BENJAMINS

TRANSLATION LIBRARY

Topics in Language Resources for Translation and Localisation

edited by
Elia Yuste Rodrigo

Topics in Language Resources for Translation and Localisation

Benjamins Translation Library (BTL)

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

EST Subseries

The European Society for Translation Studies (EST) Subseries is a publication channel within the Library to optimize EST's function as a forum for the translation and interpreting research community. It promotes new trends in research, gives more visibility to young scholars' work, publicizes new research methods, makes available documents from EST, and reissues classical works in translation studies which do not exist in English or which are now out of print.

Volume 79

Topics in Language Resources for Translation and Localisation
Edited by Elia Yuste Rodrigo

Topics in Language Resources for Translation and Localisation

Edited by

Elia Yuste Rodrigo
University of Zurich

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Topics in language resources for translation and localisation / edited by Elia Yuste Rodrigo.
p. cm. (Benjamins Translation Library, ISSN 0929-7316 ; v. 79)
Includes bibliographical references and index.
1. Translating and interpreting--Data processing. 2. Corpora (Linguistics) I. Yuste Rodrigo, Elia.
P308.T67 2008
418'.02--dc22 2008035296
ISBN 978 90 272 1688 5 (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Introduction

ELRA, the European Language Resources Association, defines Language Resources (LRs) as sets of language data and descriptions in machine readable form, which are specifically used to create, optimise or evaluate natural language processing and speech algorithms and systems, and generally, as core resources in the language services industries and localisation, for language studies, subject-area research, etc. Examples of LRs include but are not limited to written and spoken language corpora, terminology databases, computational lexica and dictionaries, software tools, etc. developed for different types of Human Language Technologies (HLT) applications, with their varied end-users in mind.

When Translation is understood as *process* rather than as *product* only, LRs play an indispensable role. Language resources such as the ones mentioned above may be of outstanding usefulness in the process of creating, standardising, leveraging, adapting... content for more than one language and culture. However, it has not been until recent years[1] that Language Resources for Translation (LR4Trans, for short) have been given the necessary attention. Since this has been the case mainly in academic and research circles, some efforts ought yet to be made to raise further awareness about LRs in general, and LRs for translation and localisation, in particular, to a wider audience in all corners of the world. Hence, the motivation number one behind this book.[2] The volume focuses on language resources from

1. Elia Yuste Rodrigo brought scholars and industry players from the areas of translation and corpus and computational linguistics together in a workshop held on 28th May 2002 in conjunction with LREC 2002 (Third International Language Resources and Evaluation Conference, Las Palmas de Gran Canaria, Spain). The workshop goal was to explore new avenues relating to language resources and technology-enhanced multilingual work and research. The term 'language resources for translation' was first formally used here. Given the attendants' interest, she would organise and chair two other workshops (second edition celebrated on 28th August 2004 as a satellite event of the 20th International Conference on Computational Linguistics – COLING at the University of Geneva, Switzerland; third edition, LR4Trans-III, held on 28th August 2006 under the auspices of LREC 2006 in Genoa, Italy).

2. Even if *Topics in Language Resources for Translation and Localisation* is the logical inheritance of the workshop series initiated by the editor, this is not a conference proceedings book and its novelty has to be emphasised. The selected contributions capture the current state-of-art in terms of research, work practices and industry standards. Much attention has been given to

several angles and in relation to current trends of multilingual content processes, being appealing to the heterogeneous readership of the Benjamins Translation Library (BTL)[3] series worldwide.

Students, educators, researchers and professionals related to the translation and localisation arena will remember that in a not so distant past there seemed to be two extremes, represented at one end by those exclusively preoccupied with the then new market tools (essentially, commercial translation memories products), and at the other end by the ones that felt somehow threatened by an increasing degree of translation automation and kept defending human translation as the only possible alternative. What LR4Trans does is to underline the interaction of all the electronic language resources, applications and technologies that may be used in (learning about) the process of translating or localising – without forgetting about all the human agents that may also be involved therein, from technical writers, domain specialists, and corporate linguists of various kinds to computational linguists and future translators, to name a few. In other words, the approach behind LR4Trans is integrative and includes data, tools and human agents, allowing for targeted yet varied discussion points. This is one of the main features of the book you are now holding in your hands, its array of innovative topics for the language professional.

A truly practical and applied linguistic book in nature that is highly connected with multilingual technologies as used in translation and localisation processes must be written in as current a fashion as possible. Nevertheless, the principles behind this volume will not outdate so rapidly; only those aspects intrinsically dependant on a technological update or a new industry standard would require a content revision in the future. It is the potential of the flexible concept of language resources for translation and localisation what the readership has to adapt to their own working scenario and set of needs. The authors and the editor remain at the reader's disposal to clarify or expand on any of the issues presented here.

What follows are the rationale behind and a summary of the twelve selected contributions that make up this book. After going through the first half of the volume, the reader might get the impression that it focuses primarily on corpora, in one way or another. Yet there are two important things to note here: First, this is a reflection of a current yet maturing research trend and, secondly, this should be a good incentive to get to know more about different types of and aspects surrounding corpora (e.g. parallel and comparable corpora, treebanks, exploitation tools, interface and other design points, potential for teaching future translators in less

expand and update the information presented in any of the workshops. Some contributions have in fact been written from scratch to better serve the needs of the here targeted wider readership.

3. http://www.benjamins.com/cgi-bin/t_seriesview.cgi?series=BTL

resourced languages, etc.). As the reading progresses, the audience will find other exciting paths along the way to satisfy their curiosity about, for example, the perception of languages resources for translation in international organisations, how language resources could coadyuvate in addressing the needs of volunteer translators and the role of language resources in localisation (both to tackle specific challenges and from the latest industry and research joint efforts).

Chapter 1, *A Comparative Evaluation of Bilingual Concordancers and Translation Memory Systems*, is a thorough analytical study by **Bowker & Barlow** of two complementary (not competing) technologies that inform the language professional about the potential advantages that each offers so that they can choose the better tool for them. Opening the book with this chapter is intentional. The less acquainted with corpora readership will benefit from a first-hand introduction to what concordancing tools and parallel corpora can do for them.

In Chapter 2, however, the reader will learn that Translation problems derived from the specificity of a language or a register require a detailed linguistic analysis, which cannot always be accomplished with the use of parallel corpora. In *Interactive Reference Grammars: Exploiting Parallel and Comparable Treebanks for Translation*, **Hansen-Schirra** points out that the grammatical slant has not yet been addressed in corpus-based translation work. She argues that monolingual and multilingual treebanks may assume the role of grammatical reference resources for professional translators and translators-to-be. A translation corpus should be annotated with more abstract kinds of linguistic information, such as semantic and discourse information.

Chapter 3 by **Bernardini** and **Castagnoli**, *Corpora for translator education and translation practice*, aims to promote an educational rather than a training perspective of corpora for student translators. These should be educated to explore the role of the corpus. E-learning materials should foster this raising-awareness factor, ideally being contrastive in focus, i.e., corpora against other resources, such as dictionaries or translation memories (TMs). Other important aspects relate to corpus construction and corpus searching, which should be made faster and more user-friendly, and ideally integrated with CAT tools. Concerning the exploitation technologies discussed, we see a complementary or integrative standpoint once more, rather than an exclusive or imposing one.

Also willing to represent an educational rather than a training voice for translators, **Maia**'s contribution and the book's Chapter 4, *Corpógrafo, V.4 – tools for educating translators*, describes the history, motivation and latest developments of a flexible tool suite, the Corpógrafo, which integrates and responds to principles of corpus linguistics, extraction and management of terminology and knowledge engineering. An online, freely available suite, this research environment for autonomous study also offers various possibilities for education in translation. It

should also be an incentive for self-discovery and improvement in other translation operational settings.

In Chapter 5, *Corpus exploitation for translation teaching and research: state of the art*, **Colominas & Badia** examine the weaknesses and strengths of current corpora interfaces and exemplifly search types that can be relevant in translation training contexts or for translation research purposes, with a view to identify the basic requirements a corpora interface should satisfy. The lack of sufficiently large corpora representative of modern languages is currently being solved by means of web corpora, their analysis providing evidence of work done regarding this matter but also of the need for further work.

Sometimes not only how to better access corpora is at stake but also how to create translation corpora for less resourced languages. In Chapter 6, *The Use of Corpora in Translator Training in the African Language Classroom: A Perspective from South Africa*, **Gauton** draws on her expertise with electronic text corpora, corpus query tools and translation memory tools to enable the African language translator to mine target language texts for possible translation equivalents, coin terms in the absence of standardised practices, and re-use existing translation to attain terminology standardisation. Future action lines include further standardisation work and the transformation of the multilingual student output site into a large and comprehensive language resource available to external parties.

In a radically different work setting but equally aware of local constraints and requirements, **de Saint Robert** pinpoints in Chapter 7, *CAT tools in international organisations: lessons learnt from the experience of the Languages Service of the United Nations Office at Geneva*, that the usefulness of such tools does not have to be taken for granted. Here translation is seen as a highly interrelated activity which has to go hand in hand with and become closer to other internal business processes. Much attention has to be given to less sophisticated tools, but suitable for the organisation's *modus operandi*, as well as to the way internal language resources are built and integrated in the workflow.

Shedding some light on *Global Content Management – Challenges and Opportunities for Creating and Using Digital Translation Resources* (Chapter 8), **Budin** discusses the convergence of content management and cross-cultural communication. After exploring the concept of content, he goes on to explain that specialised translation is currently taking place within the wider, integrative paradigm of global content management. Translation resources (e.g., translation memories and other aligned corpora, multilingual terminological resources, reference resources, etc.) are typical examples of content that needs to be managed in global action spaces.

Bey, Boutet and Kageura then present *BEYTrans: A Wiki-based Environment for Helping Online Volunteer Translators* (Chapter 9). Following major Web 2.0 advances, this research project reflects new collaborative work patterns among

volunteer translators that could open new avenues for all communities involved in translation. Leveraging the Wiki technology, BEYTrans aims at empowering online translators through system components for producing a quick, yet high-quality translation in several languages. A range of system functionalities allow them to manage the language resources themselves online, in the *wiky* fashion.

Chapter 10 is also the result of innovative research. **Cruz Lara et al.** are concerned with *Standardising the management and the representation of multilingual data: the Multilingual Information Framework*. The MLIF framework, based on a methodology of standardisation resulting from the ISO (International Standards Organisation), is being designed with a high-level common conceptual model of multilingual content in mind and as a platform allowing interoperability among several translation and localisation standards and their related tools. This interoperability is the main benefit of MLIF, which also facilitates the evaluation and comparison of multilingual resources and tools.

Kato & Arisawa's Chapter 11, *Tagging and Tracing Program Integrated Information*, introduces the reader to a software internationalisation challenge by focusing on the translation of Program Integrated Information (PII). PII is normally separated from the computer programs themselves and brought into text resource files that are translated outside the program development lab. How can this decontextualisation be compensated during the translation verification test (TVT)?

In Chapter 12, *Linguistic Resources and Localisation*, **Schäler** provides the reader with the essential definitions surrounding localisation (L10N). The discussion is followed by a real-life case study showing the use of language resources in localisation that laid the foundations for the "translation factory". The commonalities found in the L10N process, in terms of frequent updates, repetitive material, etc. may facilitate standard approaches to L10N problems. Yet the mark-up and formatting of source material and the complexity of L10N processes hinder localisation automation efforts. These have to be tackled when developing innovative L10N frameworks. We encourage every reader to keep on reading till the very end of the book and find out what IGNITE is all about.

Acknowledgements

I am exceptionally lucky to have drawn on the outstanding expertise of all the contributors who have enthusiastically believed in this exciting book. My special thanks must also go to Prof. Sue Ellen Wright (Kent State University), who did see a need for such a volume and encouraged me to submit the book proposal.

I would also like to thank all the professors and fellow lecturers and researchers from the University of Valencia, UMIST (now University of Manchester), Staffordshire University and the University of Zurich, who have forged my career as a

professional linguist and researcher. I also feel indebted to the fellow translators/localisers, terminologists, project managers and language service providers, with whom I work on a regular basis, for having made me capable of keeping abreast of what the real world does need from university scholars (and the way round) and helping me keep a *sane* balance between academia and industry in the last fourteen years.

Last but not least, I am especially thankful to the publishers' team at John Benjamins Publishing Company, in particular to Prof. Yves Gambier (University of Turku), for his valuable time and attentive responses and to Isja Conen, for her help from the very beginning.

Elia Yuste Rodrigo
(Zurich / Dublin, Winter-Spring 2008)

N.B. All the online pointers and links mentioned in this book were last consulted in January 2008.

CHAPTER 1

A comparative evaluation of bilingual concordancers and translation memory systems

Lynne Bowker and Michael Barlow
University of Ottawa / University of Auckland

Translators are turning to electronic language resources and tools to help cope with the increased demand for fast, high-quality translation. While translation memory tools are well known in the translation industry at large, bilingual concordancers appear to be familiar primarily within academic circles. In this chapter, the strengths and limitations of these two types of tool are analysed with respect to automation, search flexibility, consistency and other quality-related issues in an effort to identify those circumstances in which each could best be applied.

1. Introduction

Recent years have witnessed a number of significant changes in the translation market. For instance, largely as a result of globalisation, there has been a considerable increase in the volume of text that needs to be translated. In addition, new types of text, such as web pages, have appeared and require translation. The increased demand for translation has been accompanied by another trend: deadlines for completing translation jobs have grown shorter. This is in part because companies want to get their products onto the shelves in all corners of the world as quickly as possible. In addition, electronic documents such as web pages may have content that needs to be updated frequently. Companies want to be sure that their sites reflect the latest information, so translators are under pressure to work very quickly to ensure that the up-to-date information is displayed in all language versions of the site. This situation has been further exacerbated by the fact that in today's market, there is currently a shortage of qualified human translators (e.g., Sprung 2000: ix; Shadbolt 2002: 30–31; Allen 2003: 300).

The increase in volume coupled with shorter turnaround times and fewer workers has resulted in an immense pressure on existing translators to work more quickly, while still maintaining high quality in their work. However, these two

demands of high quality and fast turnaround are likely to be at odds with one another. Therefore, one way that some translators are trying to balance the need for high quality with the need for increased productivity is by turning to electronic resources and tools for assistance.

One type of language resource that has become very popular is the bilingual parallel corpus, which is essentially a collection of texts in one language (e.g., English) alongside their translations into another language (e.g., French). The two sets of texts must be aligned, which means that links are made between corresponding sections – often between sentences or paragraphs – in the two languages.

Bilingual parallel corpora can contain a wealth of valuable information for translators, but in order to be able to usefully exploit these resources, some kind of tool is needed. There are two main types of tools that can be used to search for and retrieve information from a bilingual parallel corpus:[1] a bilingual concordancer (BC) and a translation memory (TM) system. While these two tool types have some common goals and features, they also have a number of differences.

As we will see in the upcoming sections, BCs are sometimes considered to be "old technology" and they are not well known in the translation industry outside of academic circles. In contrast, TMs have garnered a significant amount of attention in the translation industry of late; they are very much in vogue and are considered to be leading-edge technology. Nevertheless, a number of translators have expressed frustration and disappointment when trying to apply TMs in certain contexts. It is possible that some of the frustration experienced by translators using TMs in certain situations could be alleviated by using BCs instead. While there have been numerous comparative reviews of different products, these tend to focus on comparing different products from the same category. For example, both Zerfaß (2002) and Waßmer (2005) compare a number of different TM systems, while Reppen (2001) compares two concordancers. However, to the best of our knowledge, there have not been any detailed investigations that compare BCs to TMs. The aim of this chapter is to conduct a comparative analysis of these two types of technology in an effort to determine the strengths and weaknesses of each in order to establish those situations where translators would be best served by using a TM and those where they may be better off using a BC.

Following the introduction, the chapter will be divided into four main sections. Section 2 provides some background information, including a general description of how the two types of tool work, with reference to two specific tools –

1. Note that while the same corpus data can be used with both types of tool, it is usually necessary to pre-process the corpus in a different way in order to render it readable by different tools since they may use proprietary formats.

ParaConc[2] and *SDL Trados*[3] – that are representative of the categories of BC and TM respectively. Section 3 contains a brief assessment of the place occupied by these tools within the translation industry today. Section 4 contains a more detailed comparative analysis of the features and associated advantages and disadvantages of each type of tool. Finally, Section 5 concludes with some general recommendations about which translation situations warrant the use of each type of tool.

2. General introduction to BCs and TMs

The general aim of both a BC and a TM is to allow a translator to consult, and if appropriate to "recycle" or "reuse", relevant sections of previously translated texts. In the following sections, BCs and TMs will be described with reference to *ParaConc* and *SDL Trados*, which are representative examples of these respective categories of tool.[4]

2.1 ParaConc: An example of a BC[5]

BCs, such as *ParaConc*, are fairly straightforward tools: they allow translators to search through bilingual parallel corpora to find information that might help them to complete a new translation. For example, if a translator encounters a word or expression that he does not know how to translate, he can look in the bilingual parallel corpus to see if this expression has been used before, and if so, how it was dealt with in translation.

To use *ParaConc*, the source and target texts must first be aligned, which means that corresponding text segments are linked together.[6] A semi-automatic

2. For more information on ParaConc, see http://www.athel.com/para.html

3. For more information on SDL Trados, see http://www.trados.com

4. Other examples of BCs include TransSearch, Beetext Find and MultiConcord, while other examples of TMs include Déjà Vu, STAR Transit and WordFast.

5. In fact, ParaConc could more properly be termed a multilingual concordancer, since it is possible to consult texts in up to four languages at once. However, in the context of this paper, we will refer to it as a BC and discuss its use for comparing texts in two languages.

6. A detailed description of alignment techniques is beyond the scope of this paper; however, alignment is a non-trivial matter. Problems can arise, for example, if a single source text sentence has been translated by multiple target language sentences, or vice versa, or if information has been omitted from or added to the target text (e.g., to handle cultural references). For a more detailed description, see Bowker (2002a).

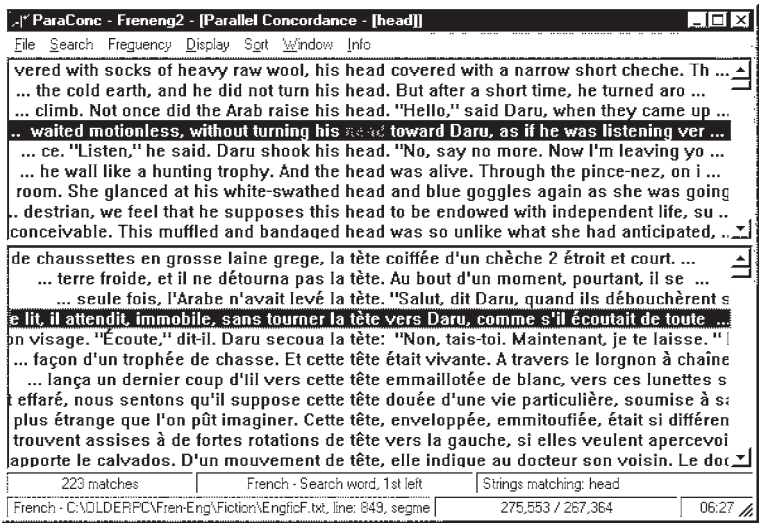

Figure 1. A ParaConc results window.

alignment utility is included in the program to prepare texts that are not already pre-aligned. The initial part of the alignment process is carried out in three stages: first the texts are aligned based on sections, if any are present in the texts, then alignment is carried out at the paragraph level, and finally at the sentence level. The software uses the formatting information in files to carry out alignment of sections and paragraphs. Alignment at the sentence level is achieved by applying the Gale-Church algorithm (Gale & Church 1993). To make adjustments to the alignment, the user can examine the aligned segments and either merge or split particular segments, as necessary. One very important thing to note is that the aligned units remain situated within the larger surrounding text.

Once the texts are aligned, the translator can consult the corpus. By choosing the basic search command, the translator can retrieve all examples of a word or phrase (or part of a word) from the corpus. As shown in Fig. 1, the search term *'head'* has been entered and all instances of *'head'* from the English corpus are displayed in the upper pane (here in a "key word in context" or KWIC format). The corresponding text segments from the French corpus are shown in the lower pane.

The concordance lines can be sorted in various ways (e.g., primarily 1st left and secondarily 1st right) in order to group similar phrases together and therefore make it easier for a translator to spot linguistic patterns. Clicking on a concordance line in the upper pane will highlight that line and also the corresponding text segment in the lower pane. Double-clicking on a line will bring up a window containing the segment within a larger context.

Suggested translations for the English *'head'* can be highlighted by positioning the cursor in the lower French results pane and clicking on the right mouse button. A possible translation of *'head'* such as *'tête'* can be entered. The program then simply highlights all instances of *'tête'* in the French results window, which can then be displayed (and sorted).

It is also possible to use a utility that presents a list of "hot words" in the French results pane, including possible translations. Some or all of the words listed can be selected and they will then be highlighted in the results.

Finally, more complex search commands can also be used if desired. Some of the possible advanced search options are: Text search, Regular expression search, Tag (part-of-speech) search, Batch search, and various heading-sensitive and context-sensitive searches. Of particular interest to translators is a Parallel search, which allows the user to enter both an English and a French search word and to retrieve only those occurrences that match both (e.g., only instances where *'head'* is translated by *'tête'* and not by *'chef'*). This type of Parallel search strategy will allow a translator to reduce the amount of "noise" or irrelevant data that is retrieved.

2.1.1 *Potential limitations of BCs*

There are a number of potential limitations that are often associated with BCs. The three principal ones are (1) the limited degree of automation; (2) the nature of the search item; and (3) the nature of the matching process.

With regard to the degree of automation, when using a BC, it is up to the translator to decide what word or expression to look up, and he then has to manually type this search pattern into the BC's search engine.

In terms of the nature of the search item, BCs are generally designed to search only for words or very short phrases. It is true that, in principle, a BC could be used to search for an entire sentence or paragraph; however, the fact that the search pattern must be manually entered tends to discourage this type of use because it would be extremely time-consuming and error prone (e.g., typos).

Finally, BCs are sometimes criticised because of the nature of the matching process that they use. By default, these tools basically search through the corpus for occurrences that match the entered search pattern *precisely*. For example, if the translator enters the search pattern *'flatbed colour scanner'* into the BC, it will retrieve only those occurrences that match that pattern exactly. It will not retrieve an example that contains differences in punctuation, spelling, morphology or word order (e.g., *'color flat-bed scanners'*). However, as noted in Section 2.1, some BCs, such as *ParaConc*, have added more advanced search features to improve the flexibility of searching.

2.2 SDL Trados: An example of a TM

Like a BC, a TM is a tool designed to help translators identify and retrieve information from a bilingual parallel corpus. However, one of the motivating factors in developing TMs was to overcome some of the seeming limitations of BCs as described in Section 2.1.1. Consequently, TMs are more automated, can search for longer segments, and employ fuzzy matching techniques.

The data contained in a conventional TM, such as *SDL Trados*,[7] are organised in a very precise way, which differs somewhat from the way in which data are stored for use with a BC. *SDL Trados* divides each text into small units known as segments, which usually correspond to sentences or sentence-like units (e.g., titles, headings, list items, table cells). The source text segments are linked to their corresponding target text segments and the resulting aligned pair of segments is known as a translation unit (TU). Each TU is extracted from the larger text and stored individually in a database. It is this database of TUs, not the original complete text, which is later searched for matches. This is an important distinguishing feature between a TM and BC.

When a TM, such as *SDL Trados*, is first acquired, its database is empty. It is up to the translator to stock the database. This can be done interactively by having the translator add each newly translated segment to the database as he works his way through the text, or it can be done by taking previously translated texts and aligning them using the accompanying automatic alignment program. It is important to note, however, that in order to ensure that the automatic alignment has been done correctly, manual verification may be required. Because the TM database stores only individual TUs rather than the complete source and target text, it is imperative that the TUs be correctly aligned if they are to be of any value to the translator.

When a translator receives a new text to translate he begins by opening this new text in the *SDL Trados* environment. *SDL Trados* proceeds to divide this new text into segments. Once this has been accomplished, the tool starts at the beginning of the new source text and automatically compares each segment to the contents of the TM database. If it finds a segment that it "remembers" (i.e., a segment that matches one that has been previously translated and stored in the TM database), it retrieves the corresponding TU from the database and shows it to the translator, who can refer to this previous translation and adopt or modify it for use in the new translation.

Of course, language is flexible, which means that the same idea can be expressed in a number of different ways (e.g., '*The filename is invalid*' / '*This file does*

7. Note that SDL Trados is actually a suite of tools that includes, among other things, an automatic aligner, a terminology manager, a term extraction system and a TM.

Table 1. Fuzzy match retrieved from the TM

Segment from new source text	The specified operation was interrupted by the system.
Fuzzy match retrieved from translation memory	EN: The specified operation was interrupted by the system. FR: L'opération a été interrompue par l'application.

not have a valid name'). Consequently, a translator cannot reasonably expect to find many exact matches for complete segments in the TM. However, it is highly likely that there will be segments in a new source text that are similar to, but not exactly the same as, segments that are stored in the TM. For this reason, *SDL Trados* also employs a powerful feature known as fuzzy matching. As shown in Table 1, a fuzzy match is able to locate segments in the TM that are an approximate or partial match for the segment in the new source text.

If more than one potential match is found for any given segment, these are ranked by the system according to the degree of similarity between the new segment to be translated and the previously translated segment found in the database. Note that the similarity in question is a superficial similarity (e.g., the number/length of character strings that the two segments have in common) and not a semantic similarity (thus '*gone*' and '*went*' will not count as similar despite the similarity in meaning of the two words). The match that the system perceives as being most similar to the new source segment is automatically pasted into the new target text. The translator can accept this proposal as is, edit it as necessary, or reject it and ask to see other candidates (if any were found).

SDL Trados also works in conjunction with terminology databases (or termbases, for short); however, it is important to note that these need to be prestocked by translators with specialised terms and their equivalents. By searching in the termbase – if one exists – *SDL Trados* can locate matches at the term level and present them to the translator. Nevertheless, there is still a level of linguistic repetition that falls between full sentences and specialised terms – repetition at the level of expression or phrase. This is in fact the level where linguistic repetition will occur most often.

Until recently, TMs permitted phrase or expression searching only though a feature that resembled a BC. In other words, a translator could manually select an expression, and the TM system would search through the database of TUs to find examples. In a recent version of *SDL Trados*, however, an auto-concordance function has been added, which, when activated will automatically go on to search for text fragments when no segment-level match is found. This has become a much-praised feature and other TM products, such as *Transit*, have begun to incorporate similar features (Hallett 2006: 9).

Once the translator is satisfied with the translation for a given segment – which can be taken directly from *SDL Trados*, adapted from an *SDL Trados* match, or created by the translator from scratch – the newly created TU can be added to the TM database and the translator can move on to the next segment. In this way, the database grows as the translator works. *SDL Trados* can also be networked so that multiple translators can search in and contribute to the same TM.

3. BCs and TMs in the translation industry

As summarised in Bowker (2004), a literature survey indicates that both BCs and TMs are widely used in academic settings for translator training. A long list of researchers (e.g., Zanettin 1998; Pearson 2000; Bernardini 2002a; Hansen & Teich 2002; Palumbo 2002; Tagnin 2002) have shown that using BCs in conjunction with parallel bilingual corpora can help students with a range of translation-related tasks, such as identifying more appropriate target language equivalents and collocations; coming to grips with difficult points of grammar (e.g., prepositions, verb tenses, negative prefixes); identifying the norms, stylistic preferences and discourse structures associated with different text types; and uncovering important conceptual information.

With regard to TMs, meanwhile, many translator trainers (e.g., DeCesaris 1996; Kenny 1999; L'Homme 1999; Austermühl 2001; Bowker 2002; Arrouart 2003) are now using TMs for tasks such as getting students to analyse and evaluate different translation solutions; helping students to learn more about inter- and intra-textual features by examining source texts and evaluating their characteristics in an effort to determine whether or not they can be usefully translated with the help of a TM; and conducting longitudinal studies of students' progress over the course of their training program.

In contrast to the academic setting, where both BCs and TMs are well known and widely used, the situation in the professional setting is somewhat different: TMs are very popular, but the existence of BCs is much less widely known. For example, TMs are discussed frequently in the professional literature. According to newsletters and programmes circulated to members, translators' associations such as the *American Translator's Association* or the *Association of Translators and Interpreters of Ontario* have provided their members with numerous opportunities – in the form of demonstrations, workshops, and professional development seminars – to learn about TMs. In addition, magazines aimed at language professionals, such as *Circuit*, *Translating Today Magazine*, *Localisation Focus* and *MultiLingual*, frequently include discussions on TMs (e.g., Lanctôt 2001; Biau Gil 2004; Musale 2004; Waßmer 2005; Hallett 2006). In those same publications, however, consid-

erably less attention has been paid to BCs. This raises the question as to why BCs appear to have received a less enthusiastic welcome in the professional world than have TMs. One factor that may have led to a difference in uptake of these two tools is the ease of access to such tools.

Firstly, it should be noted that BCs have long been known in fields such as language teaching or second-language learning (e.g., Johns 1986; Mindt 1986; Barlow 2000), but it is only more recently that their potential as translation aids has been recognised. Academics working in the field of translation are often involved in, or have colleagues who are involved in, language teaching, and as such they may have gained exposure to BCs in this way. Many of the existing BCs were initially developed by academics who work in language training[8] often as a means of helping their own students. This means that while such tools are generally very reasonably priced and may be easily accessible within the academic community, they are sometimes not widely advertised or distributed to the professional translation community because the people who have created these tools have full-time teaching jobs. In contrast, tools such as TMs, which have typically been developed in the private sector by companies that have professional full-time programmers, technical support staff and generous advertising budgets, are more actively marketed to working translation professionals. The fact that BCs do not seem to be well advertised in the professional setting may explain, in part, why language professionals and their associations seem to be more aware of the existence of TMs than they are of BCs. This situation may change in the future, however. As noted above, the use of BCs in translator training institutes has become firmly established since the late 1990s. This means that, at present, most of the translators in the workforce will have received their education during a time when BCs were not part of the translator training curriculum. However, over the coming years, the number of BC-savvy graduates will increase and they will bring to the workforce their knowledge of BCs. They will be able to share their experience with their colleagues and employers and gradually, more and more companies will have translators on staff with an understanding of such tools.

8. For example, *ParaConc* was developed by Dr. Michael Barlow, who works in the Department of Applied Language Studies and Linguistics at the University of Auckland; *MultiConcord* was developed by a consortium based in the Centre for English Language Studies at the University of Birmingham; and *TransSearch* was developed at a lab at the Université de Montréal.

4. Comparative analysis of BCs and TMs

On the surface, it may seem to be an obvious choice for a translator to select a TM over a BC since a TM includes the basic functions of a BC, as well as a number of additional features (e.g., automated searching, segment-level matching, fuzzy matching). However, if one looks beneath the surface, it seems that while TMs may be favourable in some circumstances, there are other situations where a BC may be the preferred tool. In the following sections, we will examine the strengths and weaknesses of BCs and TMs, using *ParaConc* and *SDL Trados* as representative examples of these respective categories of tools.

4.1 Automation

Automation is an oft-touted advantage of TMs. In principle, automating the search feature should speed up the process; however, this may not always be the case. As pointed out by Bédard (1995: 28), it is possible to approach automation in one of two ways: (1) an ambitious or high-tech approach, using very sophisticated and highly automated tools, such as TMs, or (2) a more modest or low-tech approach, where the tools (e.g., BCs) are simpler and require more user input.

In the case of the highly-automated approach, there can be hidden costs. Because the tools are more sophisticated, they may require a greater investment of time and effort in learning how to use them, which may prompt users to ask "What have I got myself into?". The pre-processing steps (e.g., alignment) may also be more demanding because an automated system depends more heavily on correct alignment. As noted in Section 2.2, in the case of *SDL Trados*, if a translator wishes to ensure that the alignment is absolutely correct in order to prevent misaligned TUs being presented, he must manually verify, and if necessary correct, the alignment – a process that can be extremely labour-intensive if the database is large. In contrast, since the data generated by BCs is designed for consultation by a human user, not a computer, the alignment requirements are somewhat less stringent. A certain number of alignment errors can be tolerated in a BC because the danger of "automatically" retrieving misaligned segments does not exist, and if an error does occur, the translator can simply look to the preceding or following text to find the corresponding segment because a BC does not extract the segment from its surrounding text. Because BCs can tolerate a certain margin of error, the translator need not bother to manually verify every alignment segment prior to beginning to use the tool, which can represent a significant time saving.

Another potential drawback of automation is that the system searches for all matches, even in cases where the translator may not need help with a particular passage. For example, if the auto-concordance feature in *SDL Trados* is activated, it may retrieve and display matches for phrases such as *'because of the'* or *'in order*

to', for which an experienced translator is unlikely to need assistance. This can be distracting because the fact that information has been retrieved means that the translator will probably at least have a brief look at what the system has proposed, which takes time and is disruptive to the translation process. And the return on investment is bound to be low for time spent looking at matches for segments for which no translation assistance was required in the first place. In contrast, when working with a BC, the translator initiates the searches and therefore only looks for passages for which he requires help.

In addition, the fact that many TMs, including *SDL Trados*, automatically copy and paste fuzzy matches or term matches directly into the target text can sometimes be a hindrance. In a study comparing interactive use of TMs against the use of the automatic pre-translation feature, Wallis (2006:48) found that translators generally preferred not to use automatic pre-translation, noting that depending on the amount of editing required to produce a desirable target segment, it may actually be faster for the translator to type the translation from scratch rather than editing the proposed segment. In contrast, a BC does not automatically paste any text directly into the target document, which can be a good thing or a bad thing depending on the quality of the match retrieved.

A small point, but one that is worth mentioning nonetheless is that TMs often require a great deal of user-initiated clicking in order to view or use the "automatically" retrieved information. For example, in *SDL Trados*, when working in interactive mode, the user must click in order to instruct the system to conduct a search for each new segment. Once the search has been conducted, only the highest-ranked match is automatically presented to the user, but depending on the translator's needs, this is not necessarily the match that will be the most helpful. In fact, a study by Fifer (2007:103) showed that translators found the 2nd ranked match to be more useful than the highest ranked match in close to 40% of cases. However, there are extra clicks involved in pulling up and viewing additional matches. Lastly, when the auto-concordance feature is activated, if the system does not find any sentence-level matches for the current segment, it automatically opens the concordance window and displays the results; however, in so doing, it makes the concordance window the active window, so the translator has to make a point of clicking back in the target field before starting to type, otherwise the text will be inadvertently written to the search field of the concordance window. It is true that there is also typing and clicking to be done when using a BC, but the point we want to make here is that BCs such as *ParaConc* do not profess to use automation as a time-saver. Moreover, the lack of automation may actually save time in some cases. For example, in *ParaConc*, all the matches are displayed at once and the user can peruse them at a glance instead of having to click through them.

Finally, it should be noted that not all features of TMs are in fact automated. In *SDL Trados*, for example, the termbase that is used to identify term matches must

be manually pre-stocked with term records by the translator prior to beginning a translation job. There are other tools, such as term extraction systems, that can be used to help with this task, but these tools are not without their own sets of challenges.[9] In contrast, as pointed out by Arrouart and Bédard (2001: 30), when a translator consults a parallel bilingual corpus using a BC, he has at his disposal a sort of "full-text glossary" which, by its very nature, contains countless "term records" that the translator has not yet had the time to formalise. Arrouart and Bédard go on to observe that one day, such resources may well supplant carefully managed collections of term records. Another relevant observation is made by Lauriston (1997: 180), who notes that

> For translators. . .the quantity of information provided is often more important than the quality. They are usually able to separate the wheat from the chaff and even turn the chaff into palatable solutions to a particular communications problem.

