simple tools, a commercially available term extraction tool was selected: Terminology Extractor[12], which works for French and English. It uses a dictionary to lemmatise the vocabulary of a text and produce four different output types:

- *Canonical forms:* recognised by the program and sorted by alphabetical order or by frequency; the most frequent forms are to be considered as potential terms.
- *Non words*: not recognised by the system; most of them are specialised terms.
- *Collocations*. Collocational extraction is based on a very simple principle: any sequence of at least two -- and at most ten -- words, that is repeated at least once is considered as a collocation. Stop words are discarded to avoid sequences, such as *sauvegarde de la* [save the], in which *la* is a determiner preceding the second part of the term, as in *sauvegarde de la configuration* [save the settings]. Collocates are good candidates for technical terms.
- *KWIC (key word in context):* for the combined three lists. This feature is used to extract lexico-grammatical information, on verb structures, for example.

3. Systranet: customisable dictionaries

Systran MT has been much improved in recent years (Sennelart et al. 2001). Systranet is an on-line service offered by Systran. Users have access to a dictionary manager which allows them to create and upload their own multilingual linguistically-coded dictionaries into Systran, in order to improve translation results. These multilingual dictionaries contain a list of subject-specific terms that are analyzed prior to using Systran in-house dictionaries. This feature is based on the assumption, demonstrated by Lange & Yang (1999), that domain selection and terminology restriction are beneficial to translation results.

Linguistic information, such as part-of-speech, number and gender, subcategorisation, or low-level semantics can be added to the user's dictionary entries. Once the dictionary has been compiled, its accuracy and linguistic coverage can be tested by translating subject-specific texts.

The translation results can be improved by modifying the dictionary, a recurrent process which can be continued so long as the modifications produce significant improvement. Systranet offers specific features that allow the user to see which terms have been translated using customised dictionaries, and which terms are not recognised at all. It allows the user to check whether the dictionary entries have really improved the translation

[4] http://launay.org/HOWTO/Dico.html
[5] http://wwli.com/translation/netglos/
[6] http://www-rocq.inria.fr/qui/Philippe.Deschamp/RETIF/19990316.html
[7] http://www.rfc-editor.org/rfc.html
[8] http://www.linuxdoc.org
[9] http://wall.jussieu.fr
[10] http://www.webcorp.org.uk

[11] Perl is a particularly appropriate programming language for handling word strings or finding language patterns.
[12] http://www.chamblon.com

results as expected. Another feature used to complete the dictionary is the non-word feature: all the words that have not been recognised by Systran or the user's dictionaries appear in red. They can then be integrated into the user's dictionary.

4. Experiment and methodology

We chose technical documents written by experts for experts, the Linux HOWTOs, which are the user manual of the Linux operating system. This experiment is part of a larger project that consists in translating all the new HOWTOs using MT. HOWTOs are documents of various size, describing the way to install the system and software related to it. Existing software is constantly updated and augmented, so the corresponding documents are updated and new documents are written with each new program. These documents have been translated into several languages by the various Linux communities. The French Linux community has developed a translation project[13] in which the translation is usually done by non professional, voluntary translators. People choose the document they want to translate and do the job. Today, most HOWTOs have been translated, which makes it possible to align the French translations with the English source and use them as a parallel corpus.

The task set for the experiment was to provide a complete and appropriate dictionary to translate the remaining untranslated Linux HOWTOs. This is based on the assumption that the initial dictionaries will be augmented in the light of each new text to translate. Since a comparative study of the translation results -- with and without customised dictionaries -- had to be established, each text was first translated without using any specific dictionary.

4.1. Creating the dictionaries

The methodology is a combinatorial approach, recycling data and using terminology extraction tools.

First, the three glossaries mentioned above were downloaded and converted into dictionary files, augmented with linguistic information, giving more than 500 entries. These glossaries were selected when translating a HOWTO. Then, a more complete and corpus-based approach was applied. It produced two types of dictionary: *step-one dictionary* and *step-two dictionary*.

4.1.1. Step-one dictionaries

The step-one dictionaries were created using term extraction software, corpora, and a concordancer. This sort of dictionary can be produced using large corpora, but the most efficient solution for the individual user is to apply it to the texts to be translated.

