Translation Tracking System: A tool for managing translation archives

Lynne Bowker* and Peter Bennison†

*School of Translation and Interpretation, University of Ottawa,
70 Laurier Avenue East, Room 401, Ottawa, Ontario, K1N 6N5, Canada
lbowker@uottawa.ca

†Sepro Telecom,
Dublin, Ireland
pbenn@hotmail.com

Abstract

The Translation Tracking System (TTS) is a database management tool intended to help translation researchers, translator trainers and translators to collect and organize archives of translated material. Relevant corpora can then be extracted from the archive in order to be further processed and analyzed using other natural language processing tools. This paper briefly describes the design and development of TTS, and it then goes on to explore how this tool has been successfully applied in an academic environment to help translator trainers identify areas of difficulty that have been encountered by their students. Some other applications of TTS are also discussed.

1. Introduction

Electronic corpora are playing an increasingly important role in the discipline of translation, where they are being used in a variety of ways, including 1) as resources for human translators to find solutions to specific translation- and terminology-related problems (e.g., Bowker and Pearson, 2002; Lindquist, 1999; Zanettin, 2001); 2) as resources for developing and testing machine translation systems – particularly systems using example-based and statistical models (e.g., Carl and Way, 2001); and 3) as a source of empirical data for more theoretical investigations into the nature of translation and translated text (e.g., Baker, 2001; Kenny, 2001; Laviosa, 1998). Regardless of the intended application, the collection of translation resources often begins in a haphazard and opportunistic way, with files being gathered in different formats and being indexed inconsistently. This was certainly the case for our own data collection endeavours in the not-so-distant past. Frustration with this process prompted us to work on designing and building a tool that could be used to help collate translation resources in a more systematic way. These efforts resulted in the production of the Translation Tracking System (TTS) – a tool that aims to provide a user-friendly framework for gathering and organizing translation data.

Essentially, TTS is a database application that permits users to enter (or copy) texts into a template, which also allows for the entry of relevant indexing information. The default indexing information consists of text attributes such as the source language, target language, subject field, text type, date, source, author/translator details, etc. Users can add or modify attributes as necessary. Once the data have been entered, the resulting archive can then be searched according to these indexed attributes, and texts that match the specified criteria can be extracted and exported for use with other natural language processing applications, such as part-of-speech taggers, term extraction systems, or corpus analysis tools. In short, TTS helps users to gather and manage a wide-ranging archive of translation data, but it also facilitates the export of specialized corpora for specific studies or investigations.

2. TTS design and development

As is the case for many software applications, the design and development of TTS has been an iterative process. TTS has been designed to be a straightforward and easy-to-use tool that can be employed by researchers, trainers, students and practitioners in the field of translation, as well as other users who do not necessarily have a computational background. TTS was initially implemented as a stand-alone application using Visual Basic Applications. It is currently being re-implemented using the PHP programming language so that it will be able to run over the Internet using a client-server model. In its current form, TTS has two main elements: the template, where textual and indexing information is entered, and the query form, where users formulate searches to identify and extract relevant texts from the archive. These are described in more detail in the following sections.

2.1. Template

Since the primary users of TTS are translators rather than computer scientists, the template includes a number of features intended to make it as user-friendly as possible. Firstly, the template allows users to enter as much indexing information as possible by selecting relevant criteria from drop down lists. This approach was adopted in order to minimize inconsistencies that may occur when data is entered in a free format (e.g., different spellings, typographical errors). Such inconsistencies would limit the effectiveness of queries and information retrieval. However, in cases where it may not be reasonable to create drop-down lists for some attributes, this information can be entered in a free format. Users can modify, delete or add attributes to the drop-down lists as necessary using a simple graphical user interface.

Secondly, the template provides validation, which means that users are prompted to fill in each of the attribute fields before closing the file. If a user attempts to close a file without filling in an attribute field, a message box will pop up advising the user which field(s) still need to be filled in. It is possible to override this feature in cases where no information is available for a given attribute.

When the template has been filled in, the data is saved in plain text format (.txt). This means that when texts are later exported from TTS, they will be in a format that can be easily processed by other software (e.g., taggers, corpus analysis tools). An additional benefit of using a plain text format is that this also reduces the chances of spreading viruses.

2.2. Query form

Once stored in TTS, the data can be searched using a simple query form. Using the drop-down lists of attributes, users can set the parameters of the search by selecting the relevant criteria. A query can be carried out using a single attribute or a combination of attributes. Once the user has defined the search parameters, TTS will then retrieve from the archive all the texts that conform to the specified criteria. For example, TTS can be asked to search all texts in the archive and retrieve only those that match the criteria:

- Source language: French
- Target language: English
- Publication date: 2001
- Subject field: medicine
- Text type: research article

All the texts that match these criteria are then presented to the user in a table. The user can view the full text for any item in the table by opening it from within TTS. The user can also choose to export some or all of the texts in the table to a separate file. Users can choose whether they wish to export just the texts, or whether they wish to export the texts along with a list of their corresponding indexing attributes. In addition, texts can be exported individually, or they can be concatenated into a single large file. The resulting corpus can then be processed using other software as desired by the user.