In other words, rather than being shown a single proposal at a time, as is the case with a TM, translators might prefer to see all examples of the search pattern in context and evaluate for themselves which offers the most viable solution to the translation problem at hand.

In summary, while less-automated tools such as BCs might appear to achieve less, they may be quicker to provide translators with results they can actually use, and they are likely to be more tolerant of unexpected situations. Of course, using such tools may call for a higher level of inventiveness or creativity on the part of the user, but thankfully, these are qualities that translators typically possess in abundance.

4.2 Search flexibility

It was noted in Section 2.1.1 that one of the perceived limitations of BCs is the nature of the searches that can be conducted. Typically, BCs search for occurrences in the corpus that precisely match the search pattern entered by the user. In contrast TMs can make use of a fuzzy matching technique that can identify patterns that are similar to, but do not precisely match, the source segment.

However, a fuzzy match is not a panacea. When using fuzzy matching techniques, the translator can set the sensitivity threshold of the match; in other words, the translator can decide how similar the two segments must be in order for a TU to be retrieved and displayed. Setting the appropriate sensitivity threshold can actually be quite tricky: if the threshold is set too high (e.g., 95% similarity), then potentially useful matches may be overlooked and the translator will be forced to

9. For more information on term extraction systems, see Cabré et al. (2001) or Bowker (2002a).

do unnecessary independent research. This is an information retrieval problem known as "silence". In contrast, if the sensitivity threshold is set too low (e.g., 30% similarity), then irrelevant or "noisy" segments may be erroneously retrieved and the translator will waste time weeding through the non-pertinent data. The problem is that setting the fuzzy match threshold at a maximally useful value is quite challenging. Some authors, such as Yunker (2003: 224), recommend never setting the threshold below 80%, while others, such as O'Brien (1998: 117), claim that a setting as low as 40% can still produce useful matches. Meanwhile other authors suggest that the ideal setting is somewhere in between. In addition, as noted in Section 2.2, even if a fuzzy match has a high percentage of similarity, it may not be that useful to the translator since the matching is based on surface structure similarities rather than semantic similarities. For instance, the following would be retrieved as a good match in a TM since the two segments strongly resemble each other on the surface, differing by only two characters: *File the form. / Fill the dorm.* In contrast, the following pair would not be retrieved because they are not superficially similar, though they are closely linked semantically: *File the form. / He is re-filing those forms.* Researchers such as Macklovitch and Russell (2000), Bowker (2002b) and Somers (2003), have all pointed out that a translator who is looking for an equivalent of a given segment would find the translation of a semantically-related segment to be more useful than that of a segment which bears only a superficial resemblance to the source text segment. With a BC, a translator could use his own knowledge of semantics to try to formulate more relevant queries, but with a TM, the translator has no input into the search patterns used.

Moreover, as mentioned in Section 2.1, many BCs have developed a number of additional flexible searching techniques which, though still manually initiated, can approximate to some extent the results of a fuzzy match. For example, *ParaConc* offers the possibility of using operators such as wildcards as part of a search. If used properly, these operators can increase the flexibility of a search (e.g., by finding inflected forms). However, as was the case with fuzzy matching, they can also lead to problems if they are not used rigorously. For instance, in an effort to retrieve examples of all forms of the verb *'to enter'*, a translator may input a pattern such as *'enter*'* where the * can be used to represent any string of characters. However, this pattern will also retrieve occurrences of all other words beginning with the string *'enter'* (e.g., *'enterprise'*, *'entertain'*). As a result, the translator may inadvertently be presented with irrelevant or noisy data.

The nice thing about working with a BC, however, is that the translator does have control over the search pattern that is entered, so by learning the proper search syntax and by gaining some experience, translators can learn which types of patterns are likely to produce valuable information and which are likely to waste time. When working with a TM, however, the translator has no control over the search pattern that is used. For example, as mentioned in Section 2.1, the Parallel

search option offered by *ParaConc* allows a translator to limit a search to a given word sense, whereas this cannot be achieved using a TM, which operates strictly on the basis of superficial pattern matching.

4.3 Consistency

Another highly advertised feature of TMs is that they promote consistency in translation. The question that has been raised by some translators, however, is whether this is always desirable.

Merkel (1998: 143) conducted a survey of 13 translators who were using TMs to carry out the translation of software manuals. One of the questions asked was whether they preferred consistent translations of a given source segment in two different contexts. The choice of answer was either "yes" or "no", with space for the respondent to elaborate on the motivations for his/her choice. Upon examining the completed questionnaires, Merkel (1998: 143) noted that "it became apparent that there was a need for a third response, in between 'yes' and 'no', namely a response which we can call 'doesn't matter'. This applies when the translator in the justification for the choice has indicated that the translation could be consistent, but that it would not matter whether the source segment was also translated differently." This raises an interesting point: in contrast to what many TM vendors would have us believe, while consistency may sometimes be desirable, it may not always be strictly necessary.

Furthermore, there may even be cases where consistency is not at all appropriate. For instance, the translators consulted as part of Merkel's survey warn that there is a need to evaluate a proposed match within the new context, and that it may not always be automatically acceptable. This is particularly true in the case of different structural contexts (e.g., sentence vs heading vs table cell), where caution should be used in applying a single consistent translation (Merkel 1998: 145).

4.4 Other quality-related issues

In addition to the question of consistency, other quality-related issues have been raised by translators working with TMs. One of the most significant, which was briefly introduced in Section 2.2, is the fact that TM databases store isolated segment pairs, rather than complete texts. In the words of Arrouart and Bédard (2001: 30), a TM is actually a memory of sentences out of context. Macklovitch and Russell (2000) have voiced a similar criticism noting that it is not possible to see a context that is larger than a given segment.[10]

10. Note that some recent versions of TMs, such as Fusion Translate by JiveFusion Technologies Development Corp (http://www.jivefusiontech.com), have addressed this problem by making

This can be problematic because the sentences in a text generally depend on each other in various ways. For example, when we read/write the third sentence in a text, we can refer back to information already presented in the first two sentences, which means that it is possible to use pronouns, deictic and cataphoric references, etc. However, if we take that third sentence in isolation, the antecedents of such references may not be clear.

In addition, because languages do not have a one-for-one correspondence or the same stylistic requirements, translators who are trying to convey the overall message of a text may map the information to the sentences in the target text in a way that differs from how that information was originally dispersed among the source text sentences. The result is that even if the two texts are considered to be equivalent when taken as a whole, the sentences in a translation may not depend on each other in precisely the same way in which the source text sentences do (Bédard 2000).

In order to maximise the "recyclability" of a text, a translator working with a TM may choose to structure the sentences in the target text to match those in the source text, and he may choose to avoid using pronouns or other references. According to Heyn (1998: 135) and Mogensen (2000: 28), the result may be a text that is inherently less coherent or readable, and of a lesser overall quality. This is described by Bédard (2000) as a "sentence salad" rather than a text, and by Mossop (2006: 790) as a "collage translation". This sentence salad or collage effect is exacerbated when the sentences in a TM come from a variety of different texts that have been translated by different translators. Each text and translator will have a different style, and when sentences from each are brought together, the resulting text will be a stylistic hodgepodge that is full of disparity (Delisle 2006: 162). It is highly unlikely that the source text has been created in such a fashion (i.e., by asking a variety of authors to contribute individual sentences), so it is questionable whether this approach should be used to produce a translation, which is also a text in (and of) itself.[11]

Another quality-related problem is that errors contained in TMs may come back to haunt a translator if the database is not scrupulously maintained in order to correct such errors. Lanctôt (2001: 30) provides the following account of a translator who carefully stores all his translations in a TM, but who does not update the contents to reflect corrections made by the client to the final document. When the client sends a document that closely resembles a version of a document

---

it possible to reconstitute the original texts from their constituent parts. At the present time, however, this is still a rare feature in TM tools.

11. A much more detailed exploration of the impact of TM use on "text" can be found in Bowker (2006).

previously translated the year before, the translator uses the TM and blithely reproduces the same errors in the new translation. The client is irritated because the same passages that were corrected last year need to be corrected again. This is not the kind of added value the client was looking for.

It is worth pointing out that a BC will also produce less-than-satisfactory results if the contents of the corpus are not of high quality. The main advantage offered by a BC in this regard is that it is much more straightforward to update the corpus with a corrected text than it is to fix erroneous TUs in a TM.

4.5 Translators' attitudes and satisfaction

An important point to consider with regard to any tool is whether or not the intended users enjoy working with it. As noted above, Wallis (2006: 77) recounts that translators working with TMs reported being less happy when they were required to use the pre-translation function, which automatically pasted matches directly into the target text, because they felt they had less control over the overall style and coherence of the text. Also with regard to TMs, Merkel (1998: 140) observes that some translators "fear that translation work will become more tedious and boring, and that some of the creative aspects of the job will disappear with the increasing use of translation memory tools." Merkel (1998: 141) goes on to note that there is concern that a translator who works with a TM may be reduced to somebody who simply has to press the OK button.

In a similar vein, Bédard (2000) expresses concern that translators may lose motivation when working with a TM because they risk becoming "translators of sentences" rather than "translators of texts". In order to maximise recyclability when working with a TM, translators are encouraged to translate one source text sentence by one target text sentence. However, as noted in Section 4.4, the aim of most translators is not to translate sentences, but rather to translate a message. To do this effectively, translators often need to work outside the artificial boundaries of end-of-sentence markers, and they may therefore feel constrained by the sentence-by-sentence approach imposed by TMs. In contrast, Arrouart and Bédard (2001: 30) have observed that when working with a BC, few constraints are imposed by the tool and translators therefore have more freedom to work as they wish.

Another difficulty that may be faced by translators working with TMs is that they may be biased by what the system presents. In other words, after a translator has seen a suggestion from the database, it may be difficult to think of another way of expressing that thought, so he may use the suggested translation even if it does not fit very coherently into the text as a whole. When using a BC, however, a translator is more likely to be seeking inspiration for handling a shorter term or expression, rather than a complete segment match, so he is less likely to feel

unduly influenced by the overall structure of the sentence contained in the corpus. He is also more likely to find examples of that term used in a variety of ways, so he can pick the usage that is most suitable for integration into the text as a whole. In this way, a translator feels like he is making his own decisions, rather than having someone else's decisions forced upon him.

The very fact that there are multiple ways to render a given passage in another language may also be a reason why some translators are unhappy about using a TM. Merkel (1998: 148) notes that as part of his survey, translators were presented with several different options as translations of a given passage. The choice of "best translation option" varied widely among translators, which leads him to believe that it may be difficult to encourage translators to accept suggestions from TMs. Fifer (2007: 101) reports similar findings, noting that when presented with a selection of matches from a TM database, the ten translators who were members the test group regularly identified different matches as being the most helpful for resolving a given translation difficulty.

A related problem that has to do with different working styles of translators is described by Lanctôt (2001: 30). When multiple translators are sharing a single TM over a network, it may be that translator A, for example, works by ploughing through a text to complete a full rough draft, and he then goes back over the text a second and third time to clean up any outstanding problems (e.g., terminological, stylistic). In contrast, translator B's approach is to go more slowly, doing terminological research and addressing stylistic concerns as he goes along. In Lanctôt's scenario, translator B is frustrated by the suggestions proposed by the TM – many of which were produced as part of translator A's first rough draft.

5. Concluding remarks

The aim of this paper has been to introduce and present an analysis of some of the strengths and weaknesses of two categories of tool: BCs and TMs. As noted in Section 3, although TMs are widely promoted in the translation industry, BCs are less well known and, in some cases, translators who are vaguely aware of them may erroneously believe that such tools have been completely superseded by TMs and therefore have no interest for the translation community.

It is not our intention to promote one type of tool over the other. Instead, we feel that the two technologies may be considered complementary, rather than competing, in the sense that one may be preferred in certain circumstances, while the other may be favoured in a different situation. Basically, it comes down to a translator being aware of how the two types of tool work and the potential advantages that each offers. The translator must then be able to choose the right tool for

the job at hand. What follows are some possible considerations that a translator might take into account when deciding which tool to use.

One critical factor that comes into play when choosing which tool to use is the nature of the job itself. Not all translation jobs are equal, and they will not necessarily all benefit from the same technology. Part of the frustration experienced by some translators using translation tools may result from them applying the tool in an inappropriate situation. Sometimes it may be the client who insists that a particular tool be used without really understanding that it may not be suitable, whereas in other cases, it may be the translator who is not aware that another more appropriate tool exists.

Another consideration might be the size of the job. In many cases, if a translation job amounts to just a few thousand words, this typically comes with a short deadline. And since each job is different, it may not be possible to use any tool without making some adaptation to either the tool or the corpus that it will be used to process. As pointed out by Bédard (1995: 28), by the time the tool is made operational, the deadline may be fast approaching and the cost of getting the tool to work may have exceeded the value of the job. As noted in Section 4.1, TMs typically require more in terms of a learning curve and data preparation than do BCs, so it may be that while a TM could provide a good return on investment for a large job, a BC might be a better choice for a small job.

Text type is also an important factor to consider. There are certain types of texts and writing styles that are highly conducive to being processed with a TM. In particular, texts that are a revision of a previous document (e.g., an updated version of a user manual, a re-negotiated collective agreement) are good candidates for translation with a TM because they will contain many repetitions at the sentence (or even paragraph) level. Another good candidate for use with a TM is a text where the repetitive sentences are varied (i.e., many sentences with few occurrences of each) and scattered throughout the document. However, such documents are not the only type that translators work with. Many translators are faced with texts that contain repetition primarily at the sub-sentence level. In such a case, since the manual searches initiated by the translator using a BC may be more flexible and productive than the auto-concordance search in a TM, a BC may be preferable.

The choice may also be motivated by whether the work is being done for a regular client or for a new client. If a translator works regularly for a particular client and has a corpus consisting exclusively or primarily of similar types of texts translated for that client, it may be reasonable to use a TM since presumably the "sentence salad" or "collage" effect will be lessened by the fact that the documents will all contain similar terminological and stylistic preferences. In contrast, if the job is for a new client and the corpus does not contain previous work done for that client, perhaps a BC would be a better choice since the translator could consult it

merely for inspiration without feeling constrained by choices made previously to suit other clients or text types.

The decision of whether to use a TM or a BC may also depend on the translator's preferred working style. Just as some drivers prefer driving a car with a manual transmission over one with an automatic transmission, some translators may favour a system that does a greater degree of automatic text processing (e.g., TM), while others may opt for one that does less (e.g., BC).

Another relevant issue may be the amount of experience the translator has. A translator who is very experienced may prefer the flexibility offered by a BC, which allows him to look up only those expressions for which he needs help. In contrast, a translator who is just embarking on his career may value the fact that a TM automatically makes suggestions for all types of text strings.

A final factor that may come into play could be cost. A single licence for a BC typically costs less than $200 (US), whereas a single licence for a limited version[12] of a TM retails for closer to $1000 (US). It is true that there are usually additional features present with TM software (e.g., termbase, term extraction tool), and if these features will be used, then the additional cost may be worthwhile. However, if a translator intends to use mainly the concordancing feature of a tool, then it may be preferable to purchase a more modestly priced BC.

In closing, it is also worth pointing out that technology is continuing to develop, and a new class of tools is beginning to emerge which seeks to combines the best features of both BCs and TMs in an effort to offer users the greatest amount of flexibility to meet their varying needs. As reported by Lagoudaki (2007), a survey carried out by Imperial College London[13] of 874 translators, of which 82.5% claim to use TMs, revealed some interesting findings, including suggestions for future directions in the development of TM systems. For instance, respondents indicated that they feel "segmentation should move towards translation units of a smaller size – at phrase level instead of sentence level – so that TM systems can have more chances of finding matches" (Lagoudaki 2007: 76). Another request is for a TM system to show "the actual context – some lines of text before and after the match – that will make the translator feel a greater certainty about his or her choice of the right translation" (Lagoudaki 2007: 78). Meanwhile, Gervais (2006) reports on the results of another survey – this one conducted by the translation technology development company MultiCorpora R&D – where translators were asked what changes to TM tools would be necessary in order to achieve a wider adoption and

12. "Freelance" or "lite" versions may restrict database or termbase size or may lack network capabilities. "Professional" versions may cost several thousand dollars.

13. The full Translation Memories Survey 2006 is available at http://www3.imperial.ac.uk/portallive/docs/1/7307707.pdf

greater benefit realisation from those tools. Once again, the responses included requests for easy access to full context rather than just a single segment, and the ability to effectively recycle expressions of any length (including terms, phrases, sentences and even paragraphs). A further request was for a way to eliminate the time-consuming and labour-intensive TM alignment validation step. Interestingly, these requests for improvements to TMs correspond to some of the features that are already offered by BCs.

As noted previously, some of the strengths of BCs include the possibility of searching for shorter patterns, such as phrases or expressions, and the possibility of viewing the retrieved matches in full context, which in turn allows for the possibility of having a less rigid alignment process. If these BC features are integrated into a TM system, users may well get the best of both worlds. In the opinion of Gervais (2006: 48), a tool such as MultiTrans,[14] which he refers to as a "next-generation TM", meets all these criteria, thus combining some of the strengths of both BCs and TMs. Only time will tell whether this new hybrid tool will actually replace BCs and TMs, or whether it will find its own niche as a complementary solution.

References

Allen, J. (2003) Post-editing. In Somers, H. (ed.), *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins. 297–317.

Arrouart, C. (2003) Les mémoires de traduction et la formation universitaire : quelques pistes de réflexion. In *Meta* 48 (3): 476–479.

Arrouart, C. and C. Bédard (2001) Éloge du bitexte. In *Circuit* 73, 30.

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

Barlow, M. (2000) Parallel Texts in Language Teaching. In Botley, S., T. McEnery and A. Wilson (eds) *Multilingual Corpora in Teaching and Research*, Amsterdam: Rodopi, 106–115.

Bédard, C. (1995) L'automatisation: faut-il y croire. In *Circuit* 48, 28.

Bédard, C. (1998) Ce qu'il faut savoir sur les mémoires de traduction. In *Circuit* 60, 25.

Bédard, C. (2000) Mémoire de traduction cherche traducteur de phrases… In *Traduire*. Société française des traducteurs, Paris: INIST-CNRS, Cote INIST, 186: 41–49.

Bernardini, S. (2002) Educating Translators for the Challenges of the New Millennium: The Potential of Parallel Bi-directional Corpora. In Maia, B., J. Haller & M. Ulrych (eds.) *Training the Language Services Provider for the New Millennium*, 173–186. Faculdade de Letras da Universidade do Porto.

Biau Gil, J. R. (2004) SDLX 2004. *Translating Today Magazine*, 1: 40–41.

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

14. For more information on MultiTrans, see http://www.multicorpora.com

Bowker, L. (2002b) Information Retrieval in Translation Memory Systems: Assessment of Current Limitations and Possibilities for Future Development. In *Knowledge Organisation,* 29 (3/4), 198–203.

Bowker, L. (2004) Corpus Resources for Translators: Academic Luxury or Professional Necessity? In *TradTerm*, 10: 213–247.

Bowker, L. (2006) Translation Memory and 'Text'. In Bowker, L. (ed.), *Lexicography, Terminology and Translation: Text-based Studies in Honour of Ingrid Meyer*, Ottawa: University of Ottawa Press, 175–187.

Cabré Castellví, M. T., R. Estopà Bagot and J. Vivaldi Palatresi (2001) Automatic term detection: A Review of Current Systems. In Bourigault, D., C. Jacquemin, and M.-C. L'Homme (eds.), *Recent Advances in Computational Terminology*. Amsterdam/Phildadelphia: John Benjamins, 53–87.

De Cesaris, J. (1996) Computerised Translation Managers as Teaching Aids. In Dollerup, C. & V. Appel (eds.), *Teaching Translation and Interpreting 3: New Horizons*, Amsterdam/Philadelphia: John Benjamins, 263–269.

Delisle, J. (2006) Criticizing Translations: The Notion of Disparity. In Bowker, L. (ed.) *Lexicography, Terminology and Translation: Text-based Studies in Honour of Ingrid Meyer*, Ottawa: University of Ottawa Press, 159–173.

Fifer, M. (2007) The Fuzzy Factor: An Empirical Investigation of Fuzzy Matching in the Context of Translation Memory Systems. University of Ottawa, Canada. MA dissertation.

Gale, W. A. & K. W. Church (1993) A program for aligning sentences in bilingual corpora. In *Computational Linguistics* 19: 75–102.

Gervais, D. (2006) Product profile: A corpus-based approach. In *Multilingual* 77, vol. 17, issue 1, 47–49.

Hallett, T. (2006) Transit and TRADOS: Converging functions, diverging approaches. In *Localisation Focus* 5 (4), 9–11.

Hansen, S. & E. Teich (2002) The Creation and Exploitation of a Translation Reference Corpus. In Yuste Rodrigo, E. (ed.), *Proceedings of the Workshop on Language Resources in Translation Work and Research*, Paris: ELRA/ELDA,[15] 1–4.

Heyn, M. (1998) Translation Memories: Insights and Prospects. In Bowker, L., M. Cronin, D. Kenny and J. Pearson (eds.), *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome Publishing, 123–136.

Johns, T. (1986) Microconcord: A Language Learner's Research Tool. In *System* 14 (2): 151–162.

Kenny, D. (1999) CAT Tools in an Academic Environment. In *Target* 11 (1): 65–82.

Lagoudaki, E. (2007) Translators Evaluate TM Systems – A Survey. In *MultiLingual* 18 (2), 75–78.

Lanctôt, F. (2001) Splendeurs et petites misères… des mémoires de traduction. In *Circuit* 72: 30.

L'Homme, M.-C. (1999) *Initiation à la traductique*. Brossard, Quebec: Linguatech.

Macklovitch, E. & G. Russell (2000) What's Been Forgotten in Translation Memory. In White, J. S. (ed.). *Envisioning Machine Translation in the Information Future.* 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000. Berlin: Springer Verlag, 137–146.

15. European Language Resources Association / European Language resources Distribution Agency

Merkel, M. (1998) "Consistency and Variation in Technical Translation: A Study of Translators' Attitudes," In Bowker, L., M. Cronin, D. Kenny and J. Pearson (eds.), *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome Publishing, 137–149.

Mindt, D. (1986) Corpus, Grammar and Teaching English as a Foreign Language. In Leitner, G. (ed.), *The English Reference Grammar: Language and Linguistics, Writers and Readers*, Tübingen: Niemeyer, 125–139.

Mogensen, E. (2000) Orwellian Linguistics. In *Language International* 12 (5), 28–31.

Mossop, B. (2006) Has computerisation changed translation? In *Meta* 51 (4), 787–793.

Musale, S. (2004) Getting More From Translation Memory. In *Localisation Focus* 3 (1), 9–10.

O'Brien, S. (1998) Practical Experience of Computer-Aided Translation Tools in the Localisation Industry. In Bowker, L., M. Cronin, D. Kenny & J. Pearson (eds.), *Unity in Diversity? Current Trends in Translation Studies,* Manchester: St. Jerome Publishing, 115–22.

Palumbo, G. (2002) The Use of Phraseology for Training and Research in the Translation of LSP Texts. In Maia, B., J. Haller and M. Ulrych, (eds.), *Training the Language Services Provider for the New Millennium*", Faculdade de Letras da Universidade do Porto, 199–212.

Pearson, J. (2000) Une tentative d'exploitation bi-directionnelle d'un corpus bilingue. In *Cahiers de Grammaire* 25, 53–69.

Reppen, R. (2001) Review of MonoConc Pro and WordSmith Tools. In *Language Learning and Technology* 5 (1), 32–36.

Shadbolt, D. (2002) The Translation Industry in Canada. In *Multilingual Computing and Technology* 13 (2): 30–34.

Somers, H. (2003) Translation memory systems. In Somers, H. (ed.), *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins, 31–47.

Sprung, R. C. (2000) Introduction. In Sprung, R.C. (ed.), *Translating into Success: Cutting-edge strategies for going multilingual in a global age*, Amsterdam/Philadelphia: John Benjamins, ix–xxii.

Tagnin, S. E. O. (2002) Corpora and the Innocent Translator: How Can They Help Him? In Lewandowska-Tomaszczyk, B. and M. Thelen (eds.), *Translation and Meaning, Part 6*, Maastricht: Hogeschool Zuyd, 489–496

Wallis, J. (2006) *Interactive Translation vs Pre-translation in the Context of Translation Memory Systems: Investigating the effects of translation method on productivity, quality and translator satisfaction*. University of Ottawa, Canada. MA dissertation.

Waßmer, T. (2005) Trados 7 and SDLX 2005. In *MultiLingual Computing & Technology* 16 (8), 27–32.

Yunker, J. (2003) *Beyond Borders: Web Localisation Strategies.* Indianapolis, Indiana: New Riders.

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

Zerfaß, A. (2002) Comparing Basic Features of TM tools. In *Language Technology Supplement, Multilingual Computing and Technology #51*, vol. 13, issue 7, 11–14.

CHAPTER 2

Interactive reference grammars

Exploiting parallel and comparable treebanks for translation

Silvia Hansen-Schirra
Johannes Gutenberg University Mainz

This paper discusses the role of annotated corpora as works of reference for grammatical translation problems. Within this context, the English-German CroCo Corpus and its multi-layer alignment and annotation are introduced. It is described how the corpus is exploited as interactive resource to display translation solutions for typologically problematic constructions. Additionally, the Penn and TiGer Treebanks are used as comparable corpora for English and German. The linguistic enrichment of the treebanks, i.e. their syntactic annotation, is described and corpus query techniques relevant for translation problems are shown. Relevant structures are extracted from the treebanks and translation candidates are displayed and discussed. The advantage of this technique is that translation solutions are extracted from published translations, i.e. language in use. Consequently, they are more comprehensive and inventive than dictionary entries or descriptions in grammars are. Treebanks could thus be used as an interactive reference grammar in translation education and practice.

1. Introduction

The basic idea of using corpora in translator training is that a parallel corpus consists of a more comprehensive and diverse variety of source language items and possible translation solutions than a dictionary could ever display. Thus, in translator training and practice, parallel corpora are used, for instance, for terminology look-up, for teaching the usage of collocations or as translation memories. However, the grammatical point of view has not yet been addressed in corpus-based translation work. The linguistic enrichment of parallel texts is mostly limited to sentence alignment; the exploitation facilities are restricted to string-based queries.

There are, however, translation problems which are due to typological differences between languages. The existing descriptions of these rather grammatical problems are all example-based. This is a good way to describe and define a

phenomenon; it is however not ideal for practical applications because it cannot take into account all instances to be found in the day-to-day work of translators. While example-based studies on typologically driven translation problems show the range of phenomena, they cannot give evidence on the frequency and context of their occurrences. For this purpose, an empirical corpus-based study is more suitable. And when speaking about grammatical phenomena, it would be good to have a resource where these typological characteristics are encoded and source and target language equivalents are aligned. Such a resource could then be used as an interactive reference grammar in translation education and practice.

In this chapter we explore the role of annotated corpora as works of reference for grammatical translation problems (cf. Section 2). Up to now the annotation of translation corpora, i.e., their linguistic enrichment has been carried out in order to empirically investigate the properties of translated text. Here, the English-German CroCo Corpus, aligned and annotated on several linguistic layers, is exploited as interactive resource to display translation solutions for typologically problematic constructions. Additionally, the Penn and TiGer Treebanks are used as comparable corpora for English and German.

We show how a corpus needs to be enriched (alignment, annotation), so that it becomes possible to pose queries to it that are interesting and relevant from a translation point of view. Furthermore, we show how monolingual and multilingual treebanks can then be queried with concordancing tools. We illustrate the use of these interactive reference grammars for solving the following typical translation problems that occur in translating from English into German (cf. Section 3): English seems to be far more productive concerning cleft sentences, raising constructions and deletions, while German is characterised through more word order freedom. Therefore, in the process of translating from English into German, compensations have to be found and the word order has to be adapted. On the basis of syntactic corpus annotation, relevant structures are extracted from the treebanks and translation candidates are displayed. The advantage of this technique is that translation solutions are extracted from published translations, i.e., language in use, and are thus more comprehensive and inventive than dictionary entries or descriptions in grammars are.

2. Annotated corpus resources

As for many other linguistic areas, also for translation practice and education, the primary value of employing corpora is the opportunity of investigating large amounts of data and of conducting empirical research on translations. This means that translators are able to move from the observation of text samples to the investigation of larger sets of texts in different constellations. Typically, two types of

corpus design are employed in corpus-based translation work: the parallel corpus and the comparable corpus. Parallel corpora consist of source language texts and translations of those texts into a target language. They are commonly employed for terminology look-up and for teaching the usage of collocations. Furthermore, they are used for bilingual lexicography and more recently also as training material in machine translation. Comparable corpora are collections of translations (from one or more source languages) into one target language and original texts in the target language. Comparable corpora are mainly used in translation research – they reveal properties which are characteristic for translated texts only, e.g., explicitation or simplification (for further information on corpora in translation studies see Olohan (2004)).

At the technical level, translation studies benefits from existing corpus-linguistic techniques, such as key word in context (KWIC) concordances, automatic frequency counts of words, etc. While the use of such tools has become an integral part of technical practice in corpus-based translation work, more sophisticated corpus techniques, notably tools for corpus annotation, corpus maintenance and corpus query as they have been developed for monolingual corpora, have only rarely been exploited yet. Thus, new methodological and technical challenges emerge for the discipline: Since this chapter will show how treebanks, i.e., corpora annotated with syntax trees, can be used for translation, in the following the creation of monolingual (Subsection 2.1) as well as multilingual treebanks (Subsection 2.2) will be discussed.

2.1 Monolingual treebanking

One of the first and best known treebanks is the Penn Treebank for the English language (Marcus et al. 1994), which consists of more than 1 million words of newspaper text. It contains part-of-speech tagging as well as syntactic and semantic annotation. A bracketing format is used to encode predicate argument structure and trace-filler mechanisms are used to represent discontinuous phenomena. Other treebanks for English are, for instance, the Susanne Corpus (Sampson 1995) (containing detailed part-of-speech information and phrase structure annotation), the Lancaster Parsed Corpus (Leech 1992) (representing phrase structure annotation by means of labelled bracketing) and the British part of the International Corpus of English (Greenbaum 1996) (about 1 million words of British English that were tagged, parsed and afterwards checked).

For languages other than English, a fairly well-known treebank is the Prague Dependency Treebank for Czech (Hajic 1999). It contains more than 1 million tokens and is annotated on three levels: on the morphological level (tags, lemmata, word forms), on the syntactic level (using dependency syntax) and on the tectogrammatical level (encoding functions such as actor, patient, etc.). Recently,

treebank projects for other languages have come to life as well, e.g., for French (Abeillé et al. 2000), Italian (Bosco et al. 2000), Spanish (Moreno et al. 2000), etc.

For German, the Verbmobil Treebank (Hinrichs et al. 2000) and the Tübingen Treebank (Telljohann et al. 2006) are available. However, they are rather small as reference work and restricted to spoken language (as in the case of Verbmobil). In contrast, the NEGRA/TiGer corpora (Brants et al. 2003), including 70,000 sentences, are the ideal basis for empirical investigations. For their annotation, a hybrid framework is used which combines advantages of dependency grammar and phrase structure grammar. The syntactic structure is represented by a tree. The branches of a tree may cross, allowing the encoding of local and non-local dependencies and eliminating the need for traces. This approach has considerable advantages for free-word order languages such as German, which show a large variety of discontinuous constituency types (Skut et al. 1997). The linguistic annotation of each sentence in the TiGer Treebank is represented on a number of different levels: Information on part-of-speech, morphology and lemmata is encoded in terminal nodes (on the word level). Non-terminal nodes are labelled with phrase categories. The edges of a tree represent syntactic functions. Syntactic structures are rather flat and simple in order to reduce the potential for attachment ambiguities. The distinction between arguments and adjuncts, for instance, is not expressed in the constituent structure, but is instead encoded by means of syntactic functions. Secondary edges, i.e., labelled directed arcs between arbitrary nodes, are used to encode coordination information.

Instead of having an automatic parser as pre-processor and a human annotator as postprocessor (as in the Penn Treebank project), interactive annotation with the annotation tool (Brants & Plaehn 2000) is used for the annotation process, efficiently combining automatic parsing and human annotation. The TnT tagger (Brants 2000) and the parser using Cascaded Markov Models (Brants 1999) generate small parts of the annotation, which are immediately presented visually to the human annotator, who can either accept, correct or reject it. Based on the annotator's decision, the parser proposes the next part of the annotation, which is again submitted to the annotator's judgement. This process is repeated until the annotation of the sentence is complete. The advantage of this interactive method is that the human decisions can be used by the automatic parser. Thus, errors made by the automatic parser at lower levels are corrected instantly and do not 'shine through' on higher levels. The chances grow that the automatic parser proposes correct analyses on higher levels. In order to achieve a high level of consistency and to avoid mistakes, we use a very thorough approach to the annotation: First, each sentence is annotated independently by two annotators. With the support of scripts, they afterwards compare their annotations and correct obvious mistakes. Remaining differences are submitted to a discussion between the annotators. Al-

though this process is rather time consuming, it has proven to be highly beneficial for the accuracy of the annotation.

Syntactically annotated corpora provide a wealth of information which can only be exploited with an adequate query tool and query language. Thus, the powerful search engine TiGerSearch has been developed for the NEGRA and TiGer Treebanks (Lezius 2002).

In Section 3, we will discuss how monolingual treebanks can be used as comparable reference works during the translation process. On the basis of their syntactic annotation, typical patterns can be extracted and evaluated empirically. This helps the translator to quickly find the most suitable translation solution in terms of grammatical structure.

2.2 Multilingual treebanking

Recently, parallel treebanks have proven to be useful for multilingual grammar induction, as test suites and gold standards for alignment tools and multilingual taggers and parsers as well as for the development of corpus-based machine translation systems. Despite these manifold applications, there are only very few parallel treebanks under development, for example the Prague Czech-English Dependency Treebank (Cuřín et al. 2004). For German, the above mentioned Verbmobil Treebanks (Hinrichs et al. 2000), the German-English-Swedish Sofie Treebank (Volk et al. 2006) as well as the Ptolemaios Treebank (Kuhn & Jellinghaus 2006) are available. However, all treebanks including German are rather small and they comprise one translation direction only. Sometimes it is even unclear which of the languages can be seen as source and which as target language.

In contrast to this, the CroCo Corpus built, up for investigating special properties of translations, comprises 1 million words of German originals and their English translations as well as English sources and their German translations. These sub-corpora were collected from eight registers, which are all relevant for translations: popular-scientific texts, tourism leaflets, prepared speeches, political essays on economics, fictional texts, corporate communication, instruction manuals and websites (Hansen-Schirra et al. 2006).

A characteristic feature of this corpus is the annotation and alignment of source and target texts on different linguistically motivated layers: The texts are annotated with parts of speech, morphology, phrase structure and grammatical functions. Tokenisation and part-of-speech tagging are performed for both, German and English, by TnT (Brants 2000), a statistical part-of-speech tagger. Morphological information is particularly relevant for German compared to the more analytic English language. Morphology is annotated in CroCo with MPro, a rule-based morphology tool (Maas 1998). This tool also provides an analysis of the phrase structure. Beyond this automatic annotation, syntactic functions are

currently annotated manually with the help of MMAX II (Müller & Strube 2003), a tool allowing assignment of self-defined categories and linking units.

Concerning the alignment of the texts, we do not only align sentences (which is state of the art in Translation Memories; e.g., Johansson et al. 1996) and words (which is state of the art in Machine Translation; cf. Och & Ney 2003) but also clauses. Word alignment is realised with GIZA++ (Och & Ney 2003), a statistical alignment tool. Clauses are aligned manually with the help of MMAX II (see above). Sentences are aligned using Win-Align, an alignment tool within the Translator's Workbench by Trados (Heyn 1996). Additionally, phrase alignment can be derived from word alignment in combination with the phrase chunking and syntactic functions can be mapped automatically across the parallel corpus. Each annotation and alignment layer is stored separately in a multi-layer stand-off XML representation format keeping the annotation and alignment of overlapping and/or discontinuous units in separate files. The mark-up builds on the XCES Standard.