The candidate texts were processed using Terminology Extractor. Initial candidates for headwords in the dictionaries were selected from the non-word and collocation lists. Unlike the existing glossaries, Terminology Extractor outputs do not provide French equivalents for the English words. On-line term banks, such as *Le Grand Dictionnaire Terminologique*[14] or *Termium*[15] proved insufficient for translating most terms. A corpus-driven approach was adopted to find French equivalents: the RFC corpus was used to find more information about context, the aligned HOWTO corpus was queried with the regular expressions concordancer (Wall) to find appropriate translations, as illustrated below.

The term README in the computing context is used as a noun, as shown in the following context, in which the term is the head of a subject NP:

links which Linus describes in the **README** are set up correctly. In general, if a

Figure 1. The noun *README* in context

The term *addon* was in the non word list, but by using the HOWTO corpus, we found contexts and a French translation:

The FWTK does not proxy SSL web documents but there is an **addon** for it written by Jean-Christophe Le fwtk ne route pas les documents web SSL, mais il existe un **module complémentaire** écrit par Jean-

Figure 2. The noun *addon* and its French translation

This stage was necessarily completed by using Web search engines to verify some translations found in the HOWTOs, or to deduce new translations from indirect queries. Since the documents are translated by various people who are usually not professional translators, but computing experts, the French versions of the HOWTO are not homogeneous. This means that one English term can be translated by several different words that are true synonyms in French. Only one equivalent must be chosen for the MT dictionary. Another problem is the case of borrowings. In spoken computing French, the English term is often used. Even in written texts, and especially in translations, usage leads translators to keep the English term and give the French equivalent once at the beginning of the document.

When no answer can be found in the HOWTO corpus, WebCorp can provide solutions. By looking for collocates and concordances for an English term in French language documents, possible translations can be traced back to the French sites. The collocates of *network* in French-speaking sites, for instance, allowed us to trace back *home network* and the French *réseau domestique* (Kübler, forthcoming).

4.1.2. Step-two dictionaries

Once a set of dictionaries has been produced for each HOWTO, it must be tested not only to correct possible errors in the entries, but also to add the new words that are neither in Systran's nor in the customised dictionaries. The more HOWTOs are translated, the fewer words have to be added until the dictionaries are saturated, i.e. no new word can be added to improve translation results.

Step two is illustrated with the Home-Network-Mini-HOWTO, one of the not yet translated HOWTOs. Below is an example of translation results with and without customised dictionaries:

[13] http://www.traduc.org
[14] http://www.granddictionnaire.com
[15] http://www.termium.com

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 cust. dict.	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 cust. dict.	Cette page contient un cookbook simple pour l'établissement ***Red Hat 6.***X en tant que ***passerelle Internet*** pour un ***réseau domestique*** ou un petit ***réseau de bureau***

Fig. 3: Comparing translation results with and without customised dictionaries

In the next table, the customised dictionaries were completed with the words badly or not at all translated with the first version of customised dictionaries.

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.***
Step-one dict.	Cette page contient un **cookbook** simple pour l'établissement ***Red Hat 6.***X en tant que ***passerelle Internet*** pour un ***réseau domestique*** ou un petit ***réseau de bureau***
Step-two dict.	Cette page contient ***des recettes*** simples pour ***l'installation Red Hat 6.***X en tant **que** ***passerelle Internet*** pour un ***réseau domestique*** ou un petit ***réseau de bureau***.

Fig. 4: Comparing translation results with step-one and step-two dictionaries

4.2. Translation outcome

Comparing the translation outcome with and without customised dictionaries shows encouraging results. Testing existing customised dictionaries on another text in the same subject area demonstrates that the text-based dictionaries can be reused, and that fewer headwords have to be added. Little by little, translators can add to their own dictionaries in various LSPs.

Obviously, as in any translation process, those translation results must be proofread. However, the points that need correcting are quite different from a translation done by a human being. If the MT errors are obvious and often serious, they have the advantage of always occurring in the same context. Most errors in this particular MT system are due to the same syntactic failures and can easily be corrected by the translator, once recognised.