3. TTS applications

As outlined above, TTS is a database application that can be used to manage an archive of texts. Our initial motivation for developing such a tool stemmed from our desire to study student translations with a view to identifying areas of difficulty for students and learning more about the translation process. This Student Translation Archive is the application for which TTS has been used most extensively to date, and it is described in more detail in section 3.1 below. However, other applications, such as managing translation resources for use by machines, professional translators, or translatologists are also possible. Furthermore, with some relatively simple modifications, TTS can be adapted to manage data for other types of translation-related investigations. One such adaptation will be discussed in section 3.2.

3.1. Student Translation Archive

For a number of years, foreign language teachers have been compiling and studying “learner corpora”, which are defined as textual databases of the language produced by foreign language learners (e.g., Granger 1998). Such corpora are used to identify typical characteristics of texts produced by language learners and to identify errors and problem areas that can then be addressed as part of the language learning curriculum.

Student translators can be considered as a highly specialized type of language learner/user. Although their specific needs differ from those of general language learners, we felt that a similar approach to collecting and studying the output of student translators would be highly valuable for both pedagogical and research purposes. With regard to pedagogy, a corpus of student translations can provide a means of identifying areas of difficulty that could then be integrated into the curriculum and discussed in class. In terms of research, a number of scholars (e.g., Baker, 2001; Kenny, 2001; Laviosa, 1998) have already demonstrated that translation corpora can be useful for studying the nature of professionally translated text; we believe that there is also much to be learned about translation process and product by investigating the nature of text translated by students.

One application of TTS that has already been successfully undertaken is the use of the system to facilitate the development of an archive of student translations. Using an approach based on that developed by Granger (1998) for second language learners, translator trainers at the University of Ottawa in Canada are currently using TTS to collect, organize and study translations produced by their students. Students enter their translations into TTS, and the trainers can extract a selection of texts according to desired criteria. For example, some types of corpora that have been created and extracted with the help of TTS include 1) longitudinal corpora, 2) text-specific corpora, 3) subject-specific corpora, 4) cross-subject field corpora, and 5) cross-linguistic corpora. By analyzing such corpora, translator trainers can identify the types of problems that are being encountered by their students, and they can develop course curricula that will address these issues.

3.1.1. Longitudinal corpora

Longitudinal corpora can be used to track the progress of a specific student or group of students over a given period of time (e.g., a semester, a year, or even an entire degree). By extracting and studying a longitudinal corpus, a trainer or student can see which translation-related difficulties appear to have been resolved and which are still causing problems. For instance, near the beginning of a French-to-English technical translation course, a student was identified by the trainer as having a tendency to use constructions containing prepositional phrases (e.g., head of the scanner) in places where constructions containing pre-modifiers (e.g., scanner head) would be more natural. As a result, although the student’s translations were grammatically correct, they were not idiomatic because they had not been constructed according to the norms of the sublanguage in question. The student was advised of this problem and was encouraged to use pre-modifiers instead of prepositional phrases where appropriate. Over the course of the semester, longitudinal corpora of the student’s work were extracted using TTS. The corpora were then part-of-speech tagged using the AMALGAM tagger and analyzed with the help of WordSmith Tools, a corpus analysis package that includes a concordancer.

Some of the results of the analysis of the longitudinal corpus extracted at the end of the semester are shown in table 1. This table illustrates that the texts that were translated by the student at the beginning of the semester contained a higher proportion of prepositional phrases, whereas the texts translated towards the end of the semester showed an increased use of pre-modifiers. The trainer was able to examine particular instances of both prepositional phrases and pre-modifiers in context with the help of a concordancer, and the trainer was consequently able to determine that the student was indeed learning to use pre-modifiers in appropriate places in the translated texts. With the help of the longitudinal corpus, the trainer was able to provide the student with concrete feedback and empirical evidence demonstrating the progress of the student's learning.

Text	No. of prepositional phrases	No. of pre-modifiers
Text 1 (614 words) (Jan. 2001)	11	3
Text 2 (720 words) (Feb. 2001)	10	5
Text 3 (857 words) (Mar. 2001)	7	10
Text 4 (1008 words) (Apr. 2001)	6	14

Table 1: Summary of longitudinal corpus analysis.