The architecture of the CroCo Corpus allows us to view the annotation in aligned segments and to pose queries combining different layers (Hansen-Schirra et al. 2006). The resource thus permits the analysis of a wealth of linguistic information on each level helping us to understand the interplay of the different levels and the relationship of lower level features to more abstract concepts. For parallel concordancing, query tools such as the IMS Corpus Workbench (Christ 1994) can be employed. Its corpus query processor (CQP) allows queries for words and/or annotation tags on the basis of regular expressions. For more complex queries, the annotated data is converted into a MySQL database. On this basis, an effective exploitation of different annotation and alignment layers is guaranteed. In the following, we will demonstrate how the bilingual CroCo Corpus is used to extract parallel grammatical structures helping translators to decide on typical English-German translation problems.

3. Solving translation problems with treebanks

In many cases, typological differences between languages can be translated straightforwardly without any problems. Different grammatical morphologies are, for instance, not considered as major translation problems. There are, however, typological differences that are problematic for the translation process. Typically, these are constructions which exist in one language but do not exist or are rarely used in the other. For the translation of such constructions, the translator has to compensate them in the target language. It is, however, not always easy to find an adequate translation equivalent. For this reason, a language resource including grammatical descriptions of translation pairs can help to solve translation

problems. On this basis, translation examples of typological differences can be extracted.

In the following, we explain the advantage of such a resource on the basis of Hawkins' (1986) descriptions of typological differences for English and German.

3.1 Word order variation

According to Hawkins (1986) German word order is freer than English word order. The reason for that lies in the richer morphology of German which allows various object and adverbial movements. English, however, is characterised through a very strict SVO canonical word order. This means that English word order cannot be used for pragmatic stress, e.g., distinguishing between old and new information. For translations from English into German this poses the problem that it is tempting for the translator to preserve the word order of the source language text, which results in a rather unnatural word order distribution in the German translation (i.e., with an unusual bias on SVO constructions). For this reason it is useful to extract typical word order patterns for German and English to become aware of the differences in terms of statistics and to have a look at typical constructions in the target language.

For this purpose, we use the above described English Penn Treebank as well as the German TiGer Treebank as monolingual comparable corpora and explore them with TiGerSearch (cf. Section 2.1). In our case, the Penn Treebank serves as basis of comparison for the English source language texts and the TiGer Treebank is used as comparable resource for the German translations. They constitute the typological norm for the respective language and show how the translations should behave compared to texts in the source language.

Looking at English word order in the Penn Treebank, in 2% of all sentences the verb occurs before the subject. These constructions are typically questions or inversions. All the other sentences (98%) of the corpus are SV sentences, 12% among them with an adverbial before the subject.

The situation for German is quite different: In the TiGer Treebank, in 4% of all sentences the verb occurs in first position. Again these are typically questions or the inverted *verbum dicendi* following direct speech. In 31% of all cases we can find the verb at last position which reflects the typical word order in German subordinate clauses. Although English subordinate clauses behave like main clauses (both using SVO word order), this difference is not problematic for translation since a preservation of the English word order would lead to ungrammatical sentences in German.

The rest of all TiGer sentences (i.e., 65%) is characterised through verb second word order. These are, of course, the most interesting cases because they show

which alternatives we have to form the German *Vorfeld*, i.e., the position before the finite verb. The distribution looks as follows:

- subject in *Vorfeld* position: 55%
- adverbial in *Vorfeld* position: 25%
- predicator in *Vorfeld* position: 3%
- accusative object in *Vorfeld* position: 3%
- dative object in *Vorfeld* position: 1%
- other (prepositional object, expletive etc.): 13%

This *Vorfeld* alternation is pragmatically motivated. The main difference to English is that we find fewer subjects but more adverbials in the position before the finite verb. Furthermore, depending on the information structure of a given clause, it is also possible to move objects or predicators into initial position. The following examples illustrate how adverbials are used in the *Vorfeld* position in German:

s4451: In der Hauptstadt Seoul kam es zu Zusammenstößen zwischen Studenten und der Polizei.
s4553: Hier herrscht Demokratie.
s4567: In Israel gibt es diverse ultrarechte, deren radikale Ablehnung jeder Verständigung mit den Palästinensern immer militantere Töne angenommen hat.
s4613: In Libanon schossen Freischärler aus Freude in Luft, als sie vom Tod Rabins erfuhren.
s4622: 1949 richtete ein 23jähriger im Parlamentsgebäude eine Maschinenpistole auf Regierungschef David Ben-Gurion, wurde aber schnell überwältigt.

Since the TiGer Corpus consists of newspaper texts only, here we can find many local and temporal circumstances: *In der Hauptstadt Seoul* (*In the capital Seoul*); *Hier* (*Here*); *In Israel*; *In Libanon*; *1949* (*In 1949*). These are good examples for registerially motivated language use in terms of word order.

In some cases direct and indirect objects can be found in initial position:

s25: Das Informatik-Dienstleistungsunternehmen verkauft er 1984 für 2,5 Milliarden an General Motors.
s109: Das hielte ich für moralisch außerordentlich fragwürdig.
s328: " Deutlich über 5000 " hat die SPD-Stadtregierung jetzt jeweils binnen zwölf Monaten im Visier.
s358: Seinen Höhepunkt erreichte der Aufstand der Mostazafin eine Woche später in Meschhed im Nordosten des Landes.
s14115: Repressionen bekommen auch angesehene Personen zu spüren, die andere Ideen als der 60jährige Staatschef Ben Ali vertreten.
s14148: Chardonnay , Sauvignon Blanc, Pinot Blanc und heimische Sorten bauen sie aus.

This movement of the direct object mainly happens because of pragmatic reasons: Sentence 109 is a good example for direct object movement into the *Vorfeld* position to signalise the old information of the sentence. Here, the old information is expressed through the demonstrative pronoun *Das* (*That*), which occurs here in thematic position.

Sentence 14148 illustrates, however, that also new information, in this case *Chardonnay, Sauvignon Blanc, Pinot Blanc und heimische Sorten* (*Chardonnay, Sauvignon Blanc, Pinot Blanc and local varieties*) can be found in thematic position if there is a (contrastive) stress on this information. Because of the same pragmatic reasons, indirect objects occur in initial position:

> **s27:** Ihm gelingt es aber nicht , den Koloß , der der Regierung in Washington mehr ähnelt als jeder andere Konzern, in Schwung zu bringen.
> **s60:** " Den Herren Rao und Singh gebührt ein Platz in der Geschichte ", resumiert das britische Blatt.
> **s407:** Vielen Jugendlichen bleibt nur Vereinzelung oder die Gruppe auf der Straße.
> **s2493:** Der FDP warf Blüm vour , nicht mehr zum Beschluß zu stehen.

In sentence 27 the personal pronoun *Ihm* (*him*) indicates the old information, whereas the new information can be found in the rheme. Sentence 2493 is again a good example for putting new information with contrastive stress in initial position (i.e., the German political party *FDP*).

These concordance lines clearly illustrate how information distribution is realised in German and can structurally serve as reference for English-German translators who are biased through the English SVO word order. These examples help them to choose more idiomatic target language structures, which makes the translations more typical and idiomatic for native speakers of the target language.

3.2 Raising constructions

One phenomenon for which English is more productive than German is raising. Hawkins (1986) states that the class of verbs which trigger raising is larger in English than German. This causes problems for translations from English into German since the lacking raising possibilities have to be compensated by the translator. To find examples of English raising structures which are translated into German, we use the English-German CroCo Corpus (cf. Section 2.2). Raising constructions can be found by querying the database for a finite verb followed by a direct object which is formally realised through a clause. The following subject-to-subject raising structures are excerpted from the query output of the sub-corpus of corporate communication:

(1) We continue to benefit from the strong natural gas market in North America. – Wir profitieren weiterhin von einem starken Erdgasmarkt in Nordamerika.

(2) We defined the minivan, and will continue to do so. – Wir haben den Minivan erfunden und wir werden auch künftig neue Marktsegmente definieren.

(3) . . . and attracting the best talent possible as we continue to grow our business. – . . . und werben zur Erweiterung unseres Geschäftes die besten Talente an, die wir nur finden können.

Here, one possible translation strategy becomes obvious: The verb *continue* which occurs very frequently in the sub-corpus of corporate communication is translated by using temporal adverbials in German (*weiterhin* (*furthermore*) and *künftig* (*in the future*) in examples (1) and (2)). Another solution would be to transform the verbal group into a nominal structure (example (3)). Here, the raising construction *continue to grow* is translated with the German noun *Erweiterung* (*extension*). This kind of nominalisation seems to be a typical translation strategy for English-German.

3.3 Clefting

According to Hawkins (1986), English is also more productive concerning cleft sentences. While clefting does exist in German as well, German has other options of realising information distribution patterns, e.g., by word order variation (see Section 3.1). Therefore, in the process of translating these constructions into German, the appropriate alternative has to be found. In our annotated and aligned CroCo Corpus, cleft constructions can be found by querying the database for the pronoun *it* followed by the lemma of the verb be which is again followed by the syntactic function complement including a relative pronoun. Applying this query to the sub-corpus of corporate communication, we find the following translation pairs:

(4) It is this ownership that we truly believe helped our employees to drive toward success, despite the challenges of this year. – Mit dieser Beteiligung am Unternehmen im Rücken haben unsere Mitarbeiter nach unserer Überzeugung maßgeblich zum Erfolg des Unternehmens trotz der großen Herausforderungen dieses Jahres beigetragen.

(5) It is to everyone's credit that we accomplished so much – the best year ever in our combined history. – Dem Einsatz aller ist es zu verdanken, dass wir so viel erreicht haben.

(6) In fact, it was their persistence through some very challenging days in 1998 that helped us end the year with such strong momentum. – Tatsächlich ist es ihrem Durchhaltevermögen während einiger sehr kritischer Tage 1998 zu

verdanken, dass wir das Jahr dann doch noch mit einem solch gewaltigen Erfolg beenden konnten.

Here, two options of translating English clefts into German are shown: The first English cleft sentence is nominalised using the German adverbial *Mit dieser Beteiligung* (*with this ownership*), whereas in example (5) and (6) German infinitival constructions are chosen. In the latter examples a lexical pattern for translating clefts becomes visible: The translators used *es ist jemandem/etwas zu verdanken, dass* (*it is somebody/something to thank that*) for the translation of both English cleft sentences. This might be an indicator for a good translation strategy for clefts. To find other strategies, it can be specified in the corpus query whether the clefted element is translated and thus aligned with a German adverbial, a German subject or other realisations.

Applying the same query to the sub-corpus of popular-scientific texts, the following examples can be found:

(7) … it is N that makes this one-way function reversible – Es ist dieses N, das die Einwegfunktion umkehrbar macht, …

(8) It is these properties that make them attractive as anticancer agents. – Gerade diese Eigenschaften lassen sie als Wirkstoffe gegen Krebs vielversprechend erscheinen.

(9) It is they alone that persist from one generation to the next – Nur die Gene bleiben in der Generationenabfolge erhalten.

(10) History records that it was Galileo who was foremost in establishing … – Die Geschichte belegt, daß vor allem Galilei … etablierte.

Here again, two options of translating English clefts into German are illustrated: The first English cleft sentence, which provides a description or definition of a process, is literally translated into German. So, this seems to be an appropriate context for preserving the structure of cleft sentences. Examples (8) to (10) show that for the compensation of cleft sentences it seems to be a typical translation strategy to choose a focus particle in German. Here, the particles *gerade* (*even*), *nur* (*only*) and *vor allem* (*above all*) are used to signal the syntactic focus.

The concordances discussed in the present section clearly show that there are different translation preferences for different registers. For this reason it is useful to consult corpora which are relevant in terms of their registers. Clefting is a good example for register-specific translation behaviour.

3.4 Substitution and deletion

According to Hawkins (1986), substitutions and deletions occur more frequently in English than in German. This means that the translator has to find different

realisations, otherwise the German translation would consist of an unusual high number of substitutions and deletions. Analysing English nominal substitutions in the parallel CroCo Corpus, we extract them by querying for the English word *one* using the concordance tool CQP (see Section 2.2). The following concordances are taken from the sub-corpus of fiction:

(11) As they did not know of this one either. – Genausowenig, wie sie etwas von diesem Fluss wussten.

(12) She was surprised she felt a genuine pang – an aesthetic one – that . . . – Zu ihrer Überraschung gab es ihr – aus ästhetischen Gründen – einen heftigen Stich, daß . . .

(13) At the town of Sargigora a man with a red shoe on his left foot and a green one on his right told of a seer walking at Nandul. – In dem Städtchen Sargigora erzählte uns ein Mann mit einem roten Schuh am linken und einem grünen Schuh am rechten Fuss von einem Seher in Nandul.

All of these English substitutions are translated through German lexical items: *Fluss* (*river*), *Gründen* (*reasons*) and *Schuh* (*shoe*). This means that there is a tendency to choose repetitions, synonyms, hyponyms, hyperonyms, etc. as translation strategy for English substitutions.

Searching for deletions, we query for instance noun phrases without nouns or coordinated or subordinated clauses without subjects (using the database described in Section 2.2). Surprisingly, the fact that we found more deletions in the German translations than in the English originals refutes Hawkins assumption. The following examples are taken from the sub-corpus of corporate communication:

(14) After the interviews, I told our employees that I wanted Baker Hughes to improve from being a good company to become a great one. – Nach den Gesprächen sagte ich den Mitarbeitern, dass ich Baker Hughes von einer guten Firma zu einer erstklassigen machen wolle.

(15) We want to thank shareholders for your confidence, and we will continue to do everything possible to reward that confidence. – Wir möchten den Aktionären für das uns entgegengebrachte Vertrauen danken und werden weiterhin alles Erdenkliche tun, dieses Vertrauen zu belohnen.

(16) Today, integrated functional departments, and shared ideas and technologies, are significantly improving everything we make, the way we do business, and the way we serve our customers – as this report shows. – Heute verbessern integrierte Bereiche und der Austausch von Ideen sowie Technologien nicht nur unsere Produkte, sondern auch die Art, wie wir unsere Geschäfte führen und unseren Kunden dienen.

Example (14) shows a German nominal phrase where the nominal head is deleted. In the English original a substitution is used to express the same meaning. Since substitution does not work for German (see above), deletion seems to be another strategy to translate this structure. In example (15), we find two German verbs (*danken* (*thank*) and *tun* (*do*)), the second subject is, however, deleted. In the English original the subject *we* is repeated for the second verb. The same phenomenon can be observed in example (16): The English original repeats the words *the way we*, which are deleted in the German translation. In both examples, German is more elliptical expressing the cohesive links implicitly, whereas English uses repetitions expressing the lexical cohesion more explicitly. These concordances show that the English-German corpus can also be used inversely helping translators to find compensations for German constructions which do not exist in English.

4. Concluding remarks and outlook

The need for linguistically annotated corpora is observed across all branches of linguistics, and the translation branch is no exception. There are certainly research questions which do not require such a detailed linguistic analysis, but which can be resolved using unannotated data or automatic annotation without the burden of constructing a big and complex language resource in advance. However, translation problems resulting from the specificity of a language or a register need a more detailed linguistic analysis in order to be answered.

In this chapter, we have suggested that monolingual and multilingual treebanks can assume the role of grammatical reference works for translation training and practice. In order for corpora to serve this purpose, they need to be enriched with linguistic information (Section 2). We have discussed the notions of monolingual and parallel treebanking and introduced some of the most important treebanks. In Section 3, we have shown that linguistic annotation can make a corpus a valuable resource for dealing with some typical translation problems. Offering the possibility to search for grammatical constructions (such as word order variation, clefting, raising constructions as well as substitutions and deletions) these treebanks are a much more powerful resource compared to parallel corpora of raw texts. And contrasting them to usual (printed) reference grammars, the concordances generated on the basis of the treebanks are far more comprehensive and exhaustive. Thus, the CroCo Corpus as well as the Penn and TiGer Treebanks prove to provide a wealth of information for the translation of typological language problems.

Besides typologically motivated translation solutions, also examples for register-specific language use and their appropriate translations (i.e., register-specific

translation behaviour) can be found in the treebanks. This means, however, that the corpus design should be representative in terms of size and relevant in terms of registers.

For dealing with more complex kinds of translation problems, a translation corpus should be annotated with more abstract linguistic information, e.g., semantic and discourse information. This requires more comprehensive annotation methods and more sophisticated query facilities – both of which are current research issues in computational linguistics. Future work also involves the applicability of treebanks as grammatical reference works to other language pairs.

References

Abeillé, A., Clément, L., & Kinyon, A. (2000) Building a treebank for French. In *Proceedings of the LREC 2000,* Athens, Greece, 87–94. Paris: ELRA/ELDA.

Bosco, C., Lombardo, V., Vassallo, D., & Lesmo, L. (2000). Building a treebank for Italian: A datadriven annotation schema. In *Proceedings of the LREC 2000,* Athens, Greece, 87–94. Paris: ELRA/ELDA.

Brants S., S. Dipper, P. Eisenberg, S. Hansen-Schirra, E.König, W. Lezius, C. Rohrer, G. Smith & H. Uszkoreit (2003) TIGER: Linguistic Interpretation of a German Corpus. In E. Hinrichs & K. Simov (eds.) *Journal of Language and Computation (JLAC),* Special Issue.

Brants, T. (1999) *Tagging and parsing with Cascaded Markov Models – automation of corpus annotation.* Saarbrücken, Germany: German Research Centre for Artificial Intelligence & Saarland University: Saarbrücken Dissertations in Computational Linguistics and Language Technology (vol. 6).

Brants, T. (2000) TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000.* Seattle, WA.

Brants, T. & O. Plaehn (2000) Interactive Corpus Annotation. In *Proceedings of the LREC 2000,* Athens, Greece, 87–94. Paris: ELRA/ELDA.

Christ, O. (1994) A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX 94, 3rd Conference on Computational Lexicography and Text research*. Budapest: 23–32.

Cuřín, J., M. Čmejrek, J. Havelka & V. Kuboň (2004) Building Parallel Bilingual Syntactically Annotated Corpus. In *Proceedings of The First International Joint Conference on Natural Language Processing*, Hainan Island, China, 141–146.

Greenbaum, S. (ed.), (1996) *Comparing English worldwide: The International Corpus of English.* Oxford, UK: Clarendon Press.

Hajic, J. (1999) Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova (ed.) *Issues of valency and meaning. Studies in honour of Jarmila Panevova.* Prague, Czech Republic: Charles University Press.

Hansen-Schirra, S., S. Neumann & M. Vela (2006) Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. In *Proceedings of the 5th Workshop on NLP and XML* (NLPXML-2006) at EACL 2006, Trento, 4th April 2006, 35–42.

Hawkins, J. A. (1986) *A comparative typology of English and German*. London: Croom Helm.

Heyn, M. (1996) Integrating machine translation into translation memory systems. *European Association for Machine Translation – Workshop Proceedings*. Geneva: ISSCO, University of Geneva, 111–123.

Hinrichs, E. W., J. Bartels, Y. Kawata, V. Kordoni & H. Telljohann (2000) The VERBMOBIL Treebanks. In Zühlke, W. and E. G. Schukat-Talamazzini, (eds.), *Proceedings of KONVENS 2000 Sprachkommunikation*, ITG-Fachbericht 161, 107–112. VDE Verlag.

Johansson, S. (1998) On the role of corpora in cross-linguistic research. In Johansson, S. and S. Oksefjll (eds.) *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 3–24.

Kuhn, J. and M. Jellinghaus (2006) Multilingual parallel treebanking: A lean and flexible approach. *Proceedings of the LREC 2006*. Genoa, Italy. Paris: ELRA/ELDA.

Leech, G. (1992) The Lancaster Parsed Corpus. In *ICAME Journal*, 16 (124).

Lezius, W. (2002) *Ein Werkzeug zur Suche auf syntaktisch annotierten Textkorpora.* IMS, University of Stuttgart. PhD thesis.

Maas, H. D. (1998) Multilinguale Textverarbeitung mit MPRO. *Europäische Kommunikations-kybernetik heute und morgen* '98, Paderborn.

Marcus, M., G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Gerguson, K. Katz, & B. Schasberger (1994) The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop.* San Francisco, CA: Morgan Kaufman.

Moreno, A., R. Grishman, S. López, F. Sánchez & S. Sekine (2000) A treebank of Spanish and its application to parsing. In *Proceedings of the LREC 2000*. Athens, Greece. 107–112. Paris: ELRA/ELDA.

Müller, C. and M. Strube (2003) Multi-Level Annotation in MMAX. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan. 198–107.

Och, F.J. and H. Ney (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29 (1), 19–51.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*. London: Routledge.

Sampson, G. (1995) *English for the computer. The SUSANNE corpus and analytic scheme.* Oxford, UK: Clarendon Press.

Skut, W., B. Krenn, T. Brants, & H. Uszkoreit (1997) An annotation scheme for free word order languages. In *Proceedings of ANLP-97*. Washington, D.C.

Telljohann, H., E. Hinrichs, S. Kübler & Heike Zinsmeister (2006) *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Technical report. Dept. of Linguistics, University of Tübingen, Tübingen, July 2006.

Volk, M., S. Gustafson-Capková, J. Lundborg, T. Marek, Y. Samuelsson & F. Tidström (2006) XML-based Phrase Alignment in Parallel Treebanks. In *Proceedings of the 5th Workshop on NLP and XML* (NLPXML-2006) at EACL 2006, Trento, 4th April 2006, 93–96.

CHAPTER 3

Corpora for translator education and translation practice

Silvia Bernardini and Sara Castagnoli
University of Bologna at Forlì, Italy

This article reviews the role currently played by corpora in translation teaching and practice. With regard to the former, classroom experiences involving corpus-informed approaches to translation teaching are discussed, and it is argued that such approaches should adopt an *educational* rather than a *training* attitude, giving more weight to awareness-raising uses of corpora, along with their obvious documentation roles. Examples of introductory e-learning materials about corpus use are presented which are addressed to students and professionals and which take an education-oriented view of translation teaching. With regard to the related issue of corpora in translation practice, the article presents the results of a survey that aimed to find out whether professional translators use corpora or at least know what they are. On the basis of the respondents' replies, it argues that a more widespread use of these resources is likely to depend on the availability of fast and user-friendly tools for constructing and consulting corpora, and describes some available tools that address this need.

1. Corpora and translation

Translation is in many senses an ideal field for corpus applications. One can think of few ways in which the isolation of stylistic traits and idiosyncrasies and the identification of *register* and *genre* conventions (Trosborg 1997) can be made easier than by looking at a source text against specialised and reference corpora. The browsing of target language corpora both prior to and during the production of a target text can reduce the amount of unwanted "shining through" (Teich 2003) of the source language (SL) into the target text (TT), by providing the translator with an inventory of attested "units of meaning", i.e., conventional ways of expressing specific meanings and performing specific functions in the relevant text type/variety within the target language (TL) (Tognini-Bonelli 2001: 131). Table 1 shows a simple example of the kinds of insights one can gain in this way. Given a turn of phrase typical of the wine tasting genre in Italian (*il vanigliato del legno*), a translator with a specialised corpus for the target language at her disposal can

Table 1. Snippets from a search for *vanilla* in a Web-derived bilingual comparable corpus on wine tasting (results from the English sub-component)

Original Italian	. . . **avere il sopravvento sul *vanigliato* del legno**
Original English	. . . *vanilla* and oak layers...
	. . . *vanilla* and subtle oak undertones...
	. . . *vanilla* characteristics. . .
	... oak *vanilla* nuances in dry wine...
	. . . subtle *vanilla* oak hints...
	. . . a suggestion of toasty *vanilla* oak...
	. . . hint of *vanilla* oak...
	... with *vanilla*, oak and apple notes...
	... oak barrels, it may pick up *vanilla* overtones...

extract and evaluate several likely translation candidates. In this case, the results of a simple search for *vanilla* in a small automatically-constructed corpus of web pages are presented. These provide supporting evidence for the translation of *legno* (lit. *wood*) as *oak*; they also suggest that the term *vanigliato* can be rendered as (*vanilla*) *notes*, *nuances*, or *hints*, among other possibilities. Clearly, corpus use can be particularly empowering for translators working into their L2, or tackling a specialised field they are unacquainted with. After all, and technological aids apart, these facts are not new to translators, for whom it is standard practice to rely on so-called "parallel texts", i.e., on the paper counter-part of *comparable corpora* of texts in the source and target language, matched by genre and subject matter to each other and to the text to be translated.

The last decade has seen a growing interest in the uses of corpora in translator education. Classroom experiences have shown that parallel corpora (of originals and their translations) can raise the students' awareness of professional translator strategies (Pearson 2003), that comparable corpora can help them produce more naturally-sounding translations (Zanettin 2001), and that constructing corpora can itself be a learning activity, whereby students learn to reflect on texts while acquiring the practical skills they need to build their own documentation resources (Varantola 2003). Several works have appeared that aim to provide students and professionals with practical and accessible introductions to (aspects of) corpus use. Bowker and Pearson (2002), for instance, is a book-length manual that walks the reader through the steps of building an LSP corpus, annotating it, consulting it, and applying it to different tasks (term and definition extraction, writing, translating and so forth).

If corpora are to play a role in the translation professions of tomorrow, it is important that they impact on the education of the students of today. The body of work just mentioned testifies that this is to some extent happening. However, substantial efforts still have to be put into place to make sure the majority of

translation students and teachers realise the potential of corpus work for translator education. Section 2 argues in favour of an approach that views corpus-based translation teaching and learning as *education* rather than *training*, and Section 4 provides examples of e-learning materials following the principles of such an approach. As we shall see (Section 3), professionals still appear to be largely unaware of (or unable to work with) corpora. Developing appropriate learning materials can help fill this gap. Yet in the long run widespread use of corpora in the translation profession is bound to also depend on the availability of user-friendly tools for building and consulting corpora quickly and efficiently. These are discussed in Section 5; Section 6 concludes by making suggestions about how best to tackle the challenges lying ahead.

2. Educating educators

It is common practice to speak of the instruction of future translators as "translator training". The term "training" implies that the abilities and competences to be learned are acquirable through practice with the kinds of tools and tasks one will be faced with during one's future professional career, in an environment that reproduces as closely as possible the future work environment. Widdowson (1984) contrasts the *training* framework, in which learners are prepared to solve problems that can be identified in advance through the application of pre-set or "acquired" procedures, with the *education* framework, whose aim is to develop the ability to employ available knowledge to solve new problems, and to gain new knowledge as the need arises. According to Widdowson, LSP teaching would be an example of a training setting, while general language teaching would be an example of an educational setting.

We may wonder whether translator education is in fact closer to the training or to the education end of the cline. Gouadec (2002: 32) explicitly champions the former position:

> [W]e are supposed to train people to perform clearly identified functions in clearly identified environments where they will be using clearly identified tools and "systems". [...] No serious translator training programme can be dreamt of unless the training environment emulates the work station of professional translators. [...] [T]he curriculum should [...] concentrate on emulating the actual work conditions of language services providers.

These views are certainly not unusual, and indeed are rather popular with students and prospective employers, who often lament a limited role of technology in translator education. While one is obviously sympathetic to the general issue

of technology in the translation classroom, it would be dangerous to carry these views to their extreme consequences, for two main reasons.

First, if translation skills are best taught by simulating actual work conditions, we should abandon the idea of education for translators and turn to apprenticeship instead: a professional environment should arguably provide a more appropriate setting for the simulation of actual work conditions than an academic one. Second, and more importantly, actual work conditions – and time pressure in particular – require that translator's strategies have become proceduralised, as is the case with mature professionals. Jääskeläinen (1997) finds that semi-professionals (translator trainees) show more extensive processing than both non-professionals and professionals (see also Englund Dimitrova 2005). She suggests that this may be because they are aware of the problems involved but have not yet automatised the necessary problem-solving strategies. Automatic processes are typically very efficient but rather rigid, such that there is the danger, pointed out, e.g., by Wills (1994: 144), "of problems being forced into a certain structure, because it is believed to offer a solution". In an education setting, students are still to develop the strategies that will then become proceduralised. Forcing them to work under realistic time constraints as would happen in a simulation activity could therefore work *against* the development of professionalism.

Translation instruction viewed as education, on the other hand, would make time for just the kind of activities and reflections that future professional translators will not have time for. A challenging aspect that is often neglected is how we can teach our students to identify problems in the first place. Going back to Gouadec (2002: 33), he claims that professional translators should possess, among others, the following skills:

1. Fully understand material to be translated
2. Detect, interpret and cope with cultural gaps [...]
3. Transfer information, facts, concepts [...]
4. Write and rewrite
5. Proofread
6. Control and assess quality

These skills translate into know-how; translators should know how to:

– Get the information and knowledge required
– Find the terminology
– Find the phraseology
– Translate
– Proofread
– Rewrite
– Manage their task(s)
– Manage a project (and other people)

Table 2. Titles and senses: lexicalised phrases and wordplay in the *Time Out Barcelona Guide* (Penguin)

Title	*Topic*	Senses
Get into the habit	Montserrat Monastery	1. in the habit of doing something; having a habit [...] of so doing. So *to [...] get into the habit* (OED) 2. *the habit*, monastic order or profession
Getting high	*Castells* (human towers)	1. situated far above the ground (OED) 2. *high*: under the influence of drugs or alcohol (OED)
Death on the mountain	Montjuïc (site of executions)	1. end of one's life on a mountain area 2. usually refers to climbers' accidental deaths; also a Japanese movie
On the tiles	The work of famous Catalan Architect J. M. Jujol	1. *on the tiles*: on a spree, on a debauch (OED) 2. *Josep Maria Jujol*: Catalan architect, his activity ranged from furniture designs and painting, to architecture (wikipedia)

Comparing the two lists, one notices that neither item 1 nor item 2 in the first (the "skills" list) translate into any of the know-hows in the second. In other words, there is a gap between "fully understand the material/detect any gaps etc." and "getting the information and knowledge required".

While illustrating this point with sufficient detail would take more space than is available here, a simple example can be provided. The phrases in the first column of Table 2 are taken from the *Time Out Barcelona Guide* (2002, Penguin). They are all titles of short sections devoted to different events or places, they all involve wordplay and require, to "fully understand the material to be translated", an understanding of the relationship between the facts being recounted or places being described and the standard sense of the lexicalised expression used. While the text itself no doubt provides hints for a correct interpretation of the allusions to places and events in and around Barcelona, for the wordplay to be successful the reader also has to know that the expressions used are not creative coinages but lexicalised phrases (their availability to the reader out of context is in fact a precondition for the success of the wordplay). A student who is not aware of these layers of meaning may be misled into taking such expressions as "on the tiles" and "getting high" at face value only.

While it is easy to find out about these expressions, i.e., "get the information and knowledge required" with the resources currently available to any translator, the real and often underestimated challenge lies in teaching students to identify wordplay or other types of "layered" meaning in the first place. By drawing their

attention to regularities in language performance as displayed in corpora, and making them reflect on the implications of (un)conventional usages, corpus-based activities such as those described in Sinclair (2003), Stubbs (2001) and Hoey (2005), especially if set within a translation-relevant framework, could help to fill this gap in translation pedagogy.

The didactic materials described in Section 4 below attempt to do just this, setting the treatment of the more obvious aspects of corpus work (producing and interpreting concordances, sorting results, inferring semantic preferences and prosodies from collocate sets, building small corpora etc.) within reflective activities that require students and professionals to keep (or start) asking questions about aspects of language and text that corpora can bring to light. Indeed, successful corpus work requires first and foremost an inquisitive frame of mind, a critical attitude and an ability to detect patterns, and only secondarily (some) technical skills. Yielding to the temptation of teaching the latter without providing practice to develop the former – *training* (future) translators to work with corpora, rather than *educating* them – would be a mistake.

3. Informing professionals

While sensitising students and instructors is of great importance for reaching the professionals of tomorrow, we should not forget the professionals of today. Reading about translation aids, one seldom finds references to corpora and concordancing tools. The impression that corpus use is the exception rather than the rule is confirmed by surveys attempting to find out whether professional translators are aware of the existence of corpora, and to what extent they use them in their work.

Surveying the Canadian market, Bowker (2004) finds that professional associations are aware of the existence of corpora but are generally more interested in translation memory (TM) technology, and that job advertisements never mention corpora. She suggests several possible reasons why corpora and corpus analysis have not as yet received an enthusiastic welcome in the professional world. One of these is the fact that the design, compilation and exploitation of corpora can be very time-consuming while not providing a tangible immediate increase in productivity. The success of translation memories is instead partly explainable because both their creation and consultation require minimal effort.

A more thorough investigation of the perception professional translators have of corpora was conducted in the framework of the Leonardo-funded MeLLANGE project, as part of an attempt to define user needs for learning materials on trans-

lation technology.[1] The survey aimed to collect information about the use translators make of the Web, their attitude to e-learning, and their awareness and use of corpora. Following two rounds of submissions in 2005 and 2006 – both online and on paper, mostly through translators' associations and communities – 741 questionnaires were completed by professional translators from the UK (the majority), France, Germany and Italy.[2]

In order to clarify what was meant by "corpora" and maximise the chances of receiving relevant answers, the corpus section of the questionnaire (summarised in Table 3) started with a concise and simplified definition.[3] This turned out to be essential, considering that 42% of the respondents reported never having heard about corpora. Initial questions were meant to point out the difference between using "reference materials" and using "corpora". Out of the total professional respondents, 45.7% reported collecting reference materials, more than half of them specified that they collect texts in electronic format (69.1% of those who reported collecting materials) and 49.2% of the latter declared that they read these texts (rather than *searching through* them). A slightly smaller percentage of respondents (44.2%) reported consulting corpora in their translation practice, the majority using facilities in word processors (65.7%) to search through them, with only a minority using a concordancer (17.7%). While many translators are not acquainted with corpora, or report not having the time and skills to use them, there seems to be widespread interest in learning more about them: 79.7% of respondents would be interested in a service which provides domain specific corpora, 80.0% in a tool for extracting terms from corpora, and 83.7% in learning more about their potential (MeLLANGE 2006).

Two main conclusions can be drawn from these data. First, there is a need for tailor-made learning materials addressed to translation professionals, which highlight the value added of corpora with respect to other tools and resources, and which adopt a practical (and not uncritical) perspective. Section 4 takes up

1. More information about MeLLANGE as well as the full survey results are available from http://mellange.eila.univ-paris-diderot.fr/

2. Overall, 1,015 questionnaires were returned, including those filled in by students of translation in the same countries. Here only the professionals' questionnaires are discussed, though.

3. The definition provided was: "*Corpora* are collections of texts in electronic form, usually grouped according to topic or type – contract, business letter, etc. Corpora may contain original texts in one language, comparable originals in two languages (comparable corpora), or originals and their translations (parallel corpora). Translation Memories are a special kind of parallel corpora. Corpora may be large and general, or small and specialised. They are (usually) not read cover to cover, so to speak, but searched through software programs (usually called "concordancers" or "corpus analysers"). One can list all the words contained in the corpus and see their frequencies, search for a word or expression (in context) and find out its typical patterns".