Conjunction and disjunction are two of the main problems in MT systems that have yet to be solved. The garbled translation is however easily corrected, since the errors are similar each time a conjunction or a disjunction appears in an NP context:

Source text	*Translation result*	*Correct transl.*
Your internal and external networks	votre interne et des réseaux externes	vos réseaux interne et externe
a fulltime Cable or ADSL connection	une connexion en continu d'AADSL	une connexion en continu par le câble ou l'ADSL

Fig. 5: Conjunction and disjunction in an NP context

Another characteristic of MT systems is the overgeneralisation of transfer rules which leads to errors. Again, it is quite easy to check and correct those errors, for instance, the system translates a zero article in English by a definite article in French, although, in most cases, it should be the indefinite article:

Source text	*Translation result*	*Correct transl.*
decoded by specific individuals	décodé par les individus spécifiques	décodé par des individus spécifiques

Fig. 6: An example of transfer rule overgeneralisation

4.3. Human vs machine?

We selected two HOWTO totalling 9357 words in English. The expansion coefficient (15% in French) brings the total up to 10 750, i.e. ca. 36 standardised pages. This should take a professional translator from 5 to 7 days, depending on the tools used. Systranet took less than two minutes to produce an outcome. Professional translators assess the proofreading necessary at ca. 2 days. MT can therefore be included in the set of tools professional translators can actually use.

5. Conclusion

It has been demonstrated that the quality of translation can be significantly improved by importing customised dictionaries. Individual translators can thus create their own customised dictionaries with user-friendly and publicly available resources and tools.

These dictionaries recycle already existing resources, and their upgrading is corpus-driven. Translators working in LSPs can take advantage of a customised MT system because they can obtain quickly translated texts, and proofread them in a short time, as the errors generally have similar morpho-syntactic patterns. Although considerable work needs to be done in the beginning, after processing a few documents, the dictionaries are more or less saturated, and just a few words have to be added.

Further work will focus on reusing customised dictionaries to translate cross-LSP texts, such as digital cameras. More testing on the coding of Systranet customisable dictionaries is currently being done with students to improve coding rules and their applications.

6. References

Aston, G. 2000. I corpora come risorce per la traduzione e per l'apprendimento. In Bernardini S., Zanettin F. (eds) *I corpora nella didattica della traduzione*, Bologna: Cooperativa Libraria Universitaria Editrice Bologna, 21-29.

Baker, M. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Baker, M., G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: in Honour of John Sinclair*, Amsterdam and Philadelphia: John Benjamins, 233-250.

Foucou P.-Y. et Kübler N. 2000. A Web-based Environment for Teaching Technical English. In Lou Burnard and Tony McEnery (eds.) *Rethinking Language Pedagogy: papers from the third international conference on language and teaching.* Frankfurt am Main: Peter Lang GmbH.

Kehoe, A. & A. Renouf (forthcoming) `Webcorp: Applying the Web to Linguistics and Linguistics to the Web'. In Proceedings of the WWW 2002 Conference, Honolulu, Hawaii, 7-11 May 2002.

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

Kübler N. (forthcoming-b). In Aijmer, K. (ed) forthcoming Proceedings of 21st ICAME Conference, Univ. Gothenburg, May 22-26 2002, Amsterdam & Atlanta: Rodopi.

Lang E. & Jin Yang 1999. Automatic Domain Recognition for Machine Translation. In *Proceedings of the MT Summit VII,* Singapore.

Renouf, A.J. (forthcoming). WebCorp: providing a renewable energy source for corpus linguistics, in Granger, Sylviane and Stephanie Petch-Tyson, (eds) *Extending the scope of corpus-based research: new applications, new challenges*,. Amsterdam & Atlanta: Rodopi.

Senellart, J. Dienès P., Varadi T. 2001. New Generation Systran Translation System. In *Proceedings of the MT Summit VII,* Santiago de Compostela, 18-22 September 2001.

Varantola, K. 2000. Translators, dictionaries and text corpora. In Bernardini S., Zanettin F. (eds) *I corpora nella didattica della traduzione*, Bologna: Cooperativa Libraria Universitaria Editrice Bologna, 117-133.

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, 18-22 September 2001.

Zanettin, F. 1998. Bilingual Comparable Corpora and the Training of Translators. In *Meta*, 43(4), 616-630.

Zanettin, F. 2000. Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis. In Olohan M. (ed.) *Intercultural Faultlines*. Manchester : St Jerome Publishing.