3.1.2. Text-specific corpora

Translator trainers are often interested in seeing how a particular passage in a text has been handled by the various students in a class. Such investigations permit the trainer to identify areas where the class as a whole is having difficulty, as distinct from problems that may have befallen only one or two students. This allows the trainer to appropriately orient the curriculum or class discussions in order to focus on generally problematic issues. Many translation classes typically contain between twenty and thirty students, and these students often submit their work in printed form. In a class of this size, it is cumbersome to try to identify patterns of "problem areas" when working with separate sheets of paper. In contrast, if the translations are entered into a system such as TTS, a trainer can easily export all the given translations of a particular source text, and focus in on the different renderings of a selected passage with the help of a concordancer. Table 2 illustrates how a trainer can simultaneously view all the student translations of the following extract from the French source text: "*...un pare-feu pour les internautes à haut débit...*".

By examining these different translations simultaneously, the trainer found that patterns became more visible, which made it possible to discern, for example, that the majority of the class would benefit from a discussion regarding the hyphenation of compound pre-modifiers, since only three of the eleven students who used a compound pre-modifier (e.g., "high-speed") correctly hyphenated this type of construction. In contrast, the trainer could determine that there was only one student (see table 2, line 1) who encountered difficulty with a misplaced modifier, using "broadband" to incorrectly modify "firewall" instead of "Internet users". Similarly, only a single student (see table 2, line 5) misunderstood the underlying concept represented by the French term "*haut débit*", choosing to render it as "frequent", rather than "high-speed". In cases such as these, the students could be approached independently to discuss these problems.

ST	*...un pare-feu pour les internautes à haut débit...*
1.	...a broadband firewall for Internet users...
2.	...a fire-wall for high-speed Internet users...
3.	...a firewall for broadband connection users...
4.	...a firewall for broadband surfers...
5.	...a firewall for frequent Internet users...
6.	...a firewall for high speed Internet surfers...
7.	...a firewall for high speed Internet surfers...
8.	...a firewall for high speed Internet users...
9.	...a firewall for high speed Internet users...
10.	...a firewall for high speed Internet users...
11.	...a firewall for high speed Internet users...
12.	...a firewall for high speed Internet users...
13.	...a firewall for high speed surfers...
14.	...a firewall for high-speed Internet surfers...
15.	...firewalls for broadband Internet users...
16.	...firewalls for high-speed Internet users ...

Table 2: Extracts from a text-specific corpus.

3.1.3. Subject-specific corpora

Many individual translation courses are devoted to specialized subject fields, such as medical translation or legal translation. Trainers can extract a corpus of translations pertaining to a particular subject field and examine these to determine if a problem is specific to one particular source text or if it is a difficulty that is also manifesting itself in other texts dealing with a related theme.

For example, over the course of a semester, students in a French-to-English technical translation class were required to translate two different texts on the subject of animal cloning. The first text was taken from a semi-specialized science magazine and it explained the concepts associated with animal cloning techniques in a clear and accessible manner that the majority of students were able to understand and adequately convey in their translations. The second text, which the students translated several weeks after the first, examined cloning from the point of view of ethical considerations. It described some of the same concepts as the first text, but these were expressed in a more abstract and less straightforward way. When translating the second text, many of the students seemed less sure of themselves, did not appear to understand the concepts in question, and produced much more literal translations that closely followed the syntax of the source text. Consequently, the translations of the second text were less accurate and less articulate than the translations of the first text. The fact that the first translation had been fairly accurately rendered seems to indicate that the students were generally able to comprehend the subject matter; therefore, the difficulties with the second text were more likely to be related to the language and style that were used. This prompted the trainer to spend more class time discussing stylistic features of different text types and less time focusing on explanations of the subject matter itself (i.e., animal cloning techniques).

3.1.4. Cross-subject field corpora

As noted above, many translator-training programs require students to follow courses in different types of translation (e.g., legal translation, medical translation). Using TTS, it is possible for trainers to extract corpora that span multiple subject fields in order to investigate whether the problems encountered by a student or group of students in one type of specialized translation course (e.g., economic translation) are similar to or different from the problems they are having in another type of specialized translation course (e.g., technical translation). In this way, a trainer can try to determine whether a student is having a problem that manifests itself regardless of the subject field and therefore needs to be tackled at a more global level, or whether the student is having difficulty caused by a lack of knowledge of some aspect of a particular subject field (e.g., concepts, vocabulary, syntax) and which does not manifest itself when working in other fields.

For example, in examining translations produced by an individual student taking both an economic translation course and a technical translation course, it became clear to the trainers that the student had a general difficulty in grasping the concepts of register and text type. In the economic translation class, the student had translated a section from an annual report produced by an investment company, while in the technical translation class, the student had translated an extract from a technical report describing a new type of computer operating system. Both translations contained inappropriate constructions, such as the use of the second person, the use of contractions, and the use of sentence fragments. Faced with such evidence, the trainers decided to approach the student's difficulty with register and text type at a more global level, independently of any particular subject field.