Table 3. Corpus section of MeLLANGE questionnaire (closed questions)

Question	Answer (%)	
Do you collect domain specific texts?	54.3	No
	45.7	Yes
How do you collect them? (multiple choice allowed)	69.1	In electronic form
	30.9	On paper
How do you use them? (multiple choice allowed)	50.8	Search through with software
	49.2	Read them
Do you use 'corpora' in your translation practice?	55.8	No
	44.2	Yes
If yes, do you use. . .? (multiple choice allowed)	25.7	Corpora of the target language
	22.1	Corpora of the source language
	19.4	Parallel corpora
	16.5	Domain specific corpora
	13.5	Comparable corpora
	2.8	General language corpora
What do you use to search them? (multiple choice allowed)	65.7	Search facility in word processor
	17.7	Concordancer
	14.4	Other search tools (specify: Trados, Concordance in translation memory)
	1.4	UNIX utilities
If you do not use corpora, why? (multiple choice allowed)	42.0	Never heard about them
	19.7	I don't have time to build them
	17.1	I don't know how to use a concordancer
	9.2	I can't see any advantage over *Google*
	8.2	I can't see any advantage over translation memories
	3.7	Other (1 specified – not sure if it will work with Macintosh)
Would you be interested in a service which quickly provides domain- and language-specific corpora tailored to your needs?	79.7	Yes
	20.3	No
Would you be interested in a tool for extracting terms from a domain-specific corpus?	80.0	Yes
	20.0	No
Would you be interested in learning more about the potential that corpora offer?	83.7	Yes
	16.3	No

this issue. Second, for corpora to be successful with translation professionals their construction and use has to be made substantially easier and faster, and ideally integrated into the translation workflow. Section 5 discusses available resources and future prospects in this area.

Table 4. The MeLLANGE *Corpora for translation* course

Unit	Contents
1 Overview	Introduction: why use corpora? Corpus use to understand a source text Corpus use to explore a text type Corpus use to produce a translation Comparable corpora Parallel corpora
2 Consulting corpora	Basic Advanced
3 Building your own corpus	To learn more about a text topic To produce or revise a translation
4 Encoding corpora	Introduction Structural mark-up and meta-data Linguistic annotation (manual and automatic) Querying annotated corpora Alignment
5 Applications: Term extraction	Manual Automatic

4. Learning about corpora, learning with corpora

As part of the MeLLANGE project, e-learning materials were developed that aim to familiarise students and professionals with the use of corpora for translation.[4] The aim is to cater for a double audience: highly motivated professionals, who can take the courses as fully autonomous e-learning sessions, and students of translation at undergraduate and post-graduate level, for whom we can hypothesise some class contact and some form of blended learning. In the latter scenario, the materials provide a repertory of learning contents, as well as ideas, activities, tools and corpus resources that teachers can either pick and choose from, or take as a block, simply adapting and complementing them with, e.g., assignments, forum postings, one-to-one and class-wide feedback, as well as face-to-face monitoring sessions. The contents of the course in its self-standing mode are summarised in Table 4.

Without attempting a thorough analysis, three points should be made. First of all, the teaching units are task-based: they centre on translation problems that learners have to solve through "processes of inference, deduction, practical

4. Materials relating to TM technology were also developed, which are not discussed here. Further information is available from http://mellange.eila.univ-paris-diderot.fr/public_doc.en.shtml#courses

reasoning or a perception of relationships or patterns" (Prabhu 1987:46) in corpora. The starting point is always a genuine text. To introduce learners to the notion of semantic preference (Sinclair 2004), for instance, they are given a text extract in which the word *revocation* is used (playfully) in the expression *revocation of independence*. They are then asked to find out what words follow *revocation of* in the concordance provided, drawn from the British National Corpus, and see if these words share a semantic trait that allows one to group them into a single set. After analysing the concordance and making hypotheses, the learners are asked to check these against the authors' answers. Among the things that get revoked one finds *licences*, *authorisations*, *certificates*, *concessions*, *patents* and *permissions*. In other words, *revocation* is often found together with words referring to official permits of various kinds, though one also finds revocations of other legal decrees (i.e., *constitution*, *edict*, *laws*, *order*). The conclusion is that the things that are revoked are legal acts, and particularly concessions or permits. Due to the existence of this semantic preference, that is not adhered to in the text under analysis, one is led to interpret *independence* as being framed by the author as a concession – hence the playfulness. Awareness of this writing strategy, involving exploitation of an established phraseological regularity, is crucial for a translator; yet it is well-known that intuition is unreliable when it gets to guessing collocates or detecting pattern (ir)regularities, especially with L2 speakers.

Secondly, corpus evidence is constantly compared with information obtainable from resources that the learners are likely to be more familiar with (i.e., dictionaries and the Web). Learners have to be highly motivated to keep using (and building) corpora beyond the duration of a course. Corpus use is time-consuming and cognitively demanding, and time-pressed translators and students of translation will gladly take any available shortcuts, unless we convince them that the insights they gain from the corpus cannot be obtained from any other, more user-friendly resource. For instance, in the unit on comparable corpora, students are asked to evaluate a (wrong) translation from Italian into English, and to provide an explanation of the reason why the translator's choice was not appropriate. The sentence in question, taken from the website of an Italian wine producer, is:

> *Per cogliere appieno le qualità di questo vino si consiglia di servirlo a temperatura ambiente, stappando la bottiglia un'ora prima di mescere*

This was translated as:

> *Serve it at an ambient temperature; to fully enjoy its qualities uncork the bottle about an hour before serving*

Students are asked whether the expression *ambient temperature* is an appropriate translation for "temperatura ambiente". They are encouraged to look this up in monolingual and bilingual dictionaries and on the Web. In this way they assemble

some evidence pointing at the fact that *room temperature* would seem to be a better choice (but not enough to understand exactly the differences in meaning and usage between the two expressions). They are then given concordances for "temperatura ambiente", *room temperature* and *ambient temperature*, obtained from a specialised comparable corpus of English and Italian wine tasting texts. Since both English phrases are attested in the corpus, one wonders whether both are in fact correct, i.e., whether they are simply synonymous. From this point onwards, students are taken step by step through an analysis of corpus evidence that leads them to conclude that *ambient temperature* is the actual temperature of the surrounding environment, while *room temperature* refers conventionally to a temperature of about 18–20°C (regardless of the actual temperature of the room one is in).

Lastly, the potential for serendipitous learning activities is not neglected (Bernardini 2000). When working with parallel corpora, for instance, learners are asked to identify all the different ways in which the Italian adjective "propositivo" as in "avere un ruolo propositivo" can be translated into English, drawing evidence from dictionaries, the Web and a corpus of EU Parliament texts. While dictionaries and the Web prove of little help, the corpus suggests one principal equivalent (*proactive*), and several alternative paraphrases. In the course of this activity, the students' attention is drawn to two concordance lines in which the expression *(take the) offensive* is used. While for the purposes of the current problem these lines could simply be discarded, they can also provide the starting point for further, serendipitous corpus analyses. Starting from the following question:

> *Does the adjective* offensive *seem to collocate with words that can be considered similar to those which "propositivo" collocates with?*

Students are asked to compare concordances for *offensive/take the offensive* and for "propositivo" in reference corpora of English and Italian. They soon find that the two expressions differ with regard to their semantic prosodies ("propositivo" being used in essentially positive co-texts and *offensive* being used in negative co-texts) and preferences ("propositivo" collocating with "referendum" and with abstract words such as "capacità" (*capacity*) "ruolo" (*role*), "spirito" (*spirit*), "fase" (*phase*), and *offensive* collocating with *language* and *content* as well as military terms such as *weapon*, *lineman* and *operation*). The latter task is not directly relevant to the parallel corpus activity, yet it stimulates curiosity about language and about translator strategies, and reminds students that different types of corpora can be useful for different purposes, and that one often needs to combine them (i.e., using parallel corpora to develop hypotheses and comparable corpora to test them).

Summing up, the rationale behind these materials is that the technological aspects of translation work cannot be severed from its cognitive aspects. This is because, in an education-oriented framework, one does not simply teach *about* corpora, but rather teaches to *translate with* corpora.

5. Building corpora

In order for corpora to stably enter the translators' workflow, however, one cannot simply rely on well thought-out learning materials. It was previously suggested (Section 3) that corpus construction and use should also be made easier and faster, so that these tools can compete with others that translators use in their everyday activity, such as TMs and the Web. 94.6% of the MeLLANGE questionnaire respondents reported consulting the Web through Google despite several drawbacks that most of them are aware of, such as the unhelpfulness of the sort order (20.7%), the lack of linguistic information (14.7%), the unreliability of frequency statistics (12.6%), or the inadequacy of context display (9.9%). This suggests that corpora could indeed play a role among translation tools, if remaining obstacles (especially the time needed for construction and the required search skills) were removed.

5.1 The present ...

One of the major achievements in corpus-based language learning in the past decade has been the creation of tools that allow users to consult the Web in a more linguistically-informed way, and/or that facilitate the construction of corpora from the Web. While search engines such as Google provide fast and effective retrieval of information from the Web, they are less than ideal when it gets to basic linguistic procedures such as highlighting patterns (i.e., sorting results) or selecting subsets of solutions, not to mention conducting searches for linguistically-annotated sequences (e.g., all verb lemmas preceding a certain noun lemma) (Thelwall 2005).

A solution to some of these problems has been provided by tools like *KWiCFinder* (Fletcher 2004), an online concordancer that supports regular expressions, implements concordance-like displays and functionalities (e.g., sorting), and allows off-line perusal of the retrieved texts. Along similar lines, another freely available tool, the *TextSTAT* concordancer,[5] allows one to specify a URL and retrieve a file or set of files from a single website directly from within the concordancer, thus conflating and speeding up the processes of retrieving texts and searching through them. *Corporator* (Fairon 2006) only addresses the first issue (retrieving texts from the Web): it automates the process of corpus collection and development allowing bulk retrieval of selected websites that offer RSS feeds. The corpus thus created can then be updated regularly, and is searchable with a regular concordancer.

While *KWiCFinder* is designed mainly with language learning applications in mind (searching for a given word or expression as one would search the Web),

5. http://www.niederlandistik.fu-berlin.de/textstat/software-en.html

TextSTAT only offers basic web-search facilities (i.e., it does not interact with a search engine, but simply spiders a specified URL), and *Corporator* can only retrieve pages from websites that use RSS technology, the *BootCaT* toolkit (Baroni & Bernardini 2004) was created specifically for translation students and professionals, i.e., for users who need relatively large and varied corpora (typically of about 1–2 million words), and who are likely to search the corpus repeatedly for both form- and content-oriented information within a single extended task. Starting from a series of "seeds" (search words), this set of Perl scripts provide facilities for combining the seeds into sequences, submitting queries to a search engine, retrieving URLs (for manual inspection if necessary) and eliminating duplicates. Then for each URL the text is retrieved, cleaned, and printed to a text file. This procedure can be iterated if larger corpora are required, by selecting seeds for a second round of searches from the initial corpus and repeating the various steps. These tools have been used for several projects, including the construction of Web corpora for several languages (see Sharoff 2006 and Ueyama 2006). A JavaScript implementation of the *BootCaT* toolkit which runs under MS Windows has been developed to make the tool accessible also to people with average computer skills.[6] A Web version also exists (*WebBootCaT*, Baroni et al. 2006) whereby the whole collection procedure described above is carried out via a web interface, and the resulting corpus is either downloadable for off-line consultation or loaded into the *Sketch Engine*, an online corpus query tool (Kilgarriff et al. 2004).[7]

The comparable corpus of English and Italian texts on wine tasting mentioned in Section 1 – from which the results in Table 1 were derived – was collected with *BootCaT* and used in an English to Italian translation course at the School for Translators and Interpreters of the University of Bologna, Italy. The conventions of that genre both in English and in Italian were unknown to all the students in this course. A specialised comparable corpus is indispensable to (learn to) search for genre-restricted phraseology and terminology, two of the central know-hows identified by Gouadec (Section 2, above). Given the time constraints under which translators normally operate, mastering techniques for the quick-and-dirty construction of corpus resources could be an additional asset.

6. JBootCaT, http://www.andy-roberts.net/software/jbootcat/

7. At the moment, two different versions of BootCaT exist that submit queries to either Google or Yahoo. The freely available version, running either under UNIX or Windows (through JBootCaT, see footnote 6), interacts with Google and requires users to possess a Google Web API licence key. A commercial Yahoo-based version – for which users do not need to own any API key – is integrated into and can be accessed from within the Sketch Engine (http://www.sketchengine.co.uk/, free trial accounts available at the time of writing, January 2008).

5.2 ... and the future

While the new tools at our disposal make the construction of corpora from the Web easier, certain obstacles still have to be overcome. In the long term, widespread use of corpora and corpus construction and search facilities among translators is likely to depend on their integration with Computer-Aided Translation (CAT) technology. We could envisage a tool that interacted with a web search engine to search, retrieve and morphologically annotate corpora based on user specifications. It would support regular expressions and handle subcorpora, and would provide facilities for monolingual and parallel concordancing (including alignment). Such a tool would extend the productivity of CAT systems by allowing a double search mode: automatic search for full and fuzzy matches in gold standard TMs, and manual concordancing of comparable and parallel texts for hypothesis development and testing where the TM has nothing to contribute:

> [...] translators working with texts that contain a large number of repeated segments, such as revisions, will be well served by the segment processing approach. On the other hand, translators who hope to leverage or recycle information from previous translations that are from the same subject field, but that are not revisions, may find that the bilingual concordancing approach is more productive.
>
> (Bowker 2002: 124)

Such a system would also arguably limit some of the drawbacks associated with the use of TMs. For instance, it has been observed (e.g., by Kenny 1999 and Bowker 2002) that translators using CAT software may develop a tendency to make their texts more easily recyclable within a TM, regardless of the translation brief, and that they may be led to lose sight of the notion of "text" as a consequence of a rigid subdivision into units. The possibility to search whole texts (rather than translation units) using a concordancer could positively impact on these strategies and attitudes.

While no tool currently combines all these functionalities, some form of integration seems to be underway, thanks to tools such as *MultiTrans*, a CAT package which allows one to search for strings of any length (i.e., not limited to the size of a translation unit), and, if required, displays them in full-text context. Interestingly, while the company producing this software is called *MultiCorpora*, no further mention of the words *corpus* and *corpora* can be found on the *MultiTrans* page:[8] a further proof that these are currently not buzzwords in the translation market?

8. http://www.multicorpora.ca/

6. Summing up: The future of corpora in translation

Despite achievements and enthusiasm within academic settings, corpora are still to make an impact especially on the translation profession. A number of reasons why this might be the case have been suggested, and several challenges have been identified.

There seem to be three main areas where efforts should be concentrated. First, the role of corpus work for awareness-raising purposes should be emphasised over the more obvious documentation role, and basic "translation" skills (whose development should be pursued *also* through corpus use) should be restored to their central place in translator education:

> [...] the general abilities to be taught at school [...] are the abilities which take a long time to learn: text interpretation, composition of a coherent, readable and audience-tailored draft translation, research and checking, correcting. [...] If you cannot translate with pencil and paper, then you can't translate with the latest information technology. (Mossop 1999)

Second, translator-oriented (e-)learning materials have to be provided, so as to reach those professionals who are eager to learn about/with corpora. These materials should ideally be contrastive in focus (i.e., why/when use corpora instead of the Web/TMs/dictionaries?), and include substantial practice primarily with those tools and facilities that translators (rather than linguists or language learners) are likely to find of immediate relevance (e.g., concordancing should arguably be given priority over frequency word-listing). Such practice should be embedded in translation-relevant tasks and should not neglect *serendipitous* turns encouraging the exploration of language and translation issues. Finally, corpus construction and corpus searching should be made faster and more user-friendly, and ideally integrated with CAT tools, so as to reach the largest possible number of professionals, including the less technologically enthusiastic.

References

Baroni, M. & S. Bernardini (2004) BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC 2004*. Paris: ELRA/ELDA. 1313–1316.

Baroni, M., A. Kilgarriff, J. Pomikálek, & P. Rychlý (2006) WebBootCaT: Instant Domain-specific Corpora to Support Human Translators. In *Proceedings of EAMT 2006*. Oslo. 247–252.

Bernardini, S. (2000) Systematising Serendipity: Proposals for Concordancing Large Corpora with Language Learners. In Burnard, L. & T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang. 225–234.

Bowker, L. (2004) Corpus Resources for Translators: Academic Luxury or Professional Necessity? In Tagnin, S. (ed.) *Corpora and Translation, TradTerm 10,* Special Issue. 213–247.
Bowker, L. (2002) *Computer-aided Translation Technology.* Ottawa: University of Ottawa Press.
Bowker, L. & J. Pearson (2002) *Working with Specialised Language. A Practical Guide to Using Corpora*. London: Routledge.
Englund Dimitrova, B. (2005) *Expertise and Explicitation in the Translation Process*. Amsterdam/Philadelphia: John Benjamins.
Fairon, C. (2006) Corporator: A Tool for Creating RSS-based Specialised Corpora. *Proceedings of the 2nd International Workshop on Web as corpus. EACL 2006*. Trento. 43–50.
Fletcher, W. (2004) Facilitating the Compilation and Dissemination of Ad-hoc Web Corpora. In Aston, G., S. Bernardini & D. Stewart (eds.) *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins. 273–300.
Gouadec, D. (2002) Training Translators: Certainties, Uncertainties, Dilemmas. In Maia, B., J. Haller & M. Ulrych (eds.), *Training the Language Services Provider for the New Millennium*. Oporto: Universidade do Porto. 31–41.
Hoey, M. (2005) *Lexical Priming*. London: Routledge.
Jääskeläinen, R. (1997) *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating.* Joensuu: University of Joensuu – PhD thesis.
Kenny, D. (1999) CAT Tools in an Academic Environment. In *Target*, 11(1), 65–82.
Kilgarriff, A., P. Rychlý, P. Smrz, & D. Tugwell (2004) The Sketch Engine. In *Proceedings of Euralex*. Lorient, France. 105–116.
MeLLANGE (2006) *Corpora and E-learning Questionnaire. Results Summary – Professional.* Internal document, 12th June 2006.
Mossop, B. (1999) What Should be Taught at Translation School? In Pym, A., C. Fallada, J. Ramón Biau & J. Orenstein (eds.) *Innovation and E-Learning in Translator Training*. Available online: http://isg.urv.es/library/papers/innovation_book.pdf. 20–22.
Pearson, J. (2003) Using Parallel Texts in the Translator Training Environment. In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translator Education*. Manchester: St Jerome. 15–24.
Prabhu, N. S. (1987) *Second Language Pedagogy*. Oxford: Oxford University Press.
Sharoff, S. (2006) Creating General-purpose Corpora using Automated Search Engine Queries. In Baroni, M. & S. Bernardini (eds.) *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit. 63–98.
Sinclair, J. McH. (2003) *Reading Concordances*. London: Longman.
Sinclair, J. McH. (2004) *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
Stubbs, M. (2001) *Words and Phrases*. London: Blackwell.
Teich, E. (2003) *Cross-linguistic Variation in System and Text.* Berlin: Mouton.
Thelwall, M. (2005) Creating and Using Web Corpora. In *International Journal of Corpus Linguistics*, 10 (4), 517–541.
Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.
Trosborg, A. (1997) Text Typology: Register, Genre and Text Type. In Trosborg, A. (ed.) *Text Typology and Translation*. Amsterdam/Philadelphia: John Benjamins. 3–23.
Ueyama, M. (2006) Evaluation of Japanese Web-based Reference Corpora. In Baroni, M. & S. Bernardini (eds.) *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit. 99–126.

Varantola, K. (2003) Translators and Disposable Corpora. In Zanettin, F., S. Bernardini & D. Stewart (eds.) *Corpora in Translator Education*. Manchester: St Jerome Publishing, 55–70.

Widdowson, H. G. (1984) English in Training and Education. In *Explorations in Applied Linguistics II*, Oxford: Oxford University Press. 201–212.

Wills, W. (1994) A Framework for Decision-making in Translation. In *Target*, 6(2). 131–150.

Zanettin, F. (2001) Swimming in Words. In Aston, G. (ed.) *Learning with Corpora*. Houston, TX: Athelstan. 177–197.

CHAPTER 4

CORPÓGRAFO V.4

Tools for educating translators

Belinda Maia
University of Porto and Linguateca, Portugal

It is clearly essential for future translators to learn to use the available translation technology but, given the many linguistic skills translators also need to acquire, the process needs to focus not just on the ability to use the technology, but also on encouraging a good understanding of the objectives, possibilities and limitations of the technology itself. Skilful use of the technology will come later after practice in the professional contexts in which it is needed. An understanding of the problems posed by integrating technology into both the curriculum and teaching practice led us to develop the Corpógrafo at the PoloCLUP of Linguateca[1] at the University of Porto. It allows for the building and analysis of parallel and comparable corpora, extraction and management of terminology, as well as the collection and analysis of lexical items, particularly multi-word expressions that are relevant to text, genre or discourse analysis. It is a research environment for autonomous study, but it also offers various possibilities for education in translation, text analysis and terminology management, and has the advantage over commercial software of being freely available online.

1. Introduction

The LETRAC project (1998–9)[2] was perhaps the most formal of various initiatives to drag translator education into the era of information technology and, with

1. Linguateca is a distributed language resource centre for Portuguese, with the mission to raise the quality of Portuguese language processing through the removal of difficulties for the researchers and developers involved. The leader of the project is Diana Santos and for some years she and her colleagues have provided plenty of online language resources, including a 180 million word newspaper corpus, CETEMPúblico, several smaller corpora, some of them annotated, and a parallel Portuguese/ English corpus at present containing nearly 1.5 million words in each language, COMPARA.

2. LETRAC – Language Engineering for Translation Curricula. http://www.iai.uni-sb.de/iaien/en/letrac.htm

hindsight, one can now recognise the fact that it was probably more focused on the language engineering of its title than the realities of translation curricula. It was necessary then to draw attention to the need for translators to be trained in general computer skills as well as the use of the various types of translation software and language tools, not to mention the almost infinite possibilities of obtaining information from the Internet. Today students usually arrive at university with general computer skills and are accustomed to using the Internet on a regular basis, so teaching them and how to use the translation software presents fewer problems than 10 years ago.

Technology such as translation memories, localisation tools and sub-titling programmes is proving to be a double-edged sword for translators and the institutions that educate them. The technology itself is only useful for certain types of translation, and what it offers in help, it takes away in terms of remuneration as translators are forced to negotiate new terms of work with clients who invest in translation memories in order to save on translators. Localisation programmes have simplified the translators' task so much that training students to use them now is trivial in comparison to the situation some years back. Sub-titling programmes, too, are becoming increasingly sophisticated and expensive, and, at least in Portugal, translators find it difficult to earn a living in sub-titling.

The technology is being constantly improved and, despite the fact that some companies who produce it now offer special rates to the universities, the need to constantly invest in upgrades for software that has a limited usage in the university context makes it almost impossible for many institutions to keep up with developments. There are also logistic problems in allowing for practice and project work because of the need to restrict the use of commercial software to the university campus and/or specific computer rooms. What we shall describe here is a response to this situation.

The Corpógrafo developed out of a theoretical background of corpus linguistics and the way in which this was applied to the teaching of translation and terminology. In the early stages, the Corpógrafo was primarily for terminology research, and the computer engineers involved were encouraged by the realisation that terminology is also crucial to machine and machine assisted translation, information retrieval and knowledge management. It was not prepared for any particular domain, but rather designed to adapt to any type of terminology research undertaken by individuals or small groups.

The corpora building aspect of the Corpógrafo led us to favour comparable over parallel corpora and to look for quality rather than quantity in the texts chosen. In this respect, the corpus approach was supported by a theoretical background in systemic-functional linguistics rather than the natural language processing interests of the Linguateca project. The emphasis on text analysis and classification required by the need to select for quality rather than quantity later led

to the adaptation of the existing tools for terminology work to other forms of study and research, such as more general lexical, syntactic and text analysis. For this purpose new tools have been developed and, although some imitate commercial translation software in certain respects, the objectives are pedagogical.

2. Parallel and comparable corpora – uses and limitations

Plenty of research projects and many hopes are based on the assumption that parallel corpora (and translation memories), originals aligned with their translations, must provide an ideal way of improving human and machine translation as well as term extraction. Contributors to Veronis (ed: 2000) explore the various possibilities of aligning texts as well as some applications of this methodology, with Blank's article on term extraction and Gaussier et al's one on machine-aided human translation. One has only to look at the proceedings and programmes of conferences like LREC and COLING to see that there are several projects in this area, as in a workshop at the LREC 2004 conference on 'the Amazing Utility of Parallel and Comparable Corpora', and in a special issue on the subject in the Journal of Natural Language Engineering in 2005 (Volume 11, Issue 3). Most of this research refers either to the more linguistic aspects of machine translation, or to automatic term extraction for information retrieval.

Although the judgment of what constitutes a 'good' original or a 'good' translation may appear to be subjective, in practice most teachers and professional translators recognise that it is perfectly possible to evaluate both texts qualitatively. To be of any use for research, parallel corpora must consist of good originals and good translations, must be available, and the research results must justify the effort. Although some institutions, like the European Commission, do supply some of their parallel corpora for research, and commercial companies, like Xerox, contract universities to work with their material to produce tools or terminology for their own use, the use of these parallel corpora for research is limited. Professional translation memories, which are equivalent in structure to parallel corpora but differ in function, are usually the property of a company or of a translation agency and the owners are either reluctant to share their property, or unable to do so for copyright and/or security reasons.

In nearly all other situations, the search for good parallel corpora is unrewarding, particularly if one is working with two languages that are unevenly represented in terms of documentation, as with English and Portuguese. Websites that claim to be bilingual usually turn out to be only partly, or badly, translated, or one finds that the translator has adapted the text or summarised the content for the other languages, or that one side of the website is outdated. Some sites have even

been translated by machine translation programmes and are hardly a resource for translation or terminology work.

Besides these problems, many people, and particularly experts writing articles at the cutting-edge of their research, increasingly want and need to be read in English and the original in their own language is either unavailable or non-existent, since so many write directly in English. Textbooks, usually rich in terminology, are not translated from the original English in which they were written, with the excuse that students should be encouraged to read and write English.

Comparable corpora, or texts in the same domain or in a similar genre, have several advantages over parallel ones. They are more available, and there is a greater chance that the text will probably be representative of the conventions of style and register that are acceptable in the local context. Also, if they are by experts writing in their own language, the terminology will be more acceptable than that chosen by a translator. Even so, it is usually difficult to balance the corpora in both or all the languages involved, especially if one of them is English.

A point often made when building specialised corpora is that collections of 'raw' text are not 'corpora', i.e., that for collections of text to become corpora they must be properly annotated for parts-of-speech, syntax or other purposes. Experience of working with both types of expert tells us that there is a basic lack of agreement here between the linguists, who want to study language in its multiple aspects, and the computer engineers who want to accelerate the extraction of various types of information from enormous quantities of text for which it would be unreasonable – if not impossible – to demand annotation.

There are arguments for both sides, and one of them depends on the answer to the question 'what is the ideal size of a specialised corpus?' The linguist will probably argue that quality is more important than quantity, and that 100,000 words of well-selected texts can provide better information than a million words of poorly selected texts. The computer engineer will want to extract patterns of information from enormous quantities of raw text, in an attempt to create tools that solve the problems posed by the giant 'corpus' that is the Internet. In between there is the computational linguist trying to construct the treebanks (see, for example, Hansen-Schirra's Chapter Two) and electronic dictionaries that, from a longer term perspective, are supposed help everybody by accelerating the linguistic analysis of vast quantities of text.

3. Corpora and terminology research

Terminology management skills are increasingly necessary for professional translators and translation companies, yet courses in terminology have only been considered a necessary part of translation curricula relatively recently. While in the

past terminology work was often given to junior members of translation companies or trainees, this is no longer the case.

When developing a glossary of systemic functional terms, Mathiessen (1997) quoted Firth (1948) in *Lingua* as saying:

> . . . terminology is necessitated by a system of thought ... Questions of terminology inevitably arise when new systems of thought are applied to the handling of material or events. The whole conceptual framework, the whole syntax of thought and words, should hold together systematically.

Although this definition was not made within the context of mainstream terminology, or about terminology in general, it is appropriate for descriptive terminology work that takes text corpora as the ideal, if not only, source for information. Parallel corpora are not always the best source of terminology for reasons described above. As Bennison & Bowker (2000) say, "what translators [and terminologists] need are bilingual comparable corpora and tools for accessing the information within them". The computerisation of terminology research is less automatic with monolingual and comparable corpora, but the chances are that the results will be better.

We would argue that building and using specialised corpora in teaching practice is useful not just for terminology work, but also serves as a methodology for learning about any special domain, for discovering and examining the text genres related to it and, finally, for extracting specialised terminology not only appropriate to the domain, but also to the register and text genre. The obligation to learn about complex technical and scientific domains and to distinguish between different types of texts, far from alienating translation students educated in humanities faculties, has proved motivating and has allowed for postgraduate work with the formal expert cooperation of professors from other faculties and departments.

Texts are chosen for their 'term potential' and students are encouraged to start with texts from good encyclopedias, go on to introductory and then more advanced textbooks, and only later to the documents of experts. In this way they learn about the subject gradually, extract the more general terms that provide clues to further searches for more sophisticated texts, before progressing to the state-of-the-art texts and less comprehensible terminology of the domain expert.

Corpora are very useful language resources for descriptive terminology research, but they do not solve all the problems. Each domain presents problems – and solutions – of its own. It is no coincidence that people use particular corpora to examine certain phenomena. For example, Blank (2000) used parallel corpora for term extraction with documentation on patents; Estopà et al. (2000) concentrated on Greek and Latin compounds for term extraction, but used medical texts as their corpus; Demetriou & Gaizauskas (2000) used texts from biology for automatic term retrieval from untagged text; and Conceição (2001) chose to

study reformulations of terms with pharmaceutical terminology. In each case, the domain texts were particularly suitable for studying the phenomenon under observation. In terminology work, specialisation of this sort is, almost by definition, inevitable. Even a big project, like GENOMA-KB (Cabré et al. 2004), is restricted to the subject of the human genome.

There are clearly problems of availability of text because of copyright and the related fear of plagiarism. Yet, with so much text online, the use of texts for term extraction should not constitute too much of a problem, provided that the corpora are not made publicly available and the sources properly recognised. Authors are normally flattered to know that their work is being used as expert evidence, and we are often offered the use of considerable amounts of text. However, a clue to the probable reactions can be found in the different attitudes of scientific communities as to how widely they publish online. Commercial documents that are online are hardly likely to be trying to conceal industrial secrets, and the owners may even welcome free publicity.

3.1 Other sources of terminology

Documents that do not traditionally qualify as texts for corpora, such as glossaries with detailed definitions of terms, can also be used, provided that several can be found and enough comparisons can be drawn to overcome any accusations of plagiarism. Encyclopedia entries and introductory textbooks provide simple, explanatory definitions. Couto (2003) found that there was a demand by university professors for pedagogically orientated definitions, and helped produce an interactive pedagogical glossary on corrosion (http://paginas.fe.up.pt/~mcnunes/QAE/QAE_gloss_b.htm). One also cannot throw out the baby (of good terminology collected from experts, but without reference to corpora) with the bathwater (of out-of-date, or unused, terminology perpetuated by antiquated reference material).

4. Corpora for text analysis

Parallel corpora / translation memories are often used as a resource for both human and machine (assisted) translation. Human translators look upon them as reference material, translation software exists largely to present the translator with rapid ready-made 'solutions' from their translation memories, and machine translation (MT) experts work on the possibilities of improving MT using statistical evidence from large translation memories. However, most linguists realise that parallel corpora are often skewed by the translators' adherence to the lexicon and syntactic structure of the original. Another factor, less commonly acknowledged, is that different languages / cultures sometimes have different conventions for cre-

ating text and presenting information. Rui Silva (2006) extracted and analysed phrases used to connect discourse to show the different conventions of the languages/cultures of English and Portuguese in their approach to the task of writing about art exhibitions.

A simple n-gram tool used to analyse comparable texts drew our attention to the practical possibilities offered for the collection of general language multiword expressions that might be useful to both the writers of the originals and their translators. An example of this is a database of phrases for describing university course programmes in English and Portuguese we have created to help teachers (and translators) with the presentation of their programmes in both languages, as demanded by the regulations for universities with ERASMUS mobility. This methodology can also be applied to a variety of tasks, such as extracting English phrases from native speaker academic articles in order to help the many non-native speakers of English who are obliged by circumstances to write in English. This is not exactly a new idea, but it is easier to implement using corpora and an n-gram tool than by transcribing them manually and/or relying on intuition.

These are just a few ways in which this sort of research could be developed and the tools in the Corpógrafo are still in the early stages, as will be explained below.

5. The Corpógrafo

The first version of this integrated web-based suite of tools for corpus linguistics, and terminology and text research, the GC – Gestor de Corpora, later to be rebaptised as the Corpógrafo, was developed in early 2003. Since then it has been under constant development, and new ideas are explored and developed according to the personal research interests of the people working with it. This explains the organic, and sometimes uneven, way in which it has developed, but this interactive process has led to a flexible and dynamic structure, that is more economical than creating large rigid structures that possibly contain features that may, or may not, be useful. At the end of 2006, the structure, from the user's point of view, was best represented visually in Fig. 1.

As can be seen, the focus at this point was on terminology extraction and management. Since the middle of 2007, other tools have been added that are of use for more general language analysis, as will be described below.

5.1 An integrated environment for building and using individual corpora

The first, very important, move was to overcome the limitations and frustrations of using commercial software which obliged students to work on campus by creating a web-based environment accessible by each individual using a username

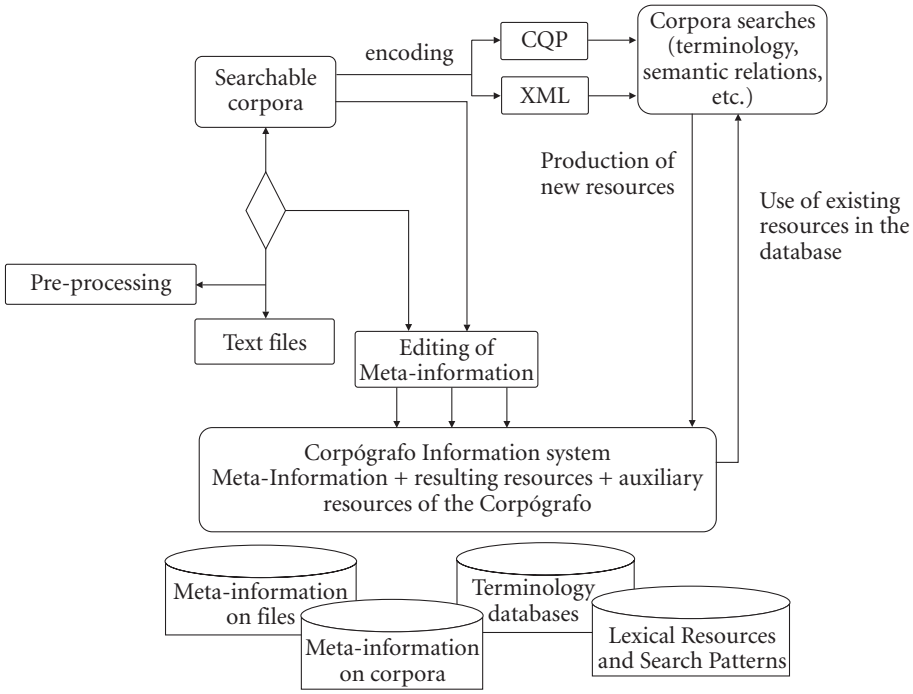

Figure 1. Structure of the Corpógrafo from the user's point of view in 2006.

and password. Each researcher or group is provided with a private space on the dedicated server on which to carry out their work. They create, analyse and experiment with their own corpora and databases and try out new ideas, but everything they do with these tools is saved on this private space. The administrator and the teacher or supervisor may use the student or researcher's username and password to enter that space, also over the Web, and provide help and advice when necessary. Otherwise, each project functions autonomously.

On acquiring a space on the server via free registration for a username and password, the user is presented with an empty 'space' in which to work, together with instructions for use. The Gestor (File Manager) allows one to:

– Import texts in various formats and upload them to the Corpógrafo;
– Register the metadata of the texts, i.e., document, authorship, source, domain and text type (this allows proper credit to be given for any information extracted, and serves as some protection from copyright problems);
– Preprocess texts by the removal of unwanted material and text correction;
– Automatically divide the text into sentence length units;
– Combine and re-combine texts to form corpora for specialised research (e.g., one can combine all the domain specific texts in one language to extract ter-

minology, or re-combine a selection of academic articles from several subject domains, in order to study the stylistic and syntactic aspects of the genre);
– Align parallel corpora;
– Store multimedia files, e.g., sound files for pronunciation, or images to relate to lexical or terminological items;
– Register the names of different users for group work.

The Pesquisa (Corpus analysis) area allows the use of the personally selected corpora to:

– Search for concordances using regular expressions;
– Search for concordances using the NooJ resources (in English, French and Portuguese – see http://www.nooj4nlp.net/).

The concordances can be viewed as whole sentences or as KWIC concordances of up to 15 words each side, including the usual left or right sorting functions.

5.2 Creating terminological and lexical / phrasal databases

The next section is called the Centro de Conhecimento (Knowledge centre) and now contains two types of database, BD Terminológicas (terminology databases) and BD Fraseológicas (phrasal databases) for the management of words and phrases. The databases have much in common, but certain aspects are specific to the analysis required.

In both cases, the first thing one must do is create a database and supply the necessary metadata. Several different databases can be created if one so wishes, but each database is designed to be multilingual so that links can be made between the data on presumed equivalents between languages.

The next step is to extract information from the corpora and enter it in the database. The information extracted from corpora automatically brings with it the metadata (authors and sources) of the texts in which it was found. The databases all allow for the insertion of information on morphology, definitions, contexts (examples taken from concordances), lexical or semantic relations, related terms or expressions, translation equivalents, and links to any relevant multimedia files.

The differences between the two types of database are in the methods of extracting terms and lexical expressions from the corpora and in ways of classifying the results.

The terminology databases allow one to:

– Extract terminological units using an n-gram tool with automatic lexical search restrictions for Portuguese (PT), English (EN), French (FR), Italian (IT) and Spanish (ES);

– Find definition candidates and semantic relations aided by lists of lexical /syntactic patterns (these tools function in several languages, but need further development).

The lexical / phrasal databases allow one to search for regular expressions and conduct searches using the NooJ part-of-speech (POS) annotated dictionaries. It is possible to mix lexical with POS data, and the objective in future is to teach people to create their own disambiguation grammars for their research using the NooJ system.[3]

At present the Corpógrafo provides the terminologist and the lexicographer with the possibility of establishing semantic relations between terms or lexical items in the database. The classical relations of synonmy / antonymy, hyperonymy / hyponymy, holonymy / meronymy, agent / instrument, agent / result, patient / instrument, cause / effect, and several others of the kind listed by Sager (1990: 34–5) are already offered as choices, as is the possibility of establishing a new relationship, giving it a name, an acronym, a description, an example and the registration of a reciprocal relationship, if appropriate. The objective here is to allow people to experiment with new semantic relations that may only be relevant to their own work.

Experiments have been made with the organisation and visualisation of semantic relations, as in Silva's work (2003) which explores various semantic relationships of interest to the sub-domain of Mortality, developed within the framework of a much larger ontology for a dictionary and thesaurus for Population Geography, which started before the Corpógrafo was available and has not yet been formalised within it. One tool we should like to provide when possible is a way of visualizing both more traditional, hierarchical, thesaurus-style ontologies as well as the multi-dimensional concept systems summarised in Kageura (1997) and Bowker (1997).

The lexical/phrasal databases offer the possibility of text and discourse analysis by classifying words and phrases according to the classification of Rhetorical Structure Theory[4] although users may create their own categories as well. This is clearly a somewhat experimental area, but we hope it will encourage further research.

Despite the emphasis on semi-automatic selection of data from the corpora, it is possible to manually edit the information in the databases and include any terms or information acquired from non-corpora sources. The databases are designed to be either monolingual or multilingual, and to include as much or as little information as needed, and they can be partially or wholly exported to other formats for a variety of uses.

3. NooJ – for information consult http://www.nooj4nlp.net/

4. For an introduction to this, see http://www.sfu.ca/rst/

5.3 Using database information to harvest further texts online

Statistics can be generated automatically from the information in the database to show the frequency and co-occurrence of the terms or expressions in the texts in the corpora. The objective, apart from providing information on the lexical or terminological relevance of the texts, is to use the information to extract new texts from the Internet. There is now a tool, similar to BootCat (Baroni & Bernardini 2004) to select the terms from this evidence and bootstrap further texts from the net.

6. Applications to translator education

It is very possible that, at this point, some people may wonder how far all this is applicable to translator education and how far it is the result of becoming involved in a natural language processing project. Why develop such tools if what the translation market needs is people trained to use the commercially available software?

Perhaps the answer lies in the distinction between 'training' and 'education'[5]. The fact that 'training' is still associated with translators reflects the long-standing attitude in academic institutions towards translation as subsidiary skill. The introduction of translation technology, rather than enhancing the position of the professional translator, both within academia and in the market, has actually contributed to the image of the translator as a 'translation machine', despite the fact that online machine translation has done something to correct the idea that it can substitute the human translator completely, at least in the foreseeable future. If universities continue to 'train' the translator to use the commercially available software, they will merely be contributing to the 'translation machine' image and emphasize skills similar to those formerly expected of secretaries trained to type or take shorthand at high levels of words per minute. On the other hand, if students are 'educated' to use technology to better understand and use language in general and translation in particular, the results will be more satisfactory for all concerned.

Another factor that exacerbates the low prestige of professional translation in the academic community is the tendency to encourage junior staff to focus their research on the more literary and cultural interests that essentially constitute Translation Studies, rather than on the growing, but still novel, areas of research into special domain language, multi-modal communication, terminology and related areas.

There have been several attempts to dignify professional translation. One of these is the European Commission's Directorate of Translation's initiative to estab-

5. Please see the work by Bernardini and Castagnoli (this volume) connecting corpora and translation *education*.

lish a European Master's in Translation (EMT),[6] which explicitly states that one of its goals is to "enhance the value of translation as a **profession** within the EU". The model for a Master's curriculum proposed at the 2006 EMT event includes training in "text/discourse analysis, terminology work, information technology for translation, linguistic awareness, and special fields and their languages", all areas that can use the technology described above. Let us hope that this initiative will encourage not just education but also research in these areas.

7. Final note

The Corpógrafo (http://www.linguateca.pt/corpografo) can be accessed through the Linguateca website and is at present freely available. For server management reasons, access is restricted to users with a username and password. These may be obtained by filling in a form on the Corpógrafo web page and returning it via e-mail. For technical reasons, it has not yet been possible to provide an English version of the Corpógrafo, but there is also a tutorial in both languages, a Portuguese/English glossary of instructions and an FAQ in English.

References

Aijmer, K & B. Altenberg (eds.) (2004) *Advances in Corpus Linguistics*. Amsterdam: Rodopi.

Almeida, S. (2006) *Pesquisa de Informação Terminológica: dos Marcadores Lexicais aos Padrões Suporte. Um Estudo no Domínio dos Riscos Naturais – Cheias.* Porto: FLUP.

Almeida, J. J. (ed.) (2003) *CP3A – Corpora Paralelos, Aplicações e Algoritmos Associados*. Braga: Universidade do Minho.

Baroni, M. and S. Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of LREC 2004*. Paris: ELRA/ELDA.

Bennison, P. and L. Bowker (2000) Designing a tool for Exploiting Bilingual Comparable Corpora. In *Proceedings of LREC 2000*. Paris: ELRA/ELDA.

Benson, M., E. Benson & R. Ilson (1986) *The BBI combinatory dictionary of English: A guide to word combinations.* Amsterdam/Philadelphia: John Benjamins.

Bernardini, S. and F. Zanettin (eds.) (2000) *I corpora nella didattica della traduzione.* Bologna: CLUEB.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Blank, I. (2000) Terminology extraction from parallel technical texts. In Véronis, J. (ed.)

Boguraev, B., J. I. Tait & M. Palmer (2005) (eds.) *Natural Language Engineering Journal*, Vol. 11, Special Issue 03. R. Mitkov, Journal Series Editor. Cambridge Journals (http://journals.cambridge.org).

6. For information see http://ec.europa.eu/dgs/translation/external_relations/universities/master_en.htm

Bowker, L. (1997) Multidimensional Classification of Concepts and Terms. In Wright, S. E. and G. Budin. *Handbook of Terminology Management*. Amsterdam/Philadelphia: John Benjamins, 133–143.

Brito, M. (2005) *Um Conceito = um Termo? Multiplicidade na relação conceito-termo numa Base de Dados Terminológica de orientação conceptual no domínio da terminologia do GPS*. Porto: FLUP.

Cabré, M. T, C. Bach, R. Estopà, J. Feliu, G. Martínez & J. Vivaldi (2004) The GENOMA-KB project : towards the integration of concepts, terms, textual corpora & entities'. In *Proceedings of LREC 2004*. Paris: ELRA/ELDA.

Casais, S. (2003) *Socioterminologia: Um estudo na área dos Materias Compósitos.* Porto: FLUP – Master's dissertation.

Conceição, M. C. (2001) *Termes et reformulations*. Presses Universitaires de Lyon.

Couto, S. (2003) A *Definição Terminológica: Problemas teóricos e práticos encontrados na construção de um glossário no domínio da Corrosão*. Porto: FLUP – Master's dissertation.

Demetriou, G. & R. Gaizauskas (2000) Automatically Augmenting Terminological Lexicons from Untagged Texts. In the *LREC 2000 Proceedings*. Paris: ELRA/ELDA.

Estopà, R, J. Vivaldi & M.T. Cabré (2000) Use of Greek and Latin forms for Term detection. In *Proceedings of LREC 2000*. Paris: ELRA/ELDA.

Halliday, M. A. K. (1984) *An Introduction to Functional Grammar*. London: Edward Arnold.

ISO 704: 2000 *Terminology work – Principles and methods*.

ISO 1087-1: 2000 *Terminology work – Vocabulary – Part 1: Theory and application*

ISO 1087-2: 2000 *Terminology work – Vocabulary – Part 2: Computer applications*

ISO 12200: 1999 *Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange*.

Jesus, C. M. de (2006) *Terminologia e Representação do Conhecimento: Uma abordagem ao Subdomínio da Geodinámica Interna: Sismologia e Vulcanismo.* Porto: FLUP.

Kageura, K. (1997) Multifaceted / multidimensional Concept Systems'. In Wright, S. E. and G. Budin. *Handbook of Terminology Management*. Amsterdam/Philadelphia: John Benjamins, 119–132.

Lewandowska-Tomasczczyk, B. (ed.) (2003) *PALC 2001 – Practical Applications of Language Corpora*. Lódz Studies in language. Frankfurt: Peter Lang.

Maia, B. (2002) Corpora for terminology extraction – the differing perspectives and objectives of researchers, teachers and language services providers. In Yuste Rodrigo, E. (ed.) *Proceedings of the Workshop Language Resources for Translation Work and Research*. LREC 2002, Las Palmas de Gran Canaria, Spain. 25–28.

Maia, B. (2003) What are comparable corpora? In *Proceedings of pre-conference workshop Multilingual Corpora: Linguistic Requirements and Technical perspectives*. Corpus Linguistics 2003, Lancaster: Lancaster University, 27–34.

Maia, B. (2003a) Using Corpora for Terminology Extraction: Pedagogical and computational approaches. In B. Lewandowska-Tomasczczyk (ed.), 147–164.

Maia, B. (2003b) Ontologies. In B. Lewandowska-Tomasczczyk (ed.), 59–68.

Maia, B. (2003c) Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora. In F. Zanettin, S. Bernardini & D. Stewart. 43–54.

Maia, B. (2003d) Ontology, Ontologies, General Language and Specialised Languages. In *Volume Comemorativo dos 25 anos do CLUP*. 23–39. Porto: CLUP.

Maia, B., and L. Sarmento (2003a) The Pedagogical and Linguistic Research Applications of the GC to Parallel and Comparable Corpora. In Almeida, J. J. (ed.).

Maia, B. and L. Sarmento. (2003b) Constructing comparable and parallel corpora for terminology extraction – work in progress. Poster presentation at *Corpus Linguistics 2003* (winners of 1st prize), Lancaster: University of Lancaster.
Mathiessen, C. (1997) *Glossary of Systemic Functional terms*. Available at http://minerva.ling.mq.edu.au/resource/VirtuallLibrary/Glossary/sysglossary.htm
Sarmento, L., B. Maia, & D. Santos. 2004. "The Corpógrafo – a Web-based environment for corpora research". In *Proceedings of LREC 2004*. Lisboa, Portugal, 25 May 2004.
Sarmento, L. & B. Maia (2003) Gestor de corpora – Um ambiente Web integrado para Linguística baseada em Corpora. In Almeida, J. J. (ed.).
Silva, M. M. da (2004) *Estruturação e Representação Gráfica de Sistemas Conceptuais: Uma abordagem no Subdomínio da Mortalidade*. Porto: FLUP – Master's dissertation.
Silva, R. (2006) *Performance and Individual Act Out: The Semantics of (Re)Building and (De)Constructing in Contemporary Artistic Discourse*. Porto: FLUP – Master's Dissertation.
Véronis, J. (2000) (ed.) *Parallel Text Processing – Alignment and Use of Translation Corpora*. Dordrecht: Kluwer.
Vintar, Š. (2001) Using Parallel Corpora for Translation-Oriented Term Extraction. In *Babel* 47:2. 121–132. Amsterdam / Philadelphia: John Benjamins.
Wright, S. E. and G. Budin (1997 /2001) *Handbook of Terminology Management.* Vols. I & II. Amsterdam / Philadelphia: John Benjamins.
Zanettin, F., S. Bernardini & D. Stewart. (2003) *Corpora in Translator Education*. Manchester: St.Jerome Publishing.

Online pointers

British National Corpus: http://www.natcorp.ox.ac.uk/
CETEMPúblico: http://www.linguateca.pt/cetempublico/
Couto, S. (2003) *Glossário de Corrosão.* Available at: http://paginas.fe.up.pt/~mcnunes/QAE/QAE_gloss_b.htm
Déjà Vu: http://www.atril.com/
LETRAC – Language Engineering for Translation Curricula: http://www.iai.uni-sb.de/iaien/en/letrac.htm
Linguateca – http://www.linguateca.pt
NooJ – http://www.nooj4nlp.net/
Rhetorical Structure Theory: http://www.sfu.ca/rst/
SDLX: http://www.sdl.com/
Star Transit: http://www.star-group.net/star-www/home/all/star-group/eng/star.html
Systran: http://www.systran.co.uk/
TRADOS: http://www.trados.com
Wordsmith Tools: http://www.lexically.net/wordsmith/index.html

Acknowledgments

We wish to thank the *Fundação para a Ciência e Tecnologia* for the grant POSI/PLP/43931/2001, co-financed by POSI, and the various members of the Linguateca project who have contributed to the formulation and development of the Corpógrafo.

CHAPTER 5

The real use of corpora in teaching and research contexts

Carme Colominas and Toni Badia
Universitat Pompeu Fabra

The relevance of corpora in translation studies has often been stressed in the literature during the last decade (Zanettin et al. 2003; Olohan 2004; Laviosa 2003). The advantages of corpora as complementary resources to dictionaries, terminologies, etc. have been recognised, and actually the use of corpora as translation resources and of corpus analysis software in general has become part of the syllabus of translation studies. However, the real use of corpora in translation studies still faces (some) practical problems/limitations, as already pointed out by Granger (2003): on the one hand, in some cases, sufficiently large corpora that are representative of modern language do not exist, and on the other, interfaces for accessing corpora are not user-friendly enough to satisfy the real needs of translation students and researchers. In this chapter we deal with these kinds of problems by discussing the weak and strong points of current corpora interfaces and referring to improvements that have already been made and that should continue to be developed in the future. The chapter ends with a revision of corpus-based applications in translation training contexts and in cross-linguistic research.

1. Requirements of corpora for translation teaching

In the last decade, the advantages of corpora used by translators in educational and training institutions have been repeatedly pointed out at a large number of congresses (PALC 2003, CULT and TALC in their various editions) and in a large number of publications (Teubert 1996; Granger (ed.) 2003; Varantola 2000). However, as becomes obvious in the training context, the real use of corpora is still limited by some practical problems to be addressed in this section.

Let us start reviewing the kind of information translation students and researchers actually need. According to our experience as translation trainers, the first information source for translation students, especially for beginners, is still mono- and bi-lingual dictionaries. Consequently, the information which they are interested in finding corpora is mainly that which is not well represented in

dictionaries. One paradigmatic example of this kind of information is when words exhibit a significant polysemy; for example, EN *facilities* or DE *Leistung*, whose translation is strongly context dependent. In such cases, the possibility to check a long list of concordances in a monolingual source corpus or in a bilingual parallel corpus, as pinpointed by Bowker and Barlow (this volume's Chapter 1) can help to provide a better understanding of the different meanings and ultimately facilitate the right translation choice. More frequently, translation students, though translating into their mother tongue, are aware that they have not selected the adequate lexical item or they are not using the right collocate. In such cases, a search in monolingual target corpora to check a list of concordances containing the surrounding lexical item(s) would undoubtedly be a most useful resource. Besides such contextual questions, translators are also often interested in new words or novel uses of words like *chat* or *malware*.

In turn, translation researchers are mainly interested in more specific aspects. For instance, they may be interested in evaluating the adequacy of bilingual dictionaries (checking the translations they offer for certain lexical items) by comparing them with data obtained from a parallel corpus, in comparing structural divergences between languages (e.g., the nominal phrase between English and Spanish), in observing the different translation possibilities of a particular construction (e.g., the translation of German modified compound nouns in Romance languages), in detecting specific features of translated texts by means of a comparable corpus, etc.

In order really to satisfy such interests, both trainee and trainer's corpora access have to fulfil two basic requirements. First of all, sufficiently large corpora which are representative of modern language should be available (ideally as files, so as to be consulted through any interface). The size requirement is especially important for researchers as they aim to study phenomena, only a sufficiently large corpora with enough occurrences (tokens) of not so common words (types) can provide the necessary statistical information (Zipf's law). In the training context, besides the size, the diversity of corpora is equally essential. Ideally a translation trainee can access the following types of corpora: large monolingual corpora in the source and target languages, bilingual corpora (either parallel or comparable, or both), and finally domain specific corpora. Apparently, due to the large number of available corpora, these requirements seem easy to satisfy. However, in spite of the fact that a lot of corpora are available, there are actually very few which satisfy the double requirement of being sufficiently large and being representative of modern language.

With respect to the size, besides the British National Corpus[1] (100 million words), there are other large corpora for major European languages like the IDS (Institut für Deutsche Sprache) corpus[2] for German with 1 billion words, the CREA[3] (Corpus de Referencia del Español Actual) for Spanish with over 200 million words and the CORIS/CODIS[4] (Dynamic Corpus or Written Italian) for Italian with 100 million words. However, there is still a lack of large corpora for other major European languages, and particularly for less-studied languages like Serbian, Polish or Basque. As far as the representativeness is concerned, until now it has been extremely difficult to build large corpora that can satisfy the demand of being representative of modern language. The problem lies in the fact that this type of resource requires a large building effort and has, at the same time, quite a short "lifetime", as it becomes outdated in a relatively short time. Even the BNC does not reflect the language of the last 15 years, so that, for instance, a neologism like *malware* has no occurrences in the corpus. This is the reason why recently the static corpus model has been substituted by the so-called *monitor corpora*, which are constantly updated to track rapid language changes; the CREA corpus, for instance, has been designed as a monitor corpus which is periodically updated so that it always represents the last twenty-five years of the history of Spanish. But taking into account the high price of making representative corpora of modern language, on the one side, and the increasing possibilities offered by the Web as a source of linguistic data (Kilgarriff and Grefenstette 2003), on the other, it seems quite reasonable to state that the future of large corpora lies in the Internet as we will see in Section 2.

In addition to the availability of large corpora that are representative of modern language, the real needs in training contexts also require quick, user-friendly access to the different corpora types (monolingual source and target corpora, as well as bilingual). This requirement stems from the fact that one of the important points often made by translation trainers/trainees and researchers when confronted with the range of electronic resources available in general, is that they recognise the potential usefulness of the data and the tools, but are unlikely to have the time to acquaint themselves with the software. This fact seems particularly true for corpora, if we consider the present state of affairs, referring to the lack of uniform interfaces for accessing resources. Interfaces differ not only in their

1. See *The British National Corpus*, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/

2. http://www.ids-mannheim.de/cosmas2/

3. REAL ACADEMIA ESPAÑOLA: Database (CORDE) [online]. *Corpus diacrónico del español.* http://www.rae.es

4. http://corpora.dslo.unibo.it/coris_ita.html

layout, but in the types of queries they allow for, and this even affects the exploitation possibilities and especially those that imply comparing the results obtained from several corpora. For instance, it is quite difficult to compare the usage of e.g., ES/CA *molar* as verb (for 'to be great', 'amazing', 'cool', etc.) in the jargon of the young Catalan and Spanish, as the available corpus in one language (CUCWeb for Catalan searches) allows for searches by lemma, whereas the one available in the other (CREA) does not. A similar problem arises when we try to compare patterns of use of a verb like *like* in the BNC and *mögen* in the German IDS corpus. Despite being one of the best available reference corpora, the BNC is not lemmatised, which considerably restricts its potential use and the possibilities of performing this kind of comparison with other languages for which a lemmatised corpus is available. In other words, the range of functionality for automated retrieval of corpora is greatly dependent on annotation, and differences between corpora in this matter limit their potential usage considerably. Besides annotation, corpora differ from each other depending on the query language used. Compare, for example, the different query syntaxes by using Xaira (to access the BNC) or Corpus Workbench. Taking into account that translation students and researchers work commonly with at least three or four different languages, they need to access constantly several URLs in order to get familiar with different interfaces and query languages and, what is worse, to face the differences in creating concordances (by form, lemma or part-of-speech (POS)), in gathering statistical information, etc., between corpora. As a result, the usefulness of resources, even when they exist, becomes far from evident for users in general, as too much time must be spent (especially by users that are not trained in query formalisms as is often the case in the context of translation) in order to familiarise themselves with the several interfaces and query languages.

The two aspects we have pointed out as the most desirable aims, that is, the availability of large and representative corpora and a more user-friendly access to the several corpora needed, are being faced nowadays by some corpus developers by means of common platforms that allow access to several corpora eventually built from the Web.

2. Internet corpora: An alternative to large corpora

In recent years the arduous and expensive task of building large corpora has found as a source of linguistic data (Kilgarriff and Grefenstette 2003) real new chances in the World Wide Web. Exploiting the Web as a corpus is becoming a real alternative to the traditional building of large corpora, as can be stated by the Internet corpora compiled at the Centre for Translation Studies of Leeds (Sharoff 2006), the OPUS collection of parallel corpora, or the CUCWeb project developed by the GLiCom

(Grup de Lingüística Computacional, UPF) jointly with the Cátedra Telefónica (UPF) aimed at obtaining a large Catalan corpus from the Web.

In Leeds, general Internet corpora for a range of languages including Chinese, French, German, Italian, Polish, Romanian, Russian and Spanish have been developed in order to cover the needs of researchers and students enrolled at their Centre. The size of all of these general corpora ranges between 100 and 200 million words, which makes them especially suitable for contrastive studies as they are supposed to be comparable. The procedure adopted in these cases for the acquisition of data is based on BootCat (Baroni & Bernardini 2004) and is described in detail in Sharoff (2004).

Following the same idea, although using a different strategy for data compilation (see Boleda et al. 2004), CUCWeb, a 166 million word corpus for Catalan, was built by crawling the Web. This project is especially relevant due to the fact that it deals with a minor language (with some 12 million speakers), similar to Serbian (also about 12 million speakers), or Swedish (9.3). Before CUCWeb, the largest annotated Catalan corpus was the CTILC corpus (Rafel 1994), containing 50 million words stemming from literary and non literary documents between 1832 and 1998, which obviously do not reflect some modern usages of the language.

Of all these efforts to build corpora from the Web, the OPUS collection[5] deserves special attention as it is a collection of parallel corpora which are, as is well known (Badia et al. 2002), much more expensive to build than any monolingual corpus. There is a general scarcity of annotated parallel corpora, but we can list a few here, such as the Europarl[6] (extracted from the proceedings of the European Parliament[7] in 11 European languages), the Canadian Hansard[8] (a parallel corpus in French and English of the proceedings of the Canadian Parliament), the International Telecommunications Union (ITU) or CRATER Corpus[9] (trilingual corpus of Spanish, French and English), the English-Norwegian Parallel Corpus[10] (2.6 million in all), the Chemnitz corpus[11] (about 2 million words), and BancTrad[12] (a parallel corpus containing texts from English, French or German and

5. http://logos.uio.no/opus/

6. http://people.csail.mit.edu/koehn/publications/europarl/

7. http://www3.europarl.eu.int/omk/omnsapir.so/calendar?APP=CRE&LANGUE=EN

8. http://spraakbanken.gu.se/pedant/parabank/node6.html

9. http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

10. http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/

11. http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus/

12. http://mutis2.upf.es/bt/english/index.htm

their respective translations into Catalan or Spanish or vice versa, about 4 million words). Due to the numerous handicaps of building such types of resources, as we can see from the examples we have mentioned, parallel corpora tend to be domain-specific or relatively small. In this context, the OPUS collection becomes much important as it is based on open source documentation that can be downloaded as files, and constitute the largest collection of translated texts available. Training institutions can benefit from the OPUS initiative (adapting the resources to their respective interfaces) and they can contribute to the initiative as well in terms of data or tools. Presently OPUS allows access to the following parallel corpora: EUconst – The European constitution; 21 languages (3 million words), OO – the OpenOffice.org corpus, (30 million), KDE – KDE system messages (20 million), KDEdoc – the KDE manual corpus (3.8 million), PHP – the PHP manual corpus (3.5 million), EUROPARL – European Parliament Proceedings 1996–2003 (296 million).

The most obvious benefit in building corpora from the Web is that they are easy and cheap to make. Furthermore, they reflect modern language use, are easily extensible and updated, and promote the technological development of non-major languages. Besides such obvious benefits, some shortcomings have already been pointed out such as the fact that not all topics, not all text types are equally available, and that such biases become far more evident across languages. This is in fact true. Internet corpora can no longer meet some of the traditional requirements made of corpora (e.g., to be balanced); however this does not need to be seen (only) as a handicap. Actually, with the exception of the BNC, most of the so-called general corpora are also heavily biased (e.g., towards newspaper texts, like the IDS or the FR German corpora, or towards literary texts like the Catalan CTILC). Internet corpora are at least representative of the language on the Web, and this can also be seen as valuable information, e.g., from a sociolinguistic perspective; consider for instance the possibilities of contrastive studies between languages and cultures by means of comparable Internet corpora. They can provide interesting and valuable information from a contrastive point of view to assess the impact of text genre and topic in Internet corpora across languages/cultures (Sharoff 2006). However, the real possibilities of carrying out such studies also depend on the possibilities of accessing the corpora through adequate interfaces.

3. The need for common interfaces to several corpora

As pointed out in Section 1 above, the usefulness of corpora resources, even when they exist, becomes not so evident for users due to the fact that too much time must be spent in order to become familiar with the several interfaces and query languages. The problem derived from the multiplicity of interfaces and query lan-

guages is increasingly being solved by the creation of uniform interfaces, especially in translation training institutions. Some examples of these are the Leeds interface (at the Centre for Translation Studies of the University of Leeds), the UFR Eila platform (at the Université Diderot, Paris 7), the BancTrad and CUCWeb corpus interfaces (at the Department of Translation and Philology of Pompeu Fabra University and Barcelona Media) or the OPUS interface (Tiedemann & Nygaard 2004). All these interfaces allow access to more than one corpus and have been developed with the aim of serving as a uniform platform to cover different user needs.

At the Centre for Translation Studies in the University of Leeds, a corpus interface[13] has been developed that allows access to Internet corpora for Chinese, French, German, Italian, Spanish, Polish and Russian. This interface allows queries with search expressions that can contain exact word forms, lemmata, POS, substrings or unknown words. The interface offers an option of simple queries akin to Google that translate into a corpus workbench query, e.g., a simple query term corresponds to a lemma, while a term in double quotes corresponds to a word form. However, queries combining POS and lemma restrictions must be written according to the Corpus Query Processor (CQP) syntax, which requires that the user must be familiar with it. A similar requirement is made by the UFR Eila and by the OPUS interface. In the former, several comparable corpora (EN-FR) from specific sub domains (water, volcanoes, mountains, etc.) can be consulted by using the syntax of regular expressions in Perl. And in the OPUS page a multilingual concordancer using the CQP is available for most of the subcorpora. Currently, searches by word, lemma and POS can be made in the "source" language. The two interfaces developed at UPF allocate several corpora. On the one hand, BancTrad[14] currently accommodates 2 monolingual corpora as well as the one multilingual parallel corpus referred to in Section 2 above. The two monolingual corpora are the BNC for English and the ECI corpus Frankfurter Rundschau for German (about 34 million words of newspaper texts). On the other hand, CUCWeb[15] allocates the Catalan corpora CUCWeb and CTILC mentioned in Section 2 above. What basically distinguishes the UPF interfaces is the fact that they are user-friendly interfaces that can be used by both non-trained and more experienced users as no knowledge of any query syntax is required. The simple mode query allows searches for words, lemmata or word strings and can be used by any untrained user. Expert mode allows queries of strings of up to 5 word units, where each unit can be a word form, lemma, part-of-speech, syntactic function or combination of any of those. The fact

13. http://corpus.leeds.ac.uk/internet.html

14. http://mutis.upf.es/bt/english/index.htm

15. http://ramsesii.upf.es/cgi-bin/CUCWeb/search-form.pl

that no knowledge of any query syntax is required is especially welcomed by relatively untrained users, like we might expect in the context of translation training (students as well as teachers). However, it also has its counterpart, as it cannot be as powerful and expressive. From this point of view, the Leeds and the DTF interfaces are examples of the effort of finding a compromise between user-friendliness and expressiveness.

As for the search for statistical information, the Leeds and the CUCWeb interfaces provide functionalities to evaluate the relative frequency of phenomena. The Leeds interface includes a function that enables the computation of collocation statistics (using mutual information, the T-score or the log likelihood score), thus providing information which can be very useful for translation tasks. In the CUCWeb interface, frequency information can be related to any of the 4 annotation levels (word, lemma, part-of-speech (POS), and syntactic function).

An interesting initiative that may lead to results in the direction indicated here is the WaCky project.[16] The WaCky initiative aims at developing tools for the use of the Web as a corpus of interest to linguists, which includes tools for crawling the Web, building, annotating, indexing and searching a corpus.

4. Translation-related exploitation possibilities of a common corpora platform

As we have seen in the last two sections, in recent years efforts have been made in order to overcome the limitations of a corpus-based methodology and to create an appropriate workbench for cross-linguistic research. We now consider some exploitation possibilities in research and translation training contexts from such a common platform that integrates different types of corpora that are at least lemmatised and annotated for part-of-speech.

4.1 Training contexts

In our experience as translation trainers, we have seen the effectiveness of corpus-based methods in translation training as underlined by Bowker (1998) confirmed. It seems unquestionable that translation training must be based on translation practice. However, the classical activities in translation training (exercise and correction) are obviously insufficient to develop the skills trainees need in order to progress in their translation competence. It seems quite reasonable to affirm that in order to prevent errors or avoid similar errors in future exercises, trainees need

16. http://wacky.sslmit.unibo.it

to be confronted with a range of examples of the same (or similar) phenomena from which they can gain some sort of generalisation. In other words, translation difficulties or errors should be illustrated through other real examples from which trainees can get a major awareness and a better understanding of the corresponding phenomena. The types of corpora that can be involved in translation training activities are mainly monolingual in the source and the target language as well as parallel and comparable corpora, depending on the aims pursued.

From a general point of view, we can consider that the two main problems with which students are often faced during translation activities into their native language can be divided into:

– comprehension difficulties
– reformulation difficulties

Comprehension difficulties may involve mainly lexical or syntactical aspects. Most lexical problems have to do with highly context dependent words like modal verbs or discourse particles; such types of words can pose translation problems for students, as their meaning is complex and highly dependent on contextual features. In order to get familiar with the different translation strategies in translating, such lexical items in different contexts, monolingual corpus in the source language or parallel corpus can be helpful.

– For instance, Catalan students faced with the translation of the German particle *erst* can reach the generalisation that the particle *erst* is mainly translated into Catalan by negating the sentence, through the concordances obtained from a monolingual or parallel annotated corpus, as we can see in Table 1.

Table 1. Some results from the parallel corpus BancTrad for the word *erst* followed by Preposition.

DE: Auch danach war Alkohol für die Mehrheit junger Leute nicht von Belang, er wurde für sie **erst ab** den 60er Jahren populär
CA: No va ser fins a partir de els anys 60 que l' alcohol es va popularitzar entre el jovent.

DE: Und das, obwohl Sie **erst seit** eineinhalb Jahren bei uns arbeiten
CA: I això , tot i fer **només** un any i mig que treballa amb nosaltres.

– Besides search for forms, mainly annotated corpora allow search for lemma. Obviously, this type of search is especially useful for searching verbs in order to obtain patterns from a contrastive point of view. Concordances containing any form of a given verb can help to remark regularities on translation equivalents like the following determined by the verb tense:

Table 2. Some results from the parallel corpus BancTrad for the lemma *mögen*.

DE: Sie **mag** den Spot, aber ich hätte gerne was mit mehr Action
CA: A ella sí que li **agrada** , però jo preferiria una mica més d' acció

DE: Er ist wie ein Kind, das schlafen **möchte**
CA: És com un infant que **voldria** dormir

DE: Und er **möchte** nach Hause, für einen Augenblick nur, nur für so lange, als es braucht, um die Worte zu sagen:
CA: Aconseguia que hom parlés de si mateix , que s' exasperés i ja no **pogués** parar .

– In a similar way, the possibility to restrict the search for both POS and lemma allows obtaining concordances like the following from which the different translations of a given verb can be extracted depending on the preposition:

Table 3. Some results from the parallel corpus BancTrad for the lemma *liegen* followed by a Preposition.

DE: Was **liegt an** meinem Mitleiden!
CA: " Què **hi fa** la meva compassió?

DE: Ein weiteres Problem **liegt in** der Tatsache, daß heute keine ausreichenden Maßnahmen getroffen werden, um den Schutz der Sprache im Alltagsleben und beispielsweise auch im Bildungswesen zu gewährleisten.
CA: Un problema addicional **rau en** el fet que avui dia no es prenen mesures suficients per a garantir la protecció de la llengua en la vida quotidiana ni en l' ensenyament .

DE: Sie **liegt auf** dem Rücken
CA: **Jeu** panxa enlaire

DE: Nach 45 Minuten ist alles vorbei, er **liegt neben** mir.
CA: A el cap de 45 minuts s' ha acabat tot , **jeu a el** meu costat

DE: Das Dorf **lag in** tiefem Schnee
CA: Una neu espessa **cobria** el poble .