3.1.5. Cross-linguistic corpora

In a similar vein, students in translator training programs often work with multiple languages or in different language directions. In such cases, cross-linguistic corpora can be extracted from TTS and used to help trainers determine whether a given student's work is subject to source language interference or whether the student encounters different types of problems when working with different languages. For instance, a trainer may decide to investigate potential problems of source language interference by extracting texts translated by a particular student that cover the same subject field but were translated from different languages (e.g., Spanish-to-English and French-to-English). A comparison of the two sets of translations may reveal that the student is having different types of problems when translating out of Spanish than when translating out of French, in which case it may be necessary to focus on source language interference, or it may turn out that a student is having similar problems regardless of the source language (e.g., perhaps the student has not grasped the concept of register), in which case it may be necessary to tackle the issue in a way that is not related to a particular source language.

For example, a student translated two technical texts on the subject of optical scanners: one from Spanish to English and one from French to English. In both cases the student's terminological choices were heavily influenced by the source text in question. For instance, the Spanish term "*scanner plano*" was translated by "flat scanner" while the French term "*numériseur à plat*" was rendered as "flat digitizer"; however, both of these terms are referring to the same concept, and in both cases, a better solution would have been a translation such as "flatbed scanner". This type of concrete evidence can be used to help the student see and correct the problem.

3.2. Other applications of TTS

Although the Student Translation Archive is the application for which TTS has been used most extensively to date, another application has been the use of TTS to manage translation resources for freelance translators. Freelance translators often work in a variety of subject fields and with a range of text types. Consequently, they are always on the lookout for any text that might be a useful resource for future translations. However, once they have built up a diverse collection of resources, they must be able to quickly pinpoint those specific elements that will be relevant for translating a given source text. TTS has proved to be extremely useful in this regard because it allows translators both to amass and organize a large archive of resource material and to quickly identify and extract only those texts that are pertinent to particular project.

TTS can also be adapted to assist with other types of data management. For example, at the University of Ottawa, a research project is currently underway that aims to investigate the evolution and current state of the translation profession in Canada by examining a corpus of translation-related job advertisements. As part of this project, researchers are collecting an archive of job advertisements and storing them using a modified version of TTS. The modifications relate primarily to the indexing attributes, which in this case include attributes such as job title, employer, location, qualifications, required skills, etc. Using TTS, researchers can extract from the archive only those advertisements for jobs requiring a BA in Translation, or only those jobs based in Ottawa, etc. The data can then be further analyzed using corpus analysis tools.

4. Concluding remarks

As noted in the introductory section of this paper, more and more people working in various areas of the translation profession are becoming aware of the value of consulting corpus-based resources. However, as these resources grow – both in size and in diversity – it is increasingly important that they be managed in a systematic fashion. The use of a tool such as a word processor to create, index, organize, and search an archive is simply not an efficient approach to managing a data collection. Organization is the key to success when constructing and consulting data resources. If such resources are not well constructed, pertinent information may be overlooked, or searches may be carried out on inappropriate material, which could produce misleading results. TTS aims to provide a user-friendly means for helping members of translation-related professions to gather and maintain their data collections in an organized manner. The flexibility of TTS has allowed it to be used to support various types of investigations relating to translation teaching, research, and practice.

5. Acknowledgements

The work described here has been partially funded by grants awarded to Lynne Bowker by the Faculty of Arts of the University of Ottawa and the University of Ottawa Research Fund. Peter Bennison would like to thank his employer, Sepro Telecom, for supporting his research.

6. References

AMALGAM tagger: http://www.comp.leeds.ac.uk/amalgam/amalgam/amalgsoft.html

Baker, Mona, 2001. Investigating the Language of Translation: A Corpus-based Approach. In P. Fernández and J.M. Bravo (eds.), *Pathways of Translation Studies*. Valladolid: Universidad de Valladolid.

Bowker, Lynne and Jennifer Pearson, 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Carl, Michael and Andy Way (eds.), 2001. *Proceedings of the MT Summit VIII Workshop on Example-Based Machine Translation*. Geneva: EAMT.

Granger, Sylviane (ed.), 1998. *Learner English on Computer*. London: Longman.

Kenny, Dorothy, 2001. *Lexis and Creativity in Translation: A Corpus-based Study*. Manchester: St. Jerome Publishing.

Laviosa, Sara, 1998. The English Comparable Corpus: A Resource and a Methodology. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (eds.), *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome Publishing.

Lindquist, Hans, 1999. In G. Anderman and M. Rogers (eds.), *Words, Text, Translation: Liber Amicorum for Peter Newmark*. Clevedon: Multilingual Matters.

WordSmith Tools: http://www1.oup.co.uk/elt/catalogue/Multimedia/WordSmithTools3.0/download.html

Zanettin, Federico, 2001. Swimming in Words: Corpora, Translation and Language Learning. In G. Aston (ed.), *Learning with Corpora*. Bologna/Houston: CLUEB/Athelstan.