– Searches for POS sequences offer multiple possibilities to practise structural divergences between languages like the following in nominal phrases between German and Catalan:

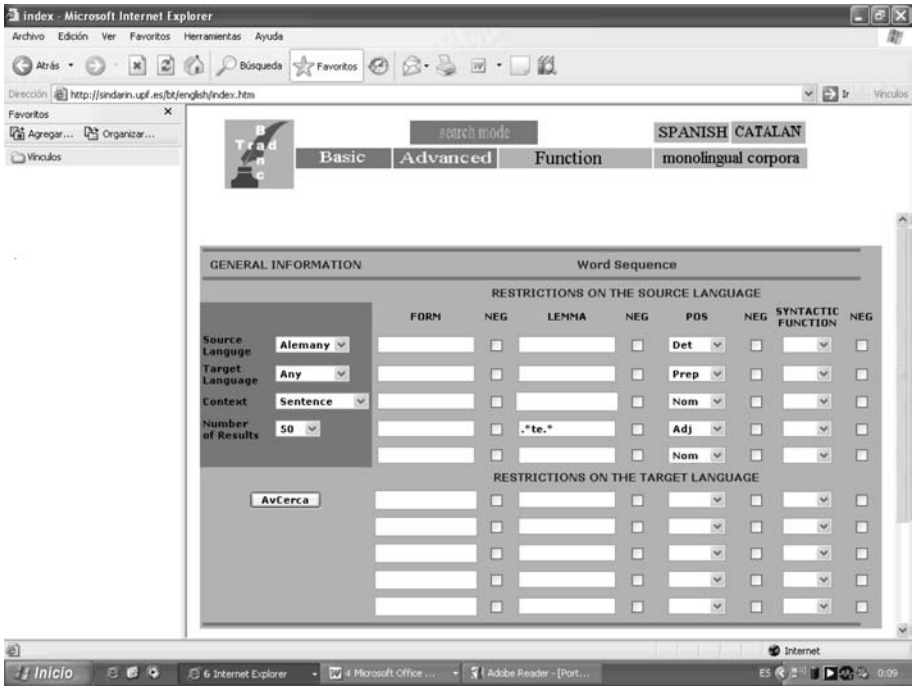

Figure 1. Search in the BancTrad corpus for the sequence Det-Prep-Noun-Adj (partially restricted to Past Participle)-Nom.

The output of such a query (partially in Table 4) can be used as source for several exercises with the aim of reinforcing the practice in translating these structures.

Table 4. Some results of the search captured in Fig. 1

DE: I.14 Instrumentarium für **eine auf Zukunftsbeständigkeit gerichtete Kommunalverwaltung**
CA: 1.14 Instruments i eines per a **la gestió urbana cap a la sostenibilitat.**

DE: 1451 errichtete er an **einem nach Süden gerichteten Chorpfeiler** des Stephansdoms eine Sonnenuhr
CA: El 1451 va construir **un rellotge de sol en un pilar** de la catedral de Sant Esteve , **encarat al sud** .

DE: Über **dem nach West-Nord-West gerichteten Zifferblatt** von I bis VIII mit römisch gotischen Ziffern befindet sich das Wappenschild des Habsburgers Erzherzog Ferdinand II von Österreich
CA: Sobre **el quadrant orientat a oest-nord-oest**, amb xifres goticoromanes i línies horàries de la I a les VIII de la tarda , hi trobem l' escut de l' arxiduc habsburg Ferran II d' Àustria .

As is known, this kind of nominal phrases constitutes a crucial difficulty in the translation from Anglo-Germanic into Romance languages as they imply non trivial restructuring tasks in the target language. Because of the diversity of modifiers (from the grammatical but also from the semantic point of view) that can be involved, it is not possible to establish so many regularities in the translation of these complex nominal phrases from German. From annotated parallel corpora, many different types of such phrases can be extracted in order to observe different translation solutions and/or to practise the translation from similar examples.

– Finally, from corpora annotated with some textual information such as reference, subject, type of text, degree of difficulty, etc. (like BancTrad) interesting searches can be done on sub-corpora in order to illustrate, for example, translation options determined by such kind of parameters. This is what happens e.g. with the preposition *nach* in German, which in legal texts often has a particular translation other than its common translation as a temporal or directional preposition:

Table 5. Restriction on text_domain: legal

DE: §52 Mitwirkung in Verfahren **nach** dem Jugendgerichtsgesetz
CA: Article 52 : Intervenció en procediments **d' acord amb** la Llei d' ordenació de els tribunals de menors

DE: §87c Örtliche Zuständigkeit für die Beistandschaft, die Amtspflegschaft, die Amtsvormundschaft und die Auskunft **nach** §58a
CA: Article 87c : la competència local per a l' assistència en l' exercici de la guarda , la curatela administrativa, la tutela administrativa i la informació **en virtut de** l' article 58a

Further exploitation possibilities of parallel corpora lie in the possibility of searching in the target language. One interesting issue, for both didactic and research purposes, is the exploration of alternative translation strategies to the canonical ones. For instance, one can be interested in translations of a given modal verb, which do not meet the translation possibilities that can be found in dictionaries. Such a query would presuppose the possibility to express restrictions (negation) over the target languages; this is actually a real possibility for a corpus processed with the CQP. Nevertheless, this is a possibility that unfortunately, as far as we know, has not yet been implemented in any interface.

As for reformulation difficulties, during the re-expression phase, obviously a large corpus in the target language will be mostly of help. Note that in training contexts, reformulation difficulties are always much more common than would be expected. Despite translating into their mother tongue, trainees are often faced with a lack of precision in and adequacy of lexical choice and, more specifically, with the question of finding the right collocate. For instance, by translating a German collocation like *seine Begabung fördern* into Catalan, trainees will easily find

the target language equivalents of the noun but will sometimes not be so sure about the verbs that collocate with the noun in such contexts. For this purpose it can be helpful to search for the relative frequency of the verbs with which the noun collocates. A search in the target monolingual corpus from the CUCWeb interface will help trainees to select the most context adequate translation:

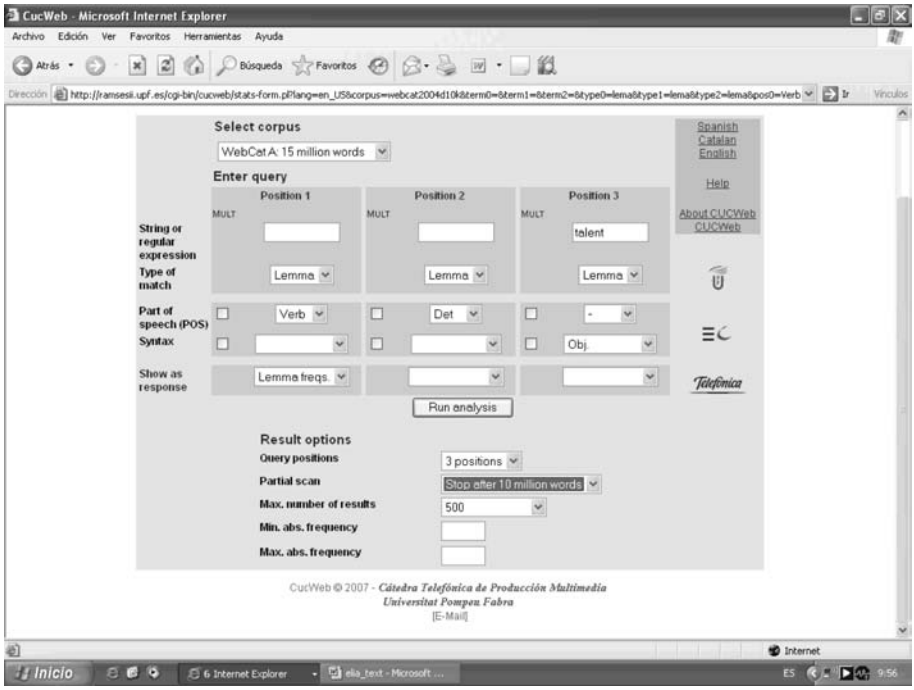

Figure 2. Frequency search from the CUCWeb interface: The lemma *talent* as an object preceded by a verb and a determiner

This search returns a list with the frequencies (Fig. 3) of the lemma corresponding to the verbs that more often collocate with *talent* as an object.

Inadequate lexical selection in the widest sense is actually one of the most frequent errors made by translation trainees. To deal with collocates, the translator must have a good deal of knowledge about the stylistic factors that enter into lexical selection and collocation in the target language. But this knowledge must be trained and a large corpus in the target language that allows this kind of statistical queries like the CUCWeb or the Leeds interface is a valuable resource for training activities in order to achieve better translation competence. In order to appreciate the benefits of having quick, easy access to such an interface, we have to consider that this kind of resource can be used by trainees during translation or assess-

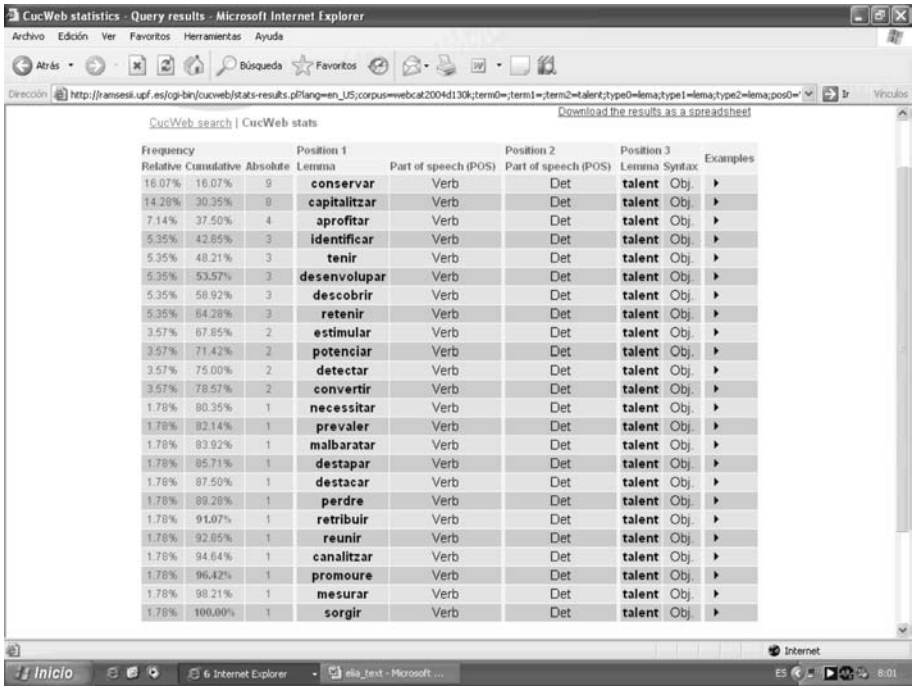

Figure 3. Results for the frequency search in the CUCWeb interface of the lemma *talent* preceded by a determiner and a verb

ment tasks but also by trainers during training tasks or for the creation of teaching contents in general.

Besides such general questions, re-expression difficulties are also determined by the degree of specialisation and by the language pair involved. As known, not all languages exhibit the same degree of representativeness and cohesion; minor languages like Catalan, Serbian or Basque are obviously not so well represented in all domains as English or German. As a consequence, when translating from such a language into a minor one, translators (and especially trainees) are often faced with difficulties related with a lack of knowledge of the right phraseology or terminology, especially for texts of a relatively specific domain. In such a context, a large corpus in the target language can obviously be of great help.

4.2 Research

Cross-linguistic studies mainly involve much more complex phenomena than those that we have seen in training contexts. We refer under cross-linguistics studies both to corpus-based translation studies (TS) as well as contrastive linguistics (CL). Under the different multilingual corpus types, as outlined by Granger

(2006), some are more relevant for TS and others for CL; however, in both types of research, some kind of comparison between languages is implied. As noted by Granger, comparison between corpora of original texts in different languages is the most important domain of CL, whereas the comparison of translated texts in different languages is the preference of TS. There is however a quite wide domain intersection between both disciplines, related to their objectives, that makes it difficult clearly to distinguish between both. In other words, researchers in both fields are equally interested in parallel and comparable corpora for their studies; the distinctive point probably lies in the objectives pursued: the objective for CL is to detect differences and similarities between languages whereas the final objective for TS is to capture the features of translation (as process and as result). We will now refer to some examples of cross-linguistic research types based on corpora:

– Cross-linguistic studies between translated and non-translated language focusing on features of *translationese* on the basis of the universals hypothesised by Baker (1993) that include simplification, explicitation, normalisation and concretisation (e.g., Puurtinen 2006), or focusing on features induced by the source language. For such studies a comparable annotated corpus is desired.
– Contrastive studies investigating expressions of a certain concept or semantic field in two different languages (e.g., Löken 1997) by means of a parallel corpus.
– Studies of comparison and translation of fixed expressions between two languages. For instance, Charteris-Black, J. (2003) proposes establishing a model for the comparison and translation of English and Malay idioms to facilitate the translator's task.
– Studies comparing the translation of events of a certain semantic field between languages. An interesting study could be made by comparing verbs which express noise by translations from a Germanic into a Romance language.
– Comparison between the tense system in two languages and how the differences are overcome in translated texts. For instance, Chuquet, H. (2003) studies how the French *imparfait* is rendered in English texts through aspectual or modal additions and showing some common features between French *imparfait* and English past tense.
– Comparison between the voice system in two languages. Davidse, K. & L. Heyvaert (2003) compare, for instance, the middle formation in English and Dutch. Other interesting studies could involve comparison of the use of the passive voice in Germanic and Romance languages.

So far, we have exemplified some of the multiple exploitation possibilities and the benefits of having at one's disposal a uniform user-friendly interface to access different types of corpora in translation training contexts. Platforms providing access to corpora for translation training should allow access to the following types of

corpora: a monolingual source, a monolingual target, and a parallel corpus. Access to other corpora types (e.g., comparables) or to adequate tools in order to build ad-hoc corpora could also be included in the platform.[17] Furthermore, as we have seen through the examples above, corpora should ideally be linguistically annotated and provide some advanced functions to extract collocations or frequency information over co-occurrences. And for parallel corpora, we have pointed out that ideally, queries should be possible on both the source as well as the target language. Part of these requirements are already satisfied by some of the current interfaces, as it has been pointed out in Section 3, which proves the efforts made as well as the efforts that must still be made in order to overcome some of the limitations that the real use of corpora still face.

5. Conclusion

We have analysed the weak and strong points of some of the current interfaces to corpora taking into account the real needs in translation training contexts (considering trainees as well as trainers and researchers). Weak points refer mainly to a lack of availability of large general corpora for some languages as well as a lack of uniform interfaces that provide all the necessary data and functionalities. The lack of sufficiently large corpora representative of modern language is currently being solved by means of web corpora, a promising alternative especially for non-major languages. In parallel, common interfaces for accessing corpora have been developed. By exemplifying search types that can be relevant in translation training contexts or for translation research purposes, we have identified the basic requirements a corpora interface should satisfy. Our analysis provides evidence of work done regarding this matter but also of the need for further work. Though the development of adequate corpora interfaces is actually not a central activity in Natural Language Processing (NLP), it is essential for practical purposes, and will help to make the NLP product accessible to a wider community.

References

Badia, T. et al. (2002) BancTrad: a web interface for integrated access to parallel annotated corpora. In Yuste Rodrigo, E. (ed.) *Proceedings of the LREC 2002 workshop on Language Resources for Translation Work and Research*, Las Palmas de Gran Canaria, Spain, 28th May 2002. Paris: ELRA/ELDA.

17. See for instance the BootCat toolkit developed by Baroni & Bernardini to bootstrap ad-hoc specialised corpora from the Internet.

Baker, M. (1993) Corpus Linguistics and Translation Studies – Implications and Applications. In Baker, M., F. Gill & E. Tognini Bonelli. *Text and Technology: In Honour of John Sinclair*, Amsterdam & Philadelphia: John Benjamins, 233–250.

Baroni, M. and S. Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of LREC* 2004. Lisbon, Portugal. Paris: ELRA/ELDA.

Boleda, G., S. Bott, C. Castillo, R. Meza, T. Badia, and V. López (2006) CUCWeb: A Catalan corpus built from the Web. In *Proceedings of the Second Workshop on the Web as a Corpus* at *EACL'06*. Trento, Italy.

Charteris-Black, J. (2003) A prototype based approach to the translation of Malay and English idioms. In Granger, S., J. Lerot and S. Petch-Tysin (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Approaches to Translation Studies, vol. 20, Amsterdam: Rodopi.

Chuquet, H. (2003) Loss and gain in English translations of the French *imparfait*. In Granger, S., J. Lerot and S. Petch-Tysin (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Approaches to Translation Studies, vol. 20, Amsterdam: Rodopi.

Granger, S. (2003) The corpus approach. A common way forward for Contrastive Linguistics and Translation Studies? In Granger, S., J. Lerot and S. Petch-Tysin (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Approaches to Translation Studies, vol. 20, Amsterdam: Rodopi.

Davidse, K. & Heyvaert, L. (2003) On the middle constructions in English and Dutch. In Granger, S., J. Lerot and S. Petch-Tysin (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Approaches to Translation Studies, vol. 20, Amsterdam: Rodopi.

Kilgariff, A. (2001) Web as corpus. In *Proceedings of the Corpus Linguistics 2001 Conference*. University centre for computer corpus research on language, Technical papers, vol. 13, Lancaster: Lancaster University.

Kilgarriff, A. and G. Grefenstette (2003) Introduction to the Special Issue on the Web as Corpus. In *Computational Linguistics*, 29 (3), 333–348.

Laviosa, S. (2003) Corpora and Translation Studies. In Granger, S., J. Lerot and S. Petch-Tysin (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Approaches to Translation Studies, vol. 20, Amsterdam: Rodopi.

Löken, B. (1997) Expressing possibility in English and Norwegian. In *ICAME Journal 21*, 43–59.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*. London: Routledge.

Puurtinen T. (2003) Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing* 18 (4), 389–406.

Rafel, J. (1994) Un corpus general de referència de la llengua catalana. In *Caplletra*, vol. 17, 219–250.

Sharoff, S. (2004) Methods and tools for development of the Russian Reference Corpus. In Archer, D., A. Wilson & P. Rayson (eds.), Corpus Linguistics around the World. Amsterdam: Rodopi.

Sharoff, S. (2006) A Uniform Interface to Large-Scale Linguistic Resources. In *Proceedings of the LREC 2006*, Genoa, Italy. Paris: ELRA/ELDA.

Teubert, W. (1996) Comparable or Parallel Corpora? In *International Journal of Lexicography* 9 (3), 38–64.

Tiedemann, J. and Nygaard, L. (2004). The OPUS corpus parallel & free. In *Proceedings of the LREC 2004*, Lisbon, Portugal. Paris: ELRA/ELDA.

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

Zanettin, F., S. Bernardini & D. Stewart (eds.) (2003) *Corpora in Translator Education.* Manchester: St Jerome Publishing.

CHAPTER 6

The use of corpora in translator training in the African language classroom

A perspective from South Africa

Rachélle Gauton
University of Pretoria

This chapter presents the translator training curriculum at the University of Pretoria as a case study to show how corpora can be used successfully in the training of African language translators, with particular reference to translating into the South African Bantu languages. These languages were marginalised and disadvantaged during the apartheid era, particularly as far as the development, elaboration and standardisation of terminology is concerned. Consequently, these languages lack (standardised) terminology in the majority of (specialist) subject fields which makes translation into these languages (and not only technical translation), an activity fraught with challenges. This chapter focuses on how training in the use of electronic text corpora, corpus query tools and translation memory tools can enable the African language translator to:

– mine existing target language texts for possible translation equivalents for source language terms that have not been lexicalised (in standardised form) in the target language;
– coin terms in the absence of clear and standard guidelines regarding term formation strategies, by making use of those term formation strategies preferred by the majority of professional translators;
– re-use existing translations in order to translate more efficiently and effectively and to attain some form of standardisation as far as terminology is concerned, given the lack of up-to-date and usable standardised terminologies in these languages.

1. Introduction

Using corpora in translator training seems to be a relatively widespread and common practice in the West (specifically in various institutions in parts of Europe and the Americas), as can be gleaned from the work of authors such as *inter alia* Bowker (1998, 2000: 46–47, 2002), Calzada Pérez (2005: 6), Fictumova (2004),

Izwaini (2003: 17), Laviosa (2003: 107–109, 2004: 15–16, 20–21), Maia (2002: 27, 2003a: 30–31, 2003b); McEnery & Xiao (2007), Varantola (2002) and Zanettin (1998, 2002). However, this does not seem to be the case on the African continent, and particularly in South Africa. As far as could be ascertained, published literature does not attest to the use of corpora in translator training at African (higher) education institutions, with the notable exception of Tiayon's (2004) article on the use of corpora in translation teaching and learning at the University of Buea, Cameroon. In South Africa too, higher education and other training institutions have generally not yet incorporated the use of electronic text corpora in their training curricula, particularly as far as translation into the African languages (including Afrikaans) are concerned. For example, Goussard-Kunz (2003) indicates that at the time of her study, translator training in the South African Department of Defence's *African language translation facilitation course* (*ALTFC*) followed contemporary trends in translator training, but without making use of electronic corpora in the training programme.

An exception to the rule is the translation curriculum of the University of Pretoria (UP), where (in 2004) courses on the application of Human Language Technology (HLT) in translation practice were established, focusing on *inter alia* the use of corpora as translation resource, translator's aid and translators' tools, with specific reference to technical translation into the official SA languages. In this chapter, therefore, the intention is to present the translator training curriculum at the University of Pretoria as a case study to show how corpora can be used successfully in the training of African language translators, with particular reference to translating into the South African Bantu languages.

First, however, a brief overview needs to be given of the language situation in South Africa.

2. The South African linguistic situation

Before the advent of the new democratic dispensation in South Africa in 1994, there were two official languages, namely Afrikaans and English. The various Bantu languages spoken in the country had no official status, except in the so-called *bantustans* that had no legitimacy outside of the *apartheid* context. Furthermore, by means of the so-called Bantu Education system, the *apartheid* regime exploited and harnessed the SA Bantu languages as vehicles to entrench white supremacy, racial domination, oppression and discrimination and to attempt to create ethnic divisions between speakers of the various Bantu languages. As Oliver & Atmore (1972: 261) put it so succinctly, the so-called Bantu Education Act of 1953 "took African education out of missionary control, and made it an instrument of government policy in reshaping men's minds." This policy eventually

played a significant role in the so-called 'Soweto uprisings' of 1976, which started off as a protest against the forced use of Afrikaans (the language of the dominant Afrikaner group) as medium of instruction in black schools.

Afrikaans is (genetically and structurally) a Germanic language that has its roots in a 17th century Dutch variety that underwent significant changes on African soil due to influence from a variety of cultural and linguistic groups that it came into contact with, developing into so-called 'Cape Dutch'. Already in the second half of the 18th century Cape Dutch moved away from European Dutch and became a language in its own right, transforming into what became known as Afrikaans (Raidt, n.d.). During the *apartheid* era, Afrikaans was seen as the language of the oppressor even though it was (and still is) the mother tongue of large groups of people who were not classified as 'white' during this time, but as so-called 'coloured' and 'black' and who were subsequently discriminated against and politically and socially disadvantaged because of this reason. Today, Afrikaans is embraced (also by many white mother-tongue speakers of the language) as an African language, born of Africa.

With the advent of democracy in South Africa in 1994, eleven official languages were recognised. In addition to the two official languages under the previous dispensation, *viz.* Afrikaans and English, the other nine official languages are the following previously disadvantaged and marginalised South African Bantu languages: four languages belonging to the Nguni group of languages, namely Zulu, Xhosa, Ndebele and Swati; three languages belonging to the Sotho group of languages, namely Sepedi, Sesotho and Tswana; plus Tsonga and Venda[1]. The South African constitution affords all eleven official languages equal status in all domains in order to provide access to everyone, irrespective of their language preference.

In reality, however, some of the official SA languages are more equal than others (to paraphrase George Orwell). There is no denying that (because of its status as international language) English tends to dominate political and public discourse. As for Afrikaans, it has well developed terminologies in most technical subject fields and a much stronger terminological tradition than the South African Bantu languages, due to the preferential treatment that this language enjoyed *vis-à-vis* the Bantu languages during the *apartheid* era. In addition to a lack of terminology in most (specialist) subject fields, the SA Bantu language translator also has to contend with the reality that the various National Language Bodies (NLBs) that replaced the *apartheid* era Language Boards, and that are the custodians of these languages charged with amongst other duties with the standardisation of the languages, cannot possibly keep up with the demand for standardised terminologies

1. The Khoi, Nama and San languages and SA sign language, although not official languages, are promoted together with the official SA languages.

needed by the Bantu language translator on a daily basis. There are woefully few technical dictionaries and terminology lists and/or glossaries available in any of the official SA Bantu languages, and this coupled with the lack of guidance regarding which terms should be regarded as standard as well as regarding term formation strategies, puts translators working into the SA Bantu languages in the unenviable position of having to create terminology when undertaking almost any translation task, and not only technical translations. Under these sets of circumstances, the translator ends up working in isolation, as there are usually no standard guidelines which can be followed or authoritative sources that can be consulted. Each translator thus effectively ends up creating his or her own terminology – on the fly, so to speak.

Another drawback of this situation is that translators working into the SA Bantu languages, often do not document their terminology for future re-use, mainly because these translators are not familiar with, and/or trained in the use of, the various electronic tools on offer that could assist them in this task. Furthermore, no easily accessible mechanism exists in South Africa yet that would allow translators to pool their resources, although various discussions have taken place towards creating such mechanisms. Should translators be able to share their resources, this would also go some way towards opening up a dialogue regarding the terminology used in various domains and could even result in achieving some form of terminology standardisation, even though this would not be an officially sanctioned standardisation process.

Despite the lack of standardised terminologies, translation and localisation into the SA Bantu languages and Afrikaans are proceeding apace. As Kruger (2004:2) points out:

> In South Africa, translation and interpreting are the main areas in which the technical registers of African languages and Afrikaans are being developed and standardised [...].

The increase in the availability of target texts in the official SA languages, particularly on the Web, creates the ideal opportunity to use these resources in translator training. This will be the topic of the next section where translator training at the University of Pretoria will be presented as a case study.

3. The use of corpora in translator training at the University of Pretoria: A case study

In 2000, translator training at the University of Pretoria was introduced at undergraduate as well as at postgraduate levels, and in 2004 a number of postgraduate modules on the application of Human Language Technology (HLT) in transla-

tion practice were added. The latter modules employ general, comparable, parallel, special purpose and DIY corpora in training student translators, and also provide students with training in using such corpora as translation resource and translator's tool.

3.1 Outline of the curriculum

The UP translation curriculum, and specifically the courses focussing on the application of HLT in translation practice, can be viewed in the yearbook of the Faculty of Humanities at the following web address: http://www.up.ac.za/academic/eng/yearbooks.html.

3.2 Available resources and infrastructure

UP's Department of African Languages which hosts the translator training courses on behalf of the School of Languages, has the following resources and infrastructure at its disposal to provide the necessary corpus-based training:

a. General (electronic) corpora for all the official SA languages.[2] UP is the only higher education institution in South Africa that possesses such large general corpora in all the official languages, and particularly in the SA Bantu languages. The respective sizes of the different corpora are as follows (see Table 1).

Table 1. Sizes of the UP general corpora (as on 2 Feb. 2003)

Language	Size in running words (tokens)
Afrikaans	4,817,239
English	12,545,938
N.Sotho	5,957,553
Ndebele	1,033,965
S.Sotho	3,159,568
Swati	316,622
Tsonga	3,533,964
Tswana	3,705,417
Venda	2,462,243
Xhosa	2,400,898
Zulu	5,001,456

2. These corpora were compiled by D J Prinsloo and/or G-M de Schryver, in collaboration with M J Dlomo and members of some of the National Lexicography Units (NLUs), except for the English corpus that was culled from the Internet by R Gauton.

These electronic corpora are all written, non-marked up and non-POS tagged raw corpora consisting of a number of sub-corpora stratified according to genre. The 5 million words untagged and unmarked running text Zulu corpus can be cited as a representative example. This corpus is organised chronologically and is stratified according to genre as follows: novels & novelettes; textbooks; short stories, essays & readers; dramas & one-act plays; religious texts; poetry; oral literature, folklore & legends; Internet files & pamphlets.

b. Large computer laboratories with Internet connections in which a series of workshops are run for the students so that they can be provided with hands-on experience of the various electronic translation resources and translators' tools. Students also take their final examination in the computer laboratory and are expected to demonstrate that they have mastered the use of the various electronic resources and tools, including the use of corpora in translating texts, usually of a technical nature.

3.3 Prerequisites for the courses, with specific reference to student profiles

A prerequisite for taking these courses is basic computer literacy, but in the South African context, this requirement is not always as straightforward as it would seem. The majority of students taking these courses come from disadvantaged backgrounds, where they have grown up without easy access to computers in the home or school environment. Although the students consider themselves to be computer literate, this is often not the case from the lecturer or translator trainer's perspective. Students would, for instance, be able to handle e-mail, have basic knowledge of how to surf the Internet and a rudimentary knowledge of a program such as Word and possibly Excel, but they will struggle in negotiating the Windows environment, working with different file formats, quickly familiarising themselves with new software, etc. The students also commonly share the belief that the African languages (particularly the Bantu languages) cannot be used as high function languages and/or in combination with cutting edge technology, and that this is solely the domain of English, and maybe Afrikaans at a stretch.

Consequently, before students are introduced to the use of corpora in translation practice, they first have to be familiarised with various online resources that can be utilised by the translator, e.g., the use of search engines, online dictionaries, thesauri, (automatic) translators, etc. Students usually realise very quickly that in a class containing students translating into some of the major languages such as French and German, students translating into a lesser used language such as Afrikaans, and students translating into the previously marginalised languages such as the Bantu languages, the available online language resources decrease as one progresses along this continuum. Also, some of the 'smaller' Bantu languages

such as Ndebele and Swati, have even less language resources on the Web than some of the other (larger) Bantu languages such as Zulu. In addition, over the years it has become clear that one of the biggest challenges for the trainee African language translator is trying to determine the exact meaning of the source language item, where the SL is usually English. It must be borne in mind that English is these students' second, third or sometimes even fourth language and that many South Africans know seven or more Bantu languages in addition to English and usually also Afrikaans. Because of the lack of terminology in the Bantu languages, students more often than not have to create translation equivalents, and in order to do so, they must be sure of the exact meaning of the English SL term. It is therefore important that students are also introduced to online English language resources that will provide them with this type of information, such as for example the excellent Visual Thesaurus. As indicated, students must be able to use search engines, specifically in accessing texts written in (or translated into) the target language. Students are usually surprised at the number of sites containing texts in the African languages that they are able to find on the net. Many of these sites are excellent sources of parallel texts in all the official SA languages. Students are also introduced to methods for finding sites and web pages that are no longer available on the net, such as using the Internet Archive Wayback Machine and the 'cached' option in Google.

As regards the profiles of the students taking these courses; they all translate from English as source language (SL) into their first language / home language, and the breakdown per target language (TL) is as follows: Afrikaans 24%, (SA Southern) Ndebele 8%, Northern Sotho 12%, Swati 4%, Tsonga 4%, Tswana 12%, Venda 28%, Zimbabwean Ndebele 4% and Zulu 4% of the total number of students.

3.4 Corpora in translator training

During the theoretical part of the course, students are familiarised with the different types of corpora, how they are compiled, what their possible uses are, etc. During the hands-on workshop sessions, students get the opportunity to apply their theoretical knowledge by building DIY Web corpora in their TL on topics such as HIV/AIDS, education, politics, etc. Students are also shown various sites that contain parallel texts in the official SA languages, and are given access to UP's large general corpora (cf. Table 1 earlier).[3]

When working with bi-/multilingual comparable corpora, students are made aware that when the sizes of English and/or Afrikaans corpora are compared with

3. See De Schryver (2002) for a discussion on African language parallel texts available on the Web. Note, however, that since the publication of this 2002 article, many more parallel texts in the official SA languages have become available on the Web.

that of Bantu language corpora, and when the sizes of corpora of conjunctively and disjunctively written Bantu languages are compared with one another, comparing the number of running words will not give an accurate representation of comparable size. For example, because of the difference in the writing systems of English and Zulu, a Zulu word such as **akakayijwayeli** corresponds to an English sentence consisting of the seven words 'he is not used to it yet'. (Cf. Gauton & De Schryver 2004: 153 and Prinsloo & De Schryver 2002).

During the workshop sessions, the corpora are then used to train students in the skills as discussed in Subsections 3.4.1 to 3.4.3.

3.4.1 *Mining for possible translation equivalents*

Students are trained in how to mine for possible translation equivalents for SL terms that are not lexicalised (in a standardised form) in the TL and that cannot therefore be found in any of the standard sources on the language. Students are taught how to obtain terminology in their TL by querying existing TL texts with (a) *WordSmith Tools* (Scott 1999) and (b) *ParaConc* (Barlow 2003). For example, students build their own DIY HIV/AIDS corpus in their TL, as well as a comparable SL corpus, and then use *WordSmith Tools* to semi-automatically extract relevant term candidates. See the example below of a list of potential Afrikaans translation equivalents obtained in this manner[4] (see Table 2).

Table 2. Afrikaans translation equivalents for HIV/AIDS SL terminology obtained with *WordSmith Tools*

N	Word	Keyness
1	VIGS	3,453.60
2	HIV	2,839.90
3	MIV	2,621.20
4	VIRUS	1,093.60
5	SEKS	665.4
6	GEÏNFEKTEER	597.4
7	OPVOEDERS	560.2
8	LEERDERS	537.8
9	BEHANDELING	513.8
10	INFEKSIE	504.3
11	KONDOOM	481.7
12	BLOED	386.3
13	SIEKTES	377.8
14	RETROVIRALE	368.5

4. Examples cited in this section are from the classroom activities of my 2004/2005 students.

Another workshop activity performed by the students is to make use of *ParaConc* to mine for possible translation equivalents by first accessing parallel texts in their source and target language combination on the Web, and then utilizing *ParaConc* to align these texts. See the *ParaConc* screenshot in this regard, illustrating aligned English-Zulu parallel texts dealing with the *South African Qualifications Authority Act (Act 58 of 1995)* (see Fig. 1).

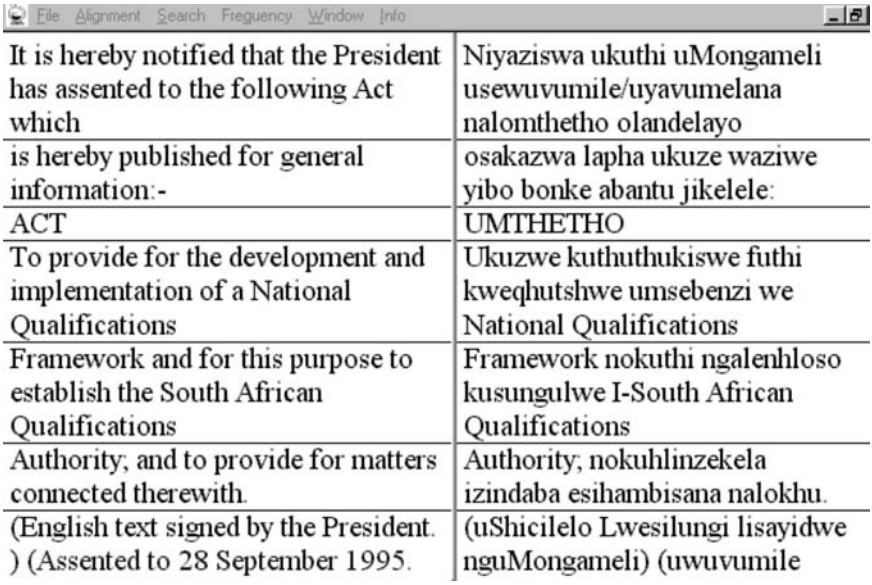

Figure 1. Screenshot of aligned English-Zulu parallel texts in *ParaConc*

For a full account of the methodology that can be followed in identifying possible African language term equivalents in (a) comparable corpora using *WordSmith Tools* and (b) in parallel corpora by utilising *ParaConc*, see Gauton & De Schryver (2004).

3.4.2 *Gaining insight into term formation strategies*

Students are trained in how to scrutinise existing TL translations in order to gain insight into term formation strategies. By studying parallel corpora and scrutinising the term formation strategies used by professional translators, trainee translators can gain insight into:

– the various term formation strategies available in their TL, and
– the preferred strategies for translating terminology into their TL.

See again Fig. 1 for the format in which TL translations can be presented for the purpose of identifying translators' strategies. A glossary such as given in Section 3.4.3 below can also be used for this purpose.

For a full exposition of the various (preferred) term formation strategies in the official SA languages, particularly in the nine official Bantu languages and in Afrikaans (English usually being the SL), the reader is referred to Gauton et al. (2003), Gauton et al. (forthcoming) and Mabasa (2005).

3.4.3 *Recycling existing translations*

Students are trained in how to recycle existing translations (their own translations and/or suitable parallel texts available on, for example, the Web) with the aid of a translation memory tool. By making use of the translation memory (TM) software programme *Déjà Vu X* (*DVX*), students are trained in how to reuse existing translations, whether their own translation work, or existing parallel texts (culled from the Internet) which are then aligned and fed into the TM. Students are also taught how to use this software to extract a glossary of the source and target language terminology used in a particular translation project. See for example the following extract from a Venda glossary based on the translation of a ST entitled *Cache and Caching Techniques*:

Table 3. Venda glossary extract

SL Word	TL Equivalent	Back Translation
cache	khetshe	cache
caching techniques	dzithekheniki dza u vhewa kha khomphiyutha | thekheniki dza kukhetshele	techniques of saving in/on the computer | technique of to cache
information	ndivhiso | mafhungo	information | news
memory	muhumbulo | memori	memory
web	webe	web

Due to space constraints, it is not feasible to give complete examples of students' work here, but see the *DVX* screenshot in Fig. 2, illustrating the penultimate step in producing a translation from English into Zulu. (This is the so-called 'pre-translation' function which allows the translator to leverage the content of his/her databases, e.g., the translation memory and the terminology database, against the source file).

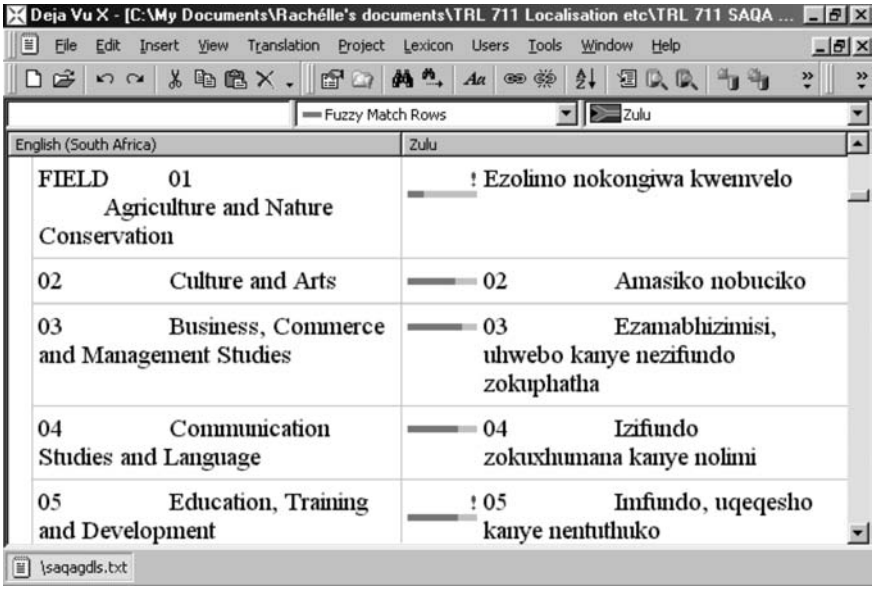

Figure 2. Screenshot of a translation from English to Zulu being done in *Déjà Vu X* (*DVX*)

At the end of the course, students have to complete a practical translation of a technical text in the computer laboratory under examination conditions and within a set timeframe, utilising the various electronic translation resources and tools that were covered in the hands-on workshop sessions. The results achieved since the inception of this course have been extremely gratifying. The 2004 student group obtained a class average of 65%, the 2005 intake a class average of 75% and the 2006 intake an average of 76%. Generally, students tend to approach the course with a certain amount of trepidation, mainly because most of the students taking this course are not that familiar with computer technology and those that are, are not familiar with the application of technology to the task of translation. However, despite (or perhaps because of) these factors, we have found the students to be totally committed to mastering these new skills, as they realise that an ability to use electronic translation resources and translator's tools are essential skills for the modern translator and that this gives them a competitive advantage when entering the job market.

4. The changing (and changed) world of the modern translator

Writing about the challenges and opportunities that localisation and the increasing availability of computer based translation tools present to the translator, Nancy A.

Locke, translator, localisation educator and writer, succeeds in getting to the heart of how the world of the modern translator has changed:

> Not long ago, a translator could make a living armed with a shelf of reference books, a sharp pencil, a passion for language and an inquisitive mind. [. . .] But technology and a rapidly globalizing marketplace are changing the way translators work and, for those who relish challenges and can adapt, radically increasing career options and creating a potentially profitable business sector. IT-savvy language professionals, language-savvy IT professionals and linguistically agile project managers with a sophisticated sense of the global economy are in demand as companies attempt to enter markets far from home. [. . .] And that's just the beginning. Translation-specific tools – the refined offspring of automatic translation research begun during the Cold War – are increasing in use. Computer-assisted translation and translation memory software and online or electronic terminologies are a far cry from leafing through the pages of a well worn dictionary.
>
> (The Globe and Mail 2003)

Murray-Smit (2003) writes as follows from the perspective of a South African translator, translating between English and Afrikaans:

> The idea of a translator not using computer-assisted methods of translation is almost unthinkable. The advantage that the computer gives the modern translator over his pen-and-paper predecessor lies in the translator having to spend less and less time on repetitive little tasks and having more time available for the actual hard creative work of translation. Some of these utilities are so widely used that translators often do not even realise that they are doing "computer-assisted" translation. Take a spell-checker, for example – although not infallible, it makes the translator's task a hundred times easier. Or think of a CD-ROM dictionary – thanks to this technology, a translator can source definitions, translations, synonyms and encyclopaedic information in seconds.

Not only has there been rapid change in the way that the modern professional translator works, but translation technology itself is changing even more rapidly. As Hutchins et al. (2005: 3) observe in the eleventh edition of the Compendium of Translation Software:

> New software products for automatic translation and for supporting translation work are appearing almost every week; there is constant updating for new versions of operating systems, and more and more companies are involved in the field.

Quah (2006: 20–21) agrees with Hutchins et al. and points out that the pace of change in the development of translation technology is so extremely rapid, that what is current today may be outdated tomorrow.

Yuste Rodrigo (2002: 35–36) refers to the 'rapidly evolving translation profession' and quotes Shreve as describing global language market needs as 'an evolution in fast-forward'. Regarding the value of incorporating translation

memory software into the daily practice of the professional translator, Benis (2003: 29) states:

> The transition to using translation memory may seem strange, but in many cases it is just as logical and beneficial as those we have made in the past from pen to keyboard and word processor. But a better analogy would be to consider how we used to do our research, making our way from one specialist consultant or library to another and filing away our findings. Now we can find almost anything we need by simply browsing from one website to another. Translation memory brings the fruits of that research all together and delivers it to us automatically as needed in a single place – our computer screens.

Already in 2001, Yuste Rodrigo discussed the increasing need in the translation market for what she refers to as the 'multitasked translator'. She summarises the profile of the modern, 'multitasked', translator as follows:

> In order to ensure optimum work prospects and client satisfaction, translators now play a number of varied tasks, which would not correspond to the traditional translator's role. Far from producing translation work only, they are expected to keep pace of information technology advancements applied to their profession, such as the latest translation technology, Internet resources for translators, etc. This enables them not only to optimise their daily linguistic activities, e.g., computer-aided terminology research and management, but also to market their translation skills, be a key team player or manage a translation project in a global setting.

Clearly the translation profession itself and the skills required by the modern professional translator are vastly different from what they were even a decade ago. It is therefore of the utmost importance that the professional African language translator does not get overtaken and left behind by the rapidly evolving translation market, with its increased emphasis on the use of translation technology (which is itself subject to continuous change and development). Having said that, SA language practitioners are only too aware of the limitations of being SA Bantu language translators, working in Africa, where the necessary infrastructure is often lacking, limiting translators' access to language technology and resources.

5. The need for language resources for the African language translator in the South African context

As indicated elsewhere (Gauton 2006) one only has to study existing translations (particularly in technical domains) to realise what the consequences are of SA Bantu language translators not having access to the necessary language technology and language resources.

Having to work in an environment where one cannot rely on the availability of standardised terminology, creates various problems for these translators. A case in point is the Zulu translations of:

a. the user interface (UI) of the Microsoft operating system Windows XP for which I was project leader; language manager responsible for creating and maintaining the language style guide and managing the terminology to ensure consistency; as well as quality controller of specifically the grammatical and linguistic correctness of the translations (cf. Gauton, 2005); and
b. the Sony Ericsson user guides for the T310, T610 and Z520i mobile phones.

A very high level of inconsistency in terminology used for the same English SL concepts can be found not only between these translations, but also within the same translation, and in the case of the Windows XP translation, within the work of the same translator. This type of situation invariably results when there is more than one translator working on a specific project, with each translator working in isolation (so to speak); i.e., working without the use of computer assisted translation (CAT) tools such as translation memory (TM) tools, terminology tools and software localisation tools, coupled with a pooling and sharing of resources within the project. However, as indicated, even when only one translator is involved and doing large amounts of translation (which is usually the case in localisation projects such as these) without the use of a translation memory tool; this usually results in terminological inconsistencies within such a translator's work. See the following representative examples (culled from the Zulu translations mentioned earlier) illustrating this point (see Table 4).

Table 4. Inconsistent use of terminology in the Zulu translations of the Windows XP UI (before consistency checking) and the published Sony Ericsson user guides

Source Language (English) Term	**Translation equivalents culled from the Zulu translations of the Windows XP UI (before consistency checking) and the published Sony Ericsson user guides**
network (*n*)	umphambo; inethiwekhi; ukuxhumana; uxhumano
e-mail (*n*)	i-imeyli; i-e-mail; i-emeyili; iposi le-elekthroniki
Settings	izinhlelo; okokuhlela; ukuhlelwa; uhlelo; izimiso
set (*v*)	hlela; misa; setha
Shortcut	indlela enqamulayo; unqamulelo; ukunqamulela; indlela emfushane; ishothikhathi; ushothikhathi
Phonebook	ibhuku locingo; ibhuku lezingcingo; incwadi yezingcingo; ibhuku lefoni; ifonibhuku
Memory	inkumbulo; imemori; isiqophi
Password	igama lokungena; igama lokuvunyelwa ukudlula; igama lokudlula; iphasiwedi
Wizard	umeluleki; umbuzimholi; iseluleki; iwizadi; umthakathi

As can be seen from Table 4, the level of inconsistency regarding the translation equivalents used for the SL terminology is unacceptably high.

Furthermore, by not being aware of terminology that has perhaps already been coined by other translators, and/or by not getting the benefit of other translators' experience and insight, the unwary translator may end up coining culturally unacceptable target language equivalents such as the term **umthakathi** 'witch' for the source language term 'wizard'. As pointed out elsewhere (Gauton 2005) the term **umthakathi** has extremely negative connotations in the Zulu culture, and is not a suitable translation equivalent for the English concept 'wizard'; particularly as used in the domain of computer studies. Whereas in the Western context a wizard is a wise and magical imaginary fairytale character that is often benign (unless of course he is characterised as an 'evil wizard'), quite the opposite applies in Zulu culture. Within Zulu culture, witches and wizards are evil creatures that practice witchcraft, are intent on doing only harm to their fellow man, and who are shunned and avoided by all wherever possible. In recent years in South Africa, a significant number of people have been persecuted, ostracised and even killed on suspicion of being witches or wizards practising witchcraft. Thus it is clearly inappropriate and culturally unacceptable to use the term **umthakathi** 'witch' to signify an interactive computer programme (i.e. a 'wizard') that fulfils the function of helping and guiding the user through complex procedures.

Another serious drawback that results from translators not having access to shared language resources, is that gross mistranslations can result, e.g., the translation of 'default' as **-nephutha / -yiphutha** '(something) that is faulty / a fault / wrong'. It must be borne in mind that the translators involved in these localisation projects are usually highly experienced translators, which underscores the necessity for translators to be able to access the kinds of language technology and resources mentioned in this chapter. What is also needed in this regard is translator training of the type that is provided at the University of Pretoria (UP), which will equip translators to take advantage of such electronic translation resources and translators' tools.

6. Conclusion

In this chapter it was shown how corpora are used with great success in the training of African language translators at the University of Pretoria, South Africa. This is the only such translator training programme in the country that I am aware of, particularly as far as the use of Bantu language corpora in the training of translators working into these languages, are concerned. This gives graduates from the UP programmes a definite competitive advantage over their peers when applying for translation positions and/or undertaking freelance translation work. Further-

more, after successful completion of this course, students working with the African languages appreciate that these languages can in fact be used as high function languages, despite there being very little or no (standardised) terminology readily available in order to produce technical translations of this kind.

In conclusion, in cooperation with the UP students, we intend to establish an (interactive) online database containing student outputs in the form of glossaries/term lists. In this way it would be possible to receive input from interested parties regarding the suitability/acceptability of the various terms and also to provide a service to other translators and language workers. In time, such a multilingual student site could become a very large, comprehensive and valuable language resource that will contribute not only to the development and elaboration of the African languages as technical languages, but also towards the standardisation of these languages.

References

Barlow, M. (2003) *ParaConc: A concordancer for parallel texts*. Houston, TX: Athelstan. See for this software also http://www.athel.com

Benis, M. (2003) Much more than memories. In *ITI bulletin*. http://www.atril.com/docs/Benis-ITI-DVX.pdf

Bowker, L. (1998) Using specialised monolingual native-language corpora as a translation resource: A pilot study. In *Meta*, 43 (4), 1–21.

Bowker, L. (2000) Towards a methodology for exploiting specialised target language corpora as translation resources. In *International Journal of Corpus Linguistics*, 5 (1), 17–52.

Bowker, L. (2002) Working together: A collaborative approach to DIY corpora. In Yuste Rodrigo, E. (ed.) *Language resources for translation work and research*, LREC 2002 Workshop Proceedings, Las Palmas de Gran Canaria, Spain, 29–32. http://www.ifi.unizh.ch/cl/yuste/postworkshop/postworkshop.htm

Calzada Pérez, M. (2005) Applying translation theory in teaching. In *New Voices in Translation Studies*, 1, 1–11. http://www.iatis.org/newvoices/current.htm

Déjà Vu X Professional. Version 7.0.238. Copyright © 1993–2003. ATRIL Language Engineering, SL.

De Schryver, G.-M. (2002) Web for/as Corpus. A Perspective for the African Languages. In *Nordic Journal of African Studies*, 11(3), pp. 266–282.

Fictumova, J. (2004) Technology-enhanced Translator Training. In Yuste Rodrigo, E. (ed.) COLING 2004 Workshop #3. *Second International Workshop on Language Resources for Translation Work, Research and Training*. The University of Geneva, Geneva, Switzerland, 28th August 2004, 31–36. http://www.ifi.unizh.ch/cl/yuste/lr4trans-2/wks_papers.html

Gauton, R., E. Taljard and G-M. de Schryver (2003) Towards Strategies for Translating Terminology into all South African Languages: A Corpus-based Approach. In G-M. de Schryver (ed.), TAMA 2003, South Africa. *Terminology in Advanced Management Applications*. 6th International TAMA Conference: Conference Proceedings. "Multilingual Knowledge and Technology Transfer". Pretoria: (SF) 2 Press, 81–88.

Gauton, R. & G-M. de Schryver (2004) Translating technical texts into Zulu with the aid of multilingual and/or parallel corpora. In *Language Matters, Studies in the Languages of Southern Africa*, 35(1) (Special issue: Corpus-based Translation Studies: Research and applications), 148–161.

Gauton, R. (2005) The anatomy of a localisation project in an African language – translating Windows XP into Zulu. In *South African Journal of African Languages*, 25 (2): 124–140.

Gauton, R. (2006) Share and Share Alike – Developing Language Resources for African Language Translators. In Roux, J. C. and S. E. Bosch (eds) *Proceedings of the Workshop on Networking the Development of Language Resources for African Languages*. LREC 2006 – Fifth International Conference on Language Resources and Evaluation. Monday, 22nd May 2006. Magazzini del Cotone Conference Centre, Genoa, Italy. Paris: ELRA / ELDA: 34–37.

Gauton, R., E. Taljard, T. A. Mabasa and L. F. Netshitomboni, (forthcoming). Translating technical (LSP) texts into the official South African languages: a corpus-based investigation of translators' strategies. (Submitted to *Language Matters*).

Google [South African localised version] http://www.google.co.za/

Goussard-Kunz, I. M. (2003) *Facilitating African language translation in the South African Department of Defence*. Pretoria, University of South Africa. Unpublished MA dissertation.

Hutchins, J., W. Hartmann, and E. Ito (2005) *Compendium of Translation Software*. http://ourworld.compuserve.com/homepages/WJHutchins/Compendium-11.pdf

Internet Archive Wayback Machine. http://web.archive.org/collections/web.html

Izwaini, S. (2003) Building specialised corpora for translation studies. In *Proceedings of the pre-conference workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*. Corpus Linguistics 2003, Lancaster University, UK, 28th–31st March 2003, 17–25. http://www.coli.uni-sb.de/muco03/izwaini.pdf

Kruger, A. (2004) *Language Matters*, Studies in the Languages of Southern Africa, 35(1) (Special issue: Corpus-based Translation Studies: Research and applications), 1–5.

Laviosa, S. (2003) Corpora and the translator. In Somers, H. (ed.) *Computers and translation: A translator's guide*. Amsterdam/Philadelphia: John Benjamins, 100–112.

Laviosa, S. (2004) Corpus-based translation studies: Where does it come from? Where is it going? In *Language Matters, Studies in the Languages of Southern Africa*, 35(1) (Special issue: Corpus-based Translation Studies: Research and applications), 6–27.

Locke, N. A. (2003) Finding the right words. Special to *The Globe and Mail*. Online. http://www.theglobeandmail.com/servlet/ArticleNews/TPStory/LAC/20030723/TRANSLATOR23

Mabasa, T. A. (2005) Translation equivalents for health/medical terminology in Xitsonga. Pretoria: University of Pretoria. Unpublished MA dissertation.

Maia, B. (2002) Corpora for Terminology Extraction – the Differing Perspectives and Objectives of Researchers, Teachers and Language Services Providers. In Yuste Rodrigo, E. (ed.) *Language resources for translation work and research, LREC 2002 Workshop Proceedings*, Las Palmas de Gran Canaria, Spain, 25–28. http://www.ifi.unizh.ch/cl/yuste/postworkshop/download.html

Maia, B. (2003a) What are comparable corpora? In *Proceedings of the pre-conference workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*. Corpus Linguistics 2003, Lancaster University, UK, 28th–31st March 2003, 27–34. http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

Maia, B. (2003b) The pedagogical and linguistic research implications of the GC to online parallel and comparable corpora. In J. J. Almeida (Ed.), CP3A – Copora Paralelos, Aplicações e Algoritmos Associados, Braga, 13 de Maio de 2003. Braga: Universidade do Minho, Portugal, 31–32. http://poloclup.linguateca.pt/docs/index.html

McEnery, A. and Z. Xiao (2007) Parallel and comparable corpora: What is happening? In M. Rogers & G. Anderman (eds.) *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters, 18–31. http://www.lancs.ac.uk/postgrad/xiaoz/publications.htm

Murray-Smit, S. (2003) *Introduction to computer aided translation*. http://www.leuce.com/translate/computeraided.html

Oliver, R. & Atmore, A. (1972) *Africa Since 1800*. Cambridge: Cambridge University Press.

Orwell, G. (1983). *Animal Farm*. Great Britain: Penguin Books.

Prinsloo, D. J. and G-M. De Schryver (2002) Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages. In *Nordic Journal of African Studies*, 1(2), 249–265.

Quah, C. K. (2006) *Translation and Technology*. Palgrave Textbooks in Translating & Interpreting. http://www.deastore.com/pdf/palgrave%20_01_2006/1403918325.pdf

Raidt, E. H. *n.d. Afrikaans – van Afrika?* Available online at http://web.archive.org/web/20050320174912/http://general.rau.ac.za/aambeeld/june1996/afrikaansvanafrika.htm

Scott, M. (1999) *WordSmith Tools, v. 3*. Oxford: Oxford University Press. See for this software also http://www.lexically.net/wordsmith/index.html

Sony Ericsson Mobile Phone T310. 2003. ©Sony Ericsson Mobile Communications AB, 2003.

Sony Ericsson Mobile Phone T610. 2003. ©Sony Ericsson Mobile Communications AB, 2002.

Sony Ericsson Z520i User Guide. ©Sony Ericsson Mobile Communications AB, 2005.

South African Qualifications Authority Act (Act 58 of 1995). Online. Available from http://www.saqa.org.za/publications/legsregs/index.htm#legs. http://web.archive.org/web/*/http://www.saqa.org.za/publications/legsregs/index.htm#legs

Tiayon, C. (2004) Corpora in translation teaching and learning. *Language Matters, Studies in the Languages of Southern Africa*, 35(1) (Special issue: Corpus-based Translation Studies: Research and applications), 119–132.

Varantola, K. (2002) Disposable corpora as intelligent tools in translation. In Tagnin, S.E.O. (ed.) *Cadernos de Tradução: Corpora e Tradução*. Florianópolis, Brasil: NUT 1/9, 71–189. http://www.cadernos.ufsc.br/online/9/krista.htm

Visual Thesaurus. http://www.visualthesaurus.com/

Yuste Rodrigo, E. (2002) Making MT Commonplace in Translation Training Curricula – Too Many Misconceptions, So Much Potential! In *Proceedings of the Machine Translation Summit VIII*, Santiago de Compostela, Spain. http://www.eamt.org/summitVIII/papers/yuste-2.pdf

Zanettin, F. (1998) Bilingual comparable corpora and the training of translators. In *Meta*, 43 (4), 616–630.

Zanettin, F. (2002) DIY Corpora: The WWW and the translator. In B. Maia, J. Haller and M. Ulrych (eds.), *Training the languae services provider for the new millennium*. Porto: Facultade de Letras, Universidade do Porto, Portugal, 239–248.

CHAPTER 7

CAT tools in international organisations

Lessons learnt from the experience of the Languages Service of the United Nations Office at Geneva

Marie-Josée de Saint Robert[1]
United Nations Office at Geneva

The language staff at the United Nations Office at Geneva (UNOG) has a very selective attitude towards language technologies despite the fact that these technologies are widely spread in the work environment of translators. Tests and pilot projects with computer-assisted translation software have been conducted over the past five years at UNOG and have amply shown that such software is neither a source of improvement of quality nor a source of improvement of quantity in translation. Obstacles to efficiency gains that have been identified prior to the introduction of CAT remained the same as those identified after its introduction. New obstacles also appeared with the introduction of CAT in the work of language staff of an international organisation that point to the following conclusion: the usefulness of work habit changes may not be found in the area in which the changes occur but in other somewhat unexpected areas. In the case of translation in international organisations, as translation is not an isolated activity, synergies with other, sometimes far related business processes are required.

1. Introduction

The United Nations may seem to be the perfect environment for the deployment of language technologies yet, surprisingly enough, translators and language staff in general do not rely heavily on tools. Technological innovations have to be user-friendly and time-saving to convince translators to use them. Between 2002 and 2004, a CAT test conducted at the United Nations Office at Geneva concluded that

1. The opinions expressed in this article are my own. This paper is an updated and expanded version of a paper presented for the First International Workshop on Language Resources for Translation Work and Research, chaired by E. Yuste Rodrigo at the LREC 2002 (Third International Conference on Language Resources and Evaluation). Las Palmas de Gran Canaria, Spain, in May 2002.

a corpus-driven translation tool would be more appropriate for translators than a sentence-driven tool such as TRADOS. A small number of individual expert MultiTrans licenses were bought in 2005 together with less powerful web licenses accessible to all translators. In 2007, the results of a new pilot project on MultiTrans showed that the CAT tool received low satisfaction scores from translators as their expectations concerning the tool did not coincide with what the tool was providing them with: their idea of having readily available automatic alignment of two language versions of the same document whenever desirable was not matched by a product which proposes a text hyperlinked to all identical or almost identical segments that appear in previously translated documents and that can be selected for direct insertion into the translation draft, in any given pair of languages. What may appear as settling for less may be linked with the very nature of the translation process in multilateral diplomatic settings where linguistic and pragmatic constraints on the one hand, and technical constraints on the other, play an important role.

2. Linguistic and pragmatic constraints

Several linguistic constraints are obstacles to the straightforward application of language technologies to translation work. Some are quite obvious, while others are specific to international organisations.

2.1 Word choice

Translation cannot be reduced to the mechanical substitution of one set of terms in one language by a similar set in another language. Subjects discussed at the United Nations are ground breaking and established terms may not be readily available in all official languages. The search for correct terms to designate new concepts is a long and often non-linear process, which varies from one language combination to another and crucially depends on the translator's mastery of term history. As a result, in all language combinations, a translator working with a CAT tool may be confronted with several alternative equivalents that he has to spend time to evaluate whereas a translator working without a CAT tool may know the ultimately established equivalents he needs right away.

2.1.1 *Level of language adequacy*

Writing, and therefore translating, for the United Nations means addressing oneself to a wide variety of document users: diplomats, scientists, media specialists, law experts, administrative officers. The linguistic distinctiveness of the United Nations resides in the level of language used, which has to be acceptable to all

readers. The sentence starting with (1) should not be translated into French by (2) no matter how common that phrase is but by (3):

(1) the report shows

(2) le rapport montre que

(3) il ressort du rapport que

Similarly the structure in (4) should be avoided and replaced by (5) or (6) but practice may be quite inconsistent in documents:

(4) the law says

(5) the law provides

(6) the law stipulates

No matter how carefully texts are prepared and processed, variation slips in that limits the usefulness of CAT tools, which would propose ad infinitum less acceptable constructions found in document corpora. Translators in haste may only keep the first instance proposed by CAT tools and disregard other instances of the same phrases found subsequently, knowing that phrases in (2) and (4) look correct from a grammatical point of view. In (2) the use with a non-animate subject of a verb in French which requires an animate subject is non-grammatical, though the mistake is quite common in every day speech; in (4) usage prevails though (5) and (6) are recognised by native speakers of English as preferred in a United Nations context. The same would apply in this particular case in French.

2.1.2 *Resistance to standardisation*

Translations serve the purpose of a specific communication need and should not be considered as models for translators to replicate across the board. Such is also the case for terminology in any target language. Mere electronic bilingual dictionaries or glossaries cannot satisfactorily capture variation, not only in the original language but also in the target language, if based upon the assumption that a notion corresponds to a term in English and one or several terms in French, for instance. Names given to human rights are a case in point. A terminologist would very happily collect the names of all rights, starting with the right to food, to adequate housing, and to education, while a translator would resent it. Such rights are indeed referred to under different names by different speakers, and a too rigid list of rights would miss the needed subtleties while discussions are still under way. Should “adequate housing” be rendered in French by “logement convenable,” “logement adéquat,” “logement suffisant,” or “logement satisfaisant,” all four equivalents being found in United Nations legal instruments or resolutions, and not by “bonnes conditions de logement” or “se loger convenablement” when the context allows or requires it? Translators want to preserve flexibility, when

present-day translation systems propagate rigidity and, as a lurking consequence, poverty of style and vocabulary. For Fernando Peral (2002), a former translator at the International Labour Organisation:

> The main operational problems of 'semi-automatic' translation [i.e., translation with the help of translation memory systems] are linked to the quality of the output and to a process of 'de-training' of the translator, who becomes less and less used to the mental process of searching for proper solutions in terms of functional equivalence and relies more and more on the machine's decisions, which inevitably affects professional development and job satisfaction.

Another type of resistance to terminology standardisation comes from the fact that document drafters do not consult the United Nations terminology database when they refer to organisations, legal instruments or political, judicial and administrative entities. They tend to coin their own translated denominations that are one-time free renderings with no official status. Denominations specific to Member States fall into that category as Member States do not adopt names for their official nomenclature in all the six languages of the United Nations.[2]

2.1.3 *Semantic adequacy*

Literal translation may lead to semantic impropriety. The correct rendering in French of the English phrase (7) is not (8) but (9):

(7) abusive sexual practices that may affect very young girls

(8) pratiques sexuelles abusives qui peuvent affecter les très jeunes filles

(9) pratiques sexuelles dont peuvent être victimes les très jeunes filles

The sentence in (8) is incorrect from the semantic point of view[3] but is grammatically correct. Maybe more accurate information on what CAT systems do is needed. In language contact situations communication may suffer from literal translation that also lead to semantic inappropriateness. In sentences (11) and (14) the message contained in sentences (10) and (13) is lost in English as the translator rendered slavishly into English the original French text. The revised versions are given in (12) and (15):[4]

(10) Les ressortissants d'Etats tiers qui résident légalement en Belgique et qui sont dans une situation transfrontalière sont également visés

2. Official languages of the United Nations are: Arabic, Chinese, English, French, Russian and Spanish.

3. Sentence (8) suggests that sexual practices are divided into two categories: abusive and non-abusive, but this dichotomy is not applicable to very young girls.

4. Examples discussed in Hobbs (2008).

(11) nationals of third States residing legally in Belgium and in a transfrontier situation are also covered

(12) nationals of third States who reside legally in Belgium but work across the border are also covered

(13) En tant que plate-forme de concertation, la nouvelle Commission nationale des droits de l'enfant est un point de rencontre, de coordination, d'échange d'idées avec les gens de terrain, un endroit fertile pour donner des impulsions à la politique des droits de l'enfant en Belgique.

(14) As a platform for dialogue, the new National Commission for the Rights of the Child is a meeting and coordination point, and a point for exchanging ideas with people in the field, a ripe breeding-ground for giving momentum to children's rights policy in Belgium.

(15) As a platform for dialogue, the new National Commission for the Rights of the Child is a forum for meetings, coordination and exchanges of ideas with people in the field, a fertile ground for ideas to boost children's rights policy in Belgium.

The quality enhancement strategy followed by translators requires that previously translated sentences easily accessible via CAT are carefully scrutinised. Revisers expect to receive explanations on why a previously translated sentence provided by the tool has not been accepted, why revisions have been made. The need to codify and re-codify practice is permanent. The state of alertness required for translating United Nations documents may not be compatible with over-reliance on precedents that may be fostered by CAT tools.

2.2 Linguistic insecurity

Document originators at the United Nations are nationals from over a hundred and twenty countries. In most cases their native language is not one of the six official languages of the Organisation, and document drafters erroneously think they have to use English, which may prevent them from using their main language, even when it is an official language, and produce better originals. Documents may also be submitted to the United Nations by officials or experts working for any of the 193 Member States that do not have either any of the official languages of the Organisation as their main language. Syntactic, semantic and morphological mistakes are therefore not rare in original documents, and in most cases only editors[5] and translators detect mistakes and rebuild faulty sentences that drafters have

5. At the United Nations, editors check original texts against editorial rules and official terminology prior to their being processed by reference assistants and translators.

left in original text. Only they are required to work in their native language that is one of the official languages. Due to lack of resources at the United Nations, only a small portion of all documents is edited prior to being translated (e.g., documents prepared by the Human Rights Council and Treaty bodies). Translators consequently do act as filters for grammatical correctness and language consistency as they work on the texts to be translated. As a result, they often improve original texts whenever the drafters or submitting officers accept their changes in the original documents. A translation memory processing straightforwardly a document to be translated prior to the perusal of a translator may not detect inappropriate use of terms or syntactic errors in the original language as errors done by non-natives are mostly unpredictable, given the number of native languages spoken at the United Nations[6] and the wide differences in language mastery from one drafter to the other. Even when an automatic term-checking system is appended to the translation memory, it may not be as efficient as a human eye either. The fear therefore is that a computer-assisted translation system may add more mistakes to the original ones, which will then be even harder to detect and correct.

2.3 Stylistic interferences

Translators have not only to follow the original text but also apply the stylistic rules of the language into which they translate. For instance, among writing styles one can mention the fact that repetitious words are not considered as poor style in English but are definitely considered poor style in French. The English sentence (16) presents a repetition of the word 'aircraft' which the French rendering in (17) would avoid:

(16) the shooting down of civil aircraft by a military aircraft

(17) la destruction d'aéronefs civils par un appareil militaire

In translating back the French sentence in (17) into English, the English-language translator has to reduce the number of stylistic variants that French requires. Hobbs (2008) shows how required usage is acquired by United Nations translators, a process based on daily practice, coaching and revising over the years. It takes decades to become an accomplished linguist at the United Nations and some translators never reach that stage. Languages outside the Indo-European family of languages, such as Arabic and Chinese, need to develop stricter quality control

6. As far as United Nations staff members are concerned, over 120 nationalities are represented, which gives an order of magnitude of linguistic variation and of the multilingual speech community within the United Nations Secretariat.

mechanisms as the risk is real to follow the original languages to the detriment of legibility or communication effectiveness in the target languages.

2.4 Functional adequacy

Each committee or body has specific ways of expressing an idea in order to reach a consensus within its respective audience or circle. Underlying references to protagonists, former meetings, and earlier decisions discussed by committee members but not explicitly mentioned in the text play an important role in editing and translation. Sometimes the reasoning of a *rapporteur*, a speaker or an author, or an amalgam of lengthy sentences couched in simple terms that are perfectly unintelligible to the outsider, i.e., someone who has not participated from the beginning in the discussions, has to be left untouched in the original. Acceptability of a translated text does not come solely from its grammatical and semantic well-formedness. It must also be appropriate within the United Nations context. A translated text must, like its original, follow a highly standardised path: it must convey the impression of having been written by a long-time member, perfectly familiar with the background in which the text has been drafted, even if it is deliberately vague or obscure. In fact most United Nations texts cannot be interpreted without prior knowledge of the particular political framework in which they appear. The sociopolitical motivation and rationale behind a text are part of the unwritten constraints imposed on communicative competence at the United Nations. Developments in artificial intelligence are not perceived to have reached this level of refinement. As Peral (2002) puts it:

> translation is based on finding 'functional equivalences' that require linguistic, intertextual, psychological and narrative competence; only human beings are capable of determining 'functional equivalences'; productivity in translation is therefore intrinsically linked to the capacity of the translator to find the adequate functional equivalence, i.e., it is based on the quality of the translator.

Occasionally functional adequacy does not coincide with grammaticality, as shown in Jastrab (1984), since political (functional) adequacy has priority over morphology and syntax. Translators and revisers are confronted with the challenge of conveying non-grammatical structures that result from negotiations from one language into another. Revising the content of translation memories or improving text corpora has to be resisted as it may lead to unwarranted changes that may have political consequences.

These constraints conflict with the concept of translation reuse for translation purposes on which most commercially available alignment tools and translation memory systems are based, especially when document traceability (i.e., the

capacity of retrieving the complete document from which a sentence is extracted by the translation memory system) is not guaranteed.

3. Technical constraints

Quality requirements are not always met in translated documents for technical reasons.

3.1 Time constraints

Non-respect of deadlines for document submission results in not allowing translation to be performed in the required conditions. Feeding translation memories with texts that have not been properly revised for lack of time appears to be useless, even when such texts are considered as basic texts in a given area. The underlying assumption is that basic texts can be improved over and over as they are cited in other texts, but no one can guarantee that it will indeed be the case, as translators are more and more required to work under emergency conditions, and as a faulty sentence may well be intentionally faulty.

This explains why most documents are not considered by translators as authoritative sources for official denominations either in the source or in the target languages. Most official names of international and national organisations, bodies and institutions are referred to under several names in various documents and sometimes even within the same document. Alignment tools and translation memories that would provide precedents in two languages to translators might perpetuate the number of variants and confusion rather than helping translators to use the right equivalent, unless quality assessment is performed, which is a rather slow and uneconomical process looked down upon in an era of search for productivity gains. The problem is even more complex when it comes to designating a body whose name may be official in one or two languages but not in other languages. Chances are that transliterated names in English, French or Spanish rarely reappear again under the same denomination unless a rather time-consuming compilation is done to provide the best possible equivalents across official languages that would be used by translators. Yet as George Steiner (1975) rightly puts it: "Languages appear to be much more resistant than originally expected to rationalization, as well as to the benefits of homogeneity and technical formalization." Languages resist because human beings resist.

3.2 Digital divides

Other technical constraints make the use of CAT systems difficult: (1) non-submission of documents in electronic form: many documents are submitted on paper with last minute manual corrections – linguistic insecurity or a changing appreciation of political requirements being the main causes of last minute changes; (2) non-availability of reference corpora: some official references may exist in one or two languages, and have to be translated into the other languages – reference documents that are considered as authoritative in one language pair may not be so in another, thus the task of building translation memories is labour-intensive, language pair by language pair; (3) scarcity of digitalised language resources in some languages: translators cannot completely switch to ready-made technological innovations – expertise in conventional research means should be kept.

3.3 CAT stumbling-blocks

Progress has to be made with CAT tools in at least two directions:

3.3.1 *The tools themselves*

CAT tools are known to be most effective with repetitive texts. So far, since at the United Nations not all texts are available in electronic form, it is hard to assess the amount of repetition to be able to ascertain whether or not CAT is an efficient tool in this environment. But other issues are still pending before CAT tools prove efficient: (1) statistics provided by the tools: no matter how repetitive a text is, efficiency gains do not match the figures provided by the tools as far as recycling is concerned: at UNOG, the proportion of potentially recyclable vs. actually recycled material was measured with MultiTrans in April 2007 (painstakingly as these figures were not easy to obtain), and the results were as follows: out of the 379,353 words of texts in text bases, 77,090 words were marked as recyclable (20.3%) and only 44,482 (11.7%) were actually recycled. Tools should provide statistics to determine what percentage of the previous texts are actually used and how often; (2) training on the tools: proper training has to be given to translators to make certain they know how to fully utilize the tools that they are given and that they can get ownership over them, while making as many requests for change and improvement to the vendors as necessary and receiving feedback and solutions from them; (3) user friendliness: skills such as problem solving, text understanding, solution evaluating and message reformulating, crucial in the translation process, require access to the full text to be translated and not only sentence by sentence or paragraph by paragraph. Sometimes, even full text and context are not sufficient to reduce ambiguity in a text as translators need to make inferences that are knowledge based and goal oriented; (4) full description of technical requirements and

infrastructure for successful implementation: equipment used in an international organisation has to be compatible with the equipment required by a particular CAT software and conversely CAT software has to respond to the needs of the translators. At UNOG, immediate access to millions of words in six languages may require special developments or adaptations that should be foreseen by the vendors. Similarly, it remains to be seen whether distributed management of translation memories can be efficiently organised on a large scale, with fifty translators having the right to update the translation memory on a permanent basis in each language pair.

3.3.2 *The perception of the consequences of using the tools*

Fears on the translators' part come from a three-fold realization: (1) managers have to be aware of the difference between potential and actual recycling and the tools should be developed to help them be aware of the differences; (2) translators should continue to be assessed for their linguistic and narrative competence and performance – their computer skills should not be considered as important as the first two skills; (3) CAT tools are generating full-system benefits rather than helping translators translate: CAT tools do enhance documentation quality and help streamline the translation workflow (e.g., by saving typists time, improving terminology consistency, identifying references, helping through alignment to spot errors in already-published texts), but do not change the nature of the work of the translator who has to convey a message from one language to another in an accurate manner, that is without mistranslations, omissions, serious shifts of emphasis (wrong nuance or shade of meaning).

4. Tools for translators

Translators at the United Nations make use of internal glossaries and terminologies developed within the specific institutional constraints.

4.1 In-house glossaries

A dictionary look-up tool commonly used by translators at the United Nations provides a list of equivalents to remind translators of all possible synonyms as is the case for "significant" in English and its possible renderings into French:

"Significant – Accusé, appréciable, assez grave/long, caractéristique, certain, considérable, de conséquence, d'envergure, de grande/quelque envergure, digne d'intérêt, d'importance, de poids, de premier plan, distinctif, efficace, élevé, éloquent, explicatif, expressif, grand, important, indicatif, instructif, intéressant,

large, louable, lourd de sens, manifeste, marquant, marqué, net, non négligeable, notable, palpable, parlant, particulier, pas indifférent, perceptible, plus que symbolique, positif, pour beaucoup, probant, qui compte, qui influe sur, réel, remarquable, représentatif, révélateur, sensible, sérieux, soutenu, significatif, spécial, substantiel, suffisant, symptomatique, tangible, valable, vaste, véritable, vraiment; a significant proportion: une bonne part; in any significant manner: un tant soit peu; not significant: guère; the developments that may be significant for: les événements qui peuvent présenter un intérêt pour; to be significant: ne pas être le fait du hasard."[7] Not only words but phrases are useful in glossaries meant for translators.

Access to validated and standardised terminology is considered more important than access to tools for document reuse other than the basic cut and paste function from documents carefully selected by the translator and not automatically provided by the system. Dictating sentences afresh, once proper terminology has been identified, is considered a less time-consuming process than reading and correcting all or a selection of all possible renderings of a sentence found in previously translated documents by a context-based translation tool. Language resources used by United Nations translators thus are primarily terminology search engines that facilitate the search for adequacy given the specific context in which the document has been drafted, rather than any previous context.

4.2 Web resources

Language resources used by translators also include online dictionaries and government and research institutions' websites that translators have learned to identify and query for information extraction and data mining. Portals have been designed to help translators locate best language and document sources on the Internet. Automatic translation offered on the Internet, such as Google Translate, can in some instances provide help to the translator. Web resources have to be used with caution. United Nations documents found on web pages other than the United Nations official document site should not be used for reference purposes as they may not be the final edited versions.

4.3 Alignment tools

Additional tools are document alignment tools by language pairs. Indexing of large text corpora for retrieval of precedents are felt preferable to tools that provide text segments, be they paragraphs, sentences or sub-units with their respective translations, but without any indication of date, source, context, originator, name of

7. Organisation des Nations Unies. Division de traduction et d'édition (2000).

translator and reviser to assess adequacy and reliability in an environment where many translators are involved. Translators at the United Nations Headquarters in New York and to a limited extent translators at the United Nations Office at Geneva a combination of tools where alignment robots (Logiterm) play an important role: aligned bilingual texts are either indexed by dtSearch for the Web and easily searchable manually by key words, or fed into the TRADOS translation memories which are operated on stand-alones as no network environment is provided.

4.4 Knowledge base

The construction of a knowledge base is envisaged to help translators perform their task in a more efficient manner. Ideally it would capture all knowledge generated by United Nations bodies and organs and various organisations and institutions working in related fields (i.e., any subject from outer space to microbiology tackled by the United Nations), and the knowledge and know-how of an experienced translator well trained in United Nations matters and that of an experienced documentalist knowing which documents are the most referred to. Such knowledge base would, for instance, predict instances where "guidelines" should be translated in French by "directives", as given by most dictionaries, and where 'principes directeurs' would be a more appropriate translation. In statistical documents at the United Nations, one finds "recommendations," a term which is translated by "recommandations" in French and refers to rules to be followed, and "guidelines", translated as "principes directeurs," which are mere indications to be taken into consideration. If the term "directives" would be used in such context, it would convey the meaning of a document of a more prescriptive nature than "recommandations" would, which are actually more binding. Such instances of translation are best captured by a knowledge base that refines contexts and provides best reference material on any topic in the text to be translated. The knowledge base would provide not only adequate referencing and documentation of the original, but also the basic understanding of any subject that arise in a United Nations document. Such knowledge base ideally would reduce the choices offered to the translator rather than list all possibilities. The easier it is for the translator to make the decisions he or she needs the faster he or she delivers.

The knowledge base would offer the translator with past alternatives, too, as in the case of "sexual harassment", translated into French by "harcèlement sexuel". Other French equivalents were tested before this rendering was coined and accepted. They may arise in a French original to be translated into other languages and thus should be retrievable: "assiduités intempestives," "avances (sexuelles) importunes," "privautés malvenues," "tracasseries à connotation sexuelle". The knowledge base would refer, too, to associated terms: "attentat à la pudeur," "outrages."

5. Conclusion

United Nations translators are very cognizant of the limitations of automated tools for translation and are more inclined to rely on easily accessible, structured information concerning the history and main issues in a particular subject matter in order to be completely free to choose the best translation equivalents. First and foremost, tools meant for translators working in environments such as the United Nations should facilitate the systematic provision, in any target language, of equivalents for key terms found in original texts submitted for translation.

References

Hobbs, G. (2008) Bringing new recruits up to speed. Paper presented at the *CIUTI Forum*, United Nations Office at Geneva, 24th–25th January 2008.

Jastrab, M.-J. (1984) Terminology standardisation at the United Nations. In *Report of the Third Annual Conference of the Center for Research and Documentation on World Language Problems*, New York, 14th December 1984.

Organisation des Nations Unies. Division de traduction et d'édition. Service français de traduction. (2000) *Vade-Mecum du traducteur (anglais-français)*, SFTR/15/Rev. 3rd September 2000.

Peral, F. (2002) The Impact of New Technologies on Language Services: Productivity Issues in Translation. Paper presented at the *Joint Inter-Agency Meeting on Computer-assisted Translation and Terminology (JIAMCATT)*, 24th–26th April 2002. World Meteorological Organisation. Geneva.

Steiner, G. (1975) *After Babel. Aspects of language and translation.* (first published in 1975, reedited in 1998 by Oxford University Press).

CHAPTER 8

Global content management

Challenges and opportunities for creating and using digital translation resources

Gerhard Budin
University of Vienna

In this chapter the concepts of content management and cross-cultural communication are combined under the perspective of translation resources. Global content management becomes an integrative paradigm in which specialised translation is taking place. Within a case study framework, we discuss the Global Content Management strategy of the Centre for Translation Studies at the University of Vienna.

1. Convergence of content management and cross-cultural communication

Two different paradigms that have previously developed independently of each other have converged into a complex area of practical activities: cross-cultural communication has become an integral part of technical communication and business communication, and content management has become a process that is complementary to communication by focusing on its semantic level, i.e., its content. Specialised translation as a form of cross-cultural communication is a content-driven process, thus digital translation resources become a crucial element in content management that takes places in a globalised marketplace.

Content management has emerged as a concept that builds upon information management and knowledge management with an additional focus on content products, such as databases, electronic encyclopedias, learning systems, etc. Due to globalised commerce and trade, such products are increasingly offered on multiple markets; therefore, they have to be adapted from a cultural perspective, which also includes the linguistic viewpoint. We will have a closer look at the concept of content, its transcultural dimension, and the role that management of digital translation resources plays in this area.

1.1 Reflections on concurrent trends

Economic globalisation had been a re-current development during several phases in modern history and several industrial revolutions and has been one of the crucial driving forces in the development of modern engineering, in particular computer technology. Together with rapid advances in telecommunications it was the basis for building databases and global information access networks such as the Internet. Visualisation techniques and constantly increasing storage capacities led to multimedia applications.

This increasingly powerful technology base has then been combined with terminology management practices in the form of termbases, with multilingual communication and translation requirements as well as with cultural adaptation strategies in the form of localisation methods. Language engineering applied to translation in the form of computer-assisted translation, translation memory systems, and machine translation, have recently been combined with localisation methods and terminology management for creating integrated workbenches.

On the economic level, international trade and commerce have increasingly required cross-cultural management and international marketing strategies tailored towards cultural conventions in local markets. This trend towards customisation of products has generated personalised products and services that are based on specific user profiles, customer satisfaction and quality management schemes. The emergence of information and knowledge management systems has been another key development in recent years. Computerisation and economic globalisation are the key drivers in a complex context of the information society, leading to interactive processes between linguistic and cultural diversity, professional communication needs in economic and industrial processes and technological developments. As a result, cross-cultural specialised communication and content management have emerged, both complex processes themselves, as dynamic and integrative action spaces in society.

2. What is content?

While terms such as data, information, knowledge, have been defined many times so that we can compare and ideally synthesize these definitions, the term *content* has not been defined so often. But since this term is essential for our discussion here, and since it is used so often in terms such as content management, eContent, content industry, etc., we have to take a closer look at what this term actually means.

In a modest attempt at distinguishing the different conceptual levels, an iterative and recursive value-adding chain emerges:

data + interpretation = information + cognitive appropriation = knowledge + collective representation (in potentially multi-media and multi-modal forms) for specific ways of utilisation = content

Each higher level of complexity integrates diverse elements of the lower level. Usability aspects are most important on the content level. All lower levels remain crucial on the higher levels, e.g., data management is still an important part of content management.

Looking at the generic concept behind the word content, we would say: Content is what is contained in a written document or an electronic medium (or other containers of such types). We would expect that any content has been created by humans with certain intentions, with goals or interests in their minds. So we can confirm that content is usually created for specific purposes (such as information, instruction, education, entertainment, arts, etc.).

Content is often created in specific domains (arts, sciences, business/industry, government, social area, education, etc.). When specific content that was originally created in a science context, for instance, it will have to be adapted and re-organised, in order to be able to re-use this content in other contexts, e.g., in secondary education or in industry.

Discussing the term content, we cannot avoid dealing with related terms such as data, information, and knowledge. As we have seen above, it is essential to have a clear distinction between the meanings of (the concepts behind) these terms. From an economic or business perspective, 'data is a set of particular and objective facts about an event or simply the structured record of a transaction' (Tiwana 2000:59f.). We derive information by condensing (summarising, eliminating noise), calculating (analysing), contextualising (relating data to concrete environments, adding historical contexts), correcting (revision of data collections on the basis of experience) and categorising data (Davenport & Prusak 1998).

Data management has always been a fundamental activity that is as important as ever. Data repositories and data sharing networks are the basic infrastructure above the technical level in order to facilitate any activity on the levels above, i.e., information management and knowledge management. The transition from information to knowledge can also be described from a systems theory point of view: a certain level of activities has to be reached, so that knowledge 'emerges' from information flows. Many knowledge management specialists warn companies not to erroneously equate information flows to knowledge flows.

In order to legitimately talk about knowledge, a number of conditions have to be met:

– Cognitive appropriation: knowledge is always the result of cognitive operations, of thinking processes. Yet knowledge is not limited to the personal, individual, subjective level. When people consciously share knowledge on the

basis of directed communication processes, it is still knowledge, either referred to as collective or shared knowledge, or as interpersonal, inter-subjective, or objective knowledge. In theories of scientific knowledge, the term 'objective knowledge' was mainly explicated by Popper (1972) and is the result of regulated research processes such as hypothesis testing, verification, proof, etc., and that is written down in science communication processes. This is the justification for libraries to talk about their knowledge repositories in the form of books that contain this type of knowledge, i.e., objective knowledge. But as mentioned above, this knowledge is also subjective knowledge in researchers when they created it and when they communicate about it or when they disseminate it to others (e.g., in teaching).

- Complexity: the level of complexity is another factor in the transition from information to knowledge. The same processes as on the previous emergence level, from data to information, are relevant: condensation of information (summarising), analysis and interpretation of information gathered, contextualisation (relating information to concrete problem solving situations, embedding and situating information in historical contexts and drawing conclusions from that, correcting (revision of data collections on the basis of experience) and categorising knowledge accordingly.
- Life span: the validity of knowledge has to be checked all the time. Again we are reminded by Popper that all knowledge is unavoidably hypothetical in nature and that no knowledge is certain for eternity. Therefore we constantly have to redefine the criteria by which we evaluate our current knowledge for its validity. Another metaphor from nuclear physics is used for knowledge, especially in scientometrics: the 'half life' of knowledge is constantly decreasing, due to the increase in knowledge dynamics, not only in science and technology, also in industry, commerce and trade, even in culture, the arts, government and public sectors, the social sector, etc.

In knowledge management, three basic steps in dealing with knowledge are distinguished (Nonaka & Takeuchi 1998; Tiwana 2000: 71ff., etc.):

- Knowledge acquisition: learning is the key for any knowledge management activity
- Knowledge sharing: the collaborative nature of knowledge is the focus
- Knowledge utilisation: knowledge management systems have to allow also informal knowledge to be dealt with, not only formalised knowledge (this is a crucial factor in evaluating knowledge technologies for their suitability in knowledge management environments.

The focus and the real goal of knowledge management are actually directed towards content, i.e., not on the formal aspects of computing, but on what is behind

the strings and codes, i.e., the concepts and the messages. When knowledge is then packaged as a product for a certain audience, presented in certain media presentation forms, then we can speak about content, which also has to be managed in specific repositories and to be processed for publishing purposes, for instance.

As soon as we introduce another dimension, that of culture and cultures, communicating content across cultural boundaries becomes a crucial issue. Since we talk about localisation as the process of culturally adapting any product to a market belonging to another culture than that of the original market of a product, content also needs to be localised when it should be presented to other cultures. Translation, as a part of the complex process of localisation, is one crucial step in this process, but not the only one. Content localisation may very well involve more than translation in the traditional sense, i.e., we might have to re-create part of that content for another culture, or at least change fundamentally the way this content is presented to a certain culture.

Since 'content' is a relational concept, we have to ask ourselves, what contains something, i.e., what is the container, and what is in this container. A book (with its table of contents), for instance, is such a container, or a database with the information entered in the records as the content. A text or a term can also be containers, with the semantics of sentences and the meaning of the term as the content. But this distinction between container and content cannot be made in a very clear-cut way. We are faced with a semiotic dilemma. Form and content always interact. The medium we choose to present certain information will have some impact on this information; the structure of the information will also lead us in the choice of an adequate medium. Usually we cannot completely separate the container from the content, the form from the content, the term from the concept, the semantics from the text, the medium from the message, etc. Despite the heuristic validity and necessity of an analytical separation, we need a synthesis in the sense of a dynamic interaction, an interactive complementarity. At the same time we also might want to transform one form of knowledge representation into another one, for certain purposes and tasks, and then have to be sure that the content of each knowledge representation does not change – a difficult task.

Similar to typologies of data, information, and knowledge, we also need a content typology.

There are different criteria for distinguishing types of content:

– the domain where specific content is created in: any field of scientific knowledge, a business branch, a profession, a form of art, a type of social activity, etc. For this type of distinction, we may also differentiate different degrees of specialisation (highly technical and scientific, mono-disciplinary or multi-disciplinary, popularised, etc., depending on the audience targeted);

– the form of representation: text, picture, personal action, etc. or the medial manifestation: website content, the 'story' of a film, of a video, a piece of music recorded, a digitised scroll, etc.

Here we see again that the form of representing content and the medium chosen to do this is constitutive for distinguishing types of content.

First of all, the purpose of the content: instruction, education, research, aesthetic and artistic purposes, etc. Secondly, the kind of content product that is designed for a particular target audience (e.g., a multimedia CD-ROM for 6-year old children to learn a foreign language, e.g., English). In addition to a content typology, we also have to look at the structures of content. In this respect, and regardless of the content type, we can make use of terminology engineering and ontology engineering. Terminologies and ontologies are the intellectual (conceptual) infrastructures of content, both implicitly (in the form of personal or subjective knowledge of the content generator), or explicitly (as objective knowledge laid down in a specific presentation form).

So we can conclude that concepts are content units (conceptual chunks) and that conceptual structures (the links among concepts) are the structures of content. Again we have to remember that the multi-dimensional content typology will determine the concrete structures of content that users will encounter in specific products.

3. Global content management

After having investigated a little bit into the concept of content, we can now look at content management and how cultural diversity determines this practice. Since the target audience of any content product is always culture-bound, i.e., belonging to one or more cultures, we can simply state that content management always has to take into account cultural factors in content design and all other processes and tasks of content management. The language(s) spoken by the target audience, social and historical factors, among many others, are examples of criteria for concrete manifestations of content management. Also at the meta-level of content management, those who are content managers are also culture-bound. Those who have designed and created content products, such as multimedia encyclopedias on CD-ROM, have to be aware that they themselves are belonging to at least one culture (in most cases, there will be one pre-dominant culture in such content management teams), and that this very fact will unavoidably determine the way the content of the product is designed.

In addition to the phases of creation and design of content, there are other key processes of content management at the processing stage:

- Analysis of existing content structures, segmentation of content into units, aggregation of content units into structures, condensation of content (summarization, abstracting, etc.), expansion of content into more detailed forms, transformation of content, etc.
- Presentation of content in different media and knowledge representation forms (see above)
- Dissemination of content on intranets or other web structures, on CD-ROMs, but also more traditionally in the form of books, etc.
- Sharing content in collaborative workspaces
- Using content for various purposes

Taking into consideration the differentiation between data, information, knowledge, and content (see above), we can make a parallel distinction between data management, information management, knowledge management, and content management. It is important to note that each management level is based on the one underneath, i.e., information management is impossible without data management, knowledge management needs both, data management and information management, and content management relies on all three levels below.

Now we should return to the aspect of cultural diversity and the way it determines content management. Global content design, accordingly, is an activity of designing content for different cultures as target groups and is cognizant of the fact that content design itself is a culture-bound process, as shown above.

From the field of cultural studies we can benefit when looking at definitions of what culture is: a specific mind set, collective thinking and discourse patterns, assumptions, world models, etc. Examples for types of culture are corporate cultures, professional, scientific cultures, notably going well beyond the national level of distinguishing cultures. Cultural diversity is both a barrier and at the same time an asset and certainly the raison d'être for translation, localisation, etc.

Global Content Management is a complex concept with a specific structure: The term element 'global' stands for all the cross-cultural activities such as translation, localisation, but also customisation, etc. 'Content' includes language resources, such as terminologies and ontologies as its infrastructures, products and their design, user documentation, but also pieces of art, etc. And the management component includes all the processes such as markup and modelling, processing, but also quality management, communication at the meta-level, etc. Usability engineering is crucial for all these components.

Content management processes cannot do without appropriate knowledge organisation and content organisation. Terminological concept systems are organised into Knowledge Organisation Systems (KOS) that can be used for this purpose of content organisation: Thesauri, Classification Systems, and other KOSs, also conceptualised and formalised as ontologies. Such ontologies may

be language-related (e.g., WordNet), domain-specific (medicine, etc.), or task-oriented (operating workflows such as in robotics). In order to establish and maintain the interoperability among heterogeneous content management systems, federation and networking of different content organisation systems are necessary in order to facilitate topic-based content retrieval and exchange of content in B2B interactions.

Global Content Management may have very different manifestations. In the area of Cultural Content Management, for instance, cultural heritage technologies have developed in order to build up digital libraries, digital archives and digital museums.

Other applications of Global Content Management systems are:

- ePublishing (single source methodologies)
- E-Learning (managing teaching content)
- Cyber Science (Collaborative Content Creation)
- Digital Cities and other Virtual Communities projects.

On the pragmatic level of maintaining content management systems we observe similar problems as on the level of knowledge management, that a corporate culture of knowledge sharing has to be developed and nurtured, that special communicative and informational skills are needed to share knowledge across cultures and that the dynamic changes in content require a management philosophy that is fully cognizant of the daily implications of these constant changes.

Translation resources such as translation memories and other aligned corpora, multilingual terminological resources, reference resources, etc. are typical examples of content that needs to be managed in such global action spaces.

4. Pragmatic issues in global content management: A case study on the Centre for Translation Studies at the University of Vienna

Study programs offered at the Centre for Translation Studies include a Bachelor in Transcultural Communication and two Master programmes in Translation (specialising into technical (or LSP) translation and literary translation) and Interpreting (specialising into conference interpreting and dialogue interpreting (incl. community interpreting) and cover 14 languages.

50% of all courses held are implementing a blended learning strategy that we have developed and that covers all phases of content creation and content utilisation. This strategy explicitly addresses the global aspect of learning content (cross-cultural learning), the technological dimension of using language technologies and knowledge engineering methods for creating and using digital learning objects, and the social dimension of the interaction between teachers and learners

with the additional role of tutors who are trained to offer E-Learning support to both, teachers and learners.

The Media Lab of the Centre operates several servers with content repositories including audio content (recorded speeches that are used for interpreting classes and radio broadcasts in different languages, TV broadcasts and video recordings, web content, full text corpora (written texts), lexical resources (glossaries, terminology databases), as well as digital learning objects (content units specifically designed for use in an E-Learning environment). Using a single sourcing approach, these central repositories are used by many different people for different purposes (research, publishing, and teaching) and in different learning contexts (such as courses).

In the context of a university-wide project called PHAIDRA (Permanent Hosting, Archiving anf Indexing of Digital Resources and Assets), all content objects (that are indeed digital assets) are systematised in a taxonomy of content object types and annotated with metadata that are essential for enhanced search and retrieval functionalities. The collaborative aspect of content management is of particular importance in this context: starting at the personal, individual level, all researchers, teachers and students are able to manage their personal digital objects (their research papers, their annotations on learning objects and on published content, etc.). In the PHAIDRA context any kind of group can be established when their members decide to work collaboratively on a certain content object, such as a joint research paper or a learning object that is collaboratively created and updated. The next level of collaborative work is a stable organisational (sub-)unit or a specific strand of work, such as all teachers in the Spanish program, all teachers offering interpreting courses, the whole faculty staff, etc. Beyond the faculty level the PHAIDRA approach also allows to provide access to local content to colleagues at other faculties as well as to other research groups in other universities.

Our approach to Global Content Management geared towards the many different use contexts in translation studies is based on a highly granular, i.e., conceptual approach. Terminological entries are the most granular units. They are organised in a terminology database that can be hyper-linked to all text resources that contain the terms documented in the termbase.

Fig. 1 shows an entry of our terminology database on risk management. In English, German and French basic terms of risk management are documented with definitions, referenced to multiple sources and indexed according to a subject-specific classification system. The multiple purposes of this database are:

– project communication: the database is a deliverable in a European project on risk management (WIN – Wide Area Information Network on risk management in Europe, coordinated by Alcatel, with a work package "human language interoperability" coordinated by the University of Strasbourg (Gertrud

Greciano). It is being extended to more languages and the underlying data model is also being enhanced. The database will further be used as terminological input into a project-wide ontology for risk management
- web-based access to multilingual risk terminology for various target groups (such as risk experts in different domains such as geology, environmental protection, biology, civil engineering, remote sensing, etc., translators and other language professionals, information managers, students, etc.) – there is a web interface for this database
- learners' resource: the database serves as a learners' resource in a number of courses (e.g., on terminology management in order to study the aspects of terminological data modelling, or for translation courses where the database serves as a look-up resource for translation work in different language combinations and translation directions.
- referencing framework: a large multilingual text repository on the topic of risk management and natural hazards has been created, with these texts being analysed with a term extraction tool in order to collect data on the use of these terms in real-life contexts

The database is currently being integrated into the E-Learning environment at the Centre for Translation Studies so that it can be used in the specific learning contexts in many different courses offered in the study programmes.

Another example is given for complex learning objects that have been designed and created for use in various translation studies courses. Within the Mellange project (Multilingual E-Learning for Language Engineering – a Leonardo da Vinci II project coordinated by University of Paris Denis Diderot) a whole course program for a European Master on Translation Technologies with the relevant learning content organised in learning objects have been created. Figure 2 is an example for one of these learning objects that are being created for E=Learning in the field of translation technologies. The example shows a course unit on information management for translators in the form of a structured hypertext-based sequenced learning unit. All these SCORM objects have been created in a collaborative way by expert teachers from the members of the Mellange consortium. The E=Learning environment used for the project is Moodle. Figure 2 shows the use of this learning object on Moodle in the Mellange project.

The integration of E-Learning, knowledge management, content management, and communication management is a crucial aspect of our blended learning strategy. For us, E-Learning is a multilingual, cross-cultural process. Members of learning communities, teachers, tutors, etc. increasingly have different cultural and linguistic backgrounds. This aggravates communication problems caused by the specific teacher-learner situation by adding another dimension of cross-cultural communication and its countless pitfalls that most communica-

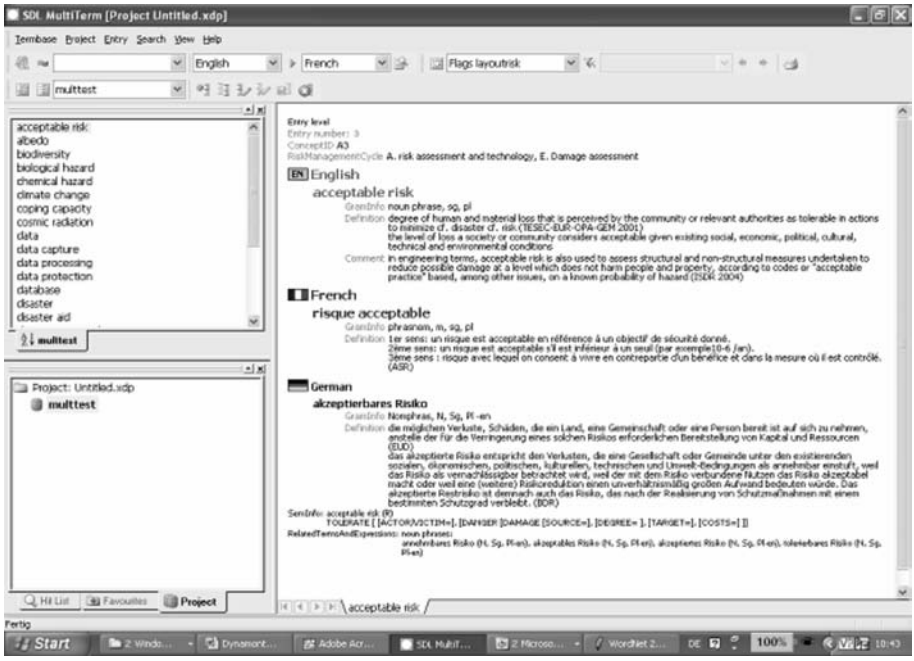

Figure 1. An entry in the terminology database on risk management

tion partners are not even aware of. Analysing previous and ongoing E-Learning projects (e.g., Budin 2006) as well as by looking at numerous other concrete projects it seems necessary to generalise from individual activities for formulating E-Learning strategies for content development.

In the process of further expanding and developing the E-Learning environment at the Centre for Translation Studies, the following processes are identified requiring specific support from the PHAIDRA initiative:

– Multilingual co-operative work for collaborative student work (including collaborative annotation of shared content such as learning objects, collaborative writing translations as well as research papers, collaborative glossary preparation, etc.)
– Cross-cultural and cross-disciplinary collaborative work (group work with students from different countries) with specific support requirements such as meta-communicative mediation and annotation functions to mediate between diverse cultural groups, to explain in the required degree of explicitness specialised content to members of other disciplines or professions
– The multiple (re-)use of multilingual language resource corpora: the re-use of corpora is an essential element of the content strategy. At the same time there is also a dilemma that requires specific solutions: on the one hand language

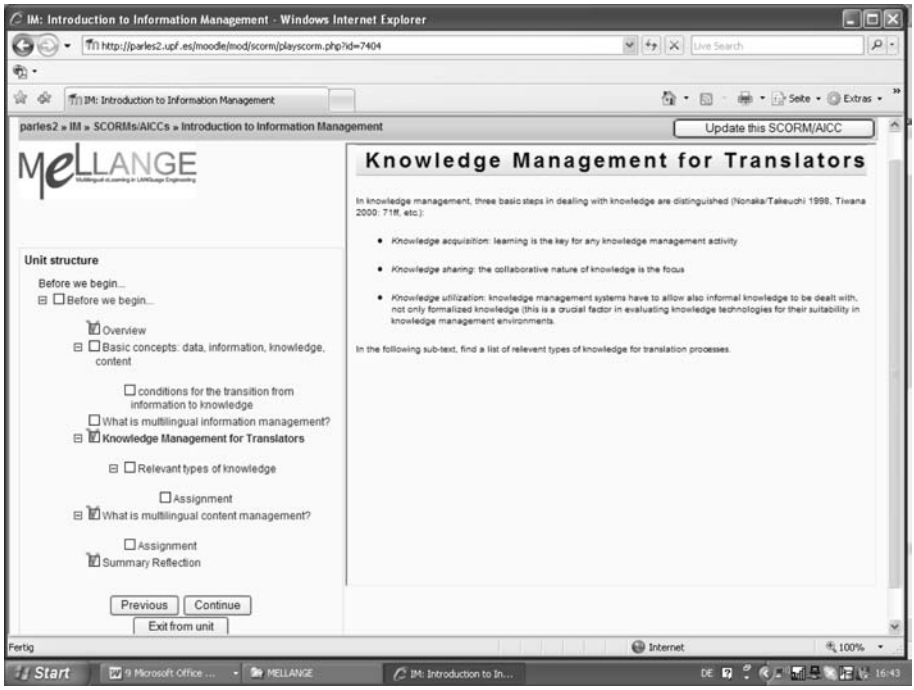

Figure 2. A SCORM learning object on Moodle on the topic of Information Management

resources should be highly re-usable, which essentially means that they should be as "neutral" as possible in relation to any specific learning context or learning goal, on the other hand learning objects should be customised and adapted as much as possible to clearly defined learning contexts and learning goals. The role of a separate annotation layer as well as a communication layer becomes obvious – the learning object stays unchanged, but the annotation layer pre-customises it to a specific E-2Learning context and a specific learning goal and the communication layer enables the users (teachers, students, tutors) to post-customise it to their learning activities.

– Modelling different competence levels for curriculum design for Bologna-type study programmes such as bachelor and master degrees: competence modelling is an important perspective for curriculum design as well as for the design of learning objects governed by specified learning goals. For the translation professions and other language professions this has become an important research area
– Search and navigation support across various content repositories
– Resource assessment for quality assurance
– Enhancing terminological coherence and consistency across all content

– Text mining with ontology systems, terminology extraction from language resource corpora, metadata harvesting from learning content, multimedia content management and content repository management.

The following process model has been proposed (Budin 2005) as the starting point for strategy development for E-Learning environments. The aspect of interactivity is seen as crucial for E-Learning in the future. Interaction design has become an important principle for learning design (interactive learning) and in fact for all modules of an E-Learning environment. Interactivity is also crucial for linking the four different dimensions of the model to each other. All four dimensions only make sense in an interactive model as part of the whole. Therefore it is mandatory that E-Learning environments show all these dimensions, none of them can be eliminated or simply "forgotten" as it frequently happens these days. The steering dimension is the left one, i.e., didactic design and learning management, this includes workflows of learning processes that are monitored, managed, and supported by teachers, also in exploratory autonomous learning situations. Knowledge management is an important aspect for E-Learning environments, but at the moment only few academic organisations have explicit knowledge management strategies. In that respect corporate E-Learning traditions are far more advanced by integrating knowledge management and E-Learning processes. Designing and using tools for hypermedia communication and for collaborative learning have become very important processes for supporting the social dimension of learning. The fourth dimension is obviously another crucial one, i.e., multilingual content development and content repository management. All four dimensions are linked to each other in dynamic ways, as Fig. 3 shows:

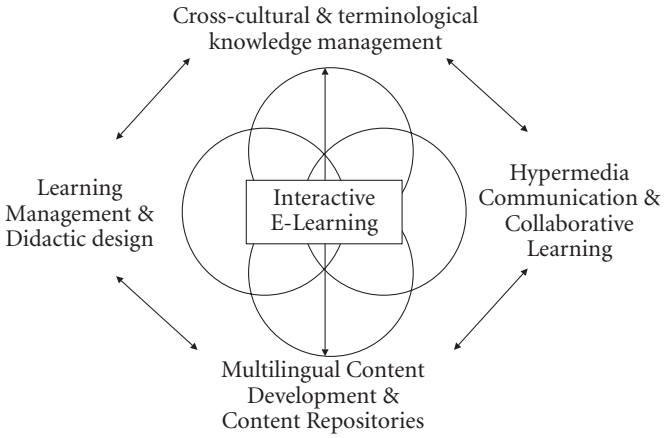

Figure 3. An Integrated Process Model of Interactive E-Learning (Budin 2005)

5. Outlook

On the technological level a number of enabling technologies for global content management have emerged that are converging into Semantic Web technologies. Intelligent information agents are integrated into such systems. They are combined with knowledge organisation systems (in particular multilingual ontologies). Semantic interoperability has also become a major field of research and development in this respect.

In the field of the so-called content industry different business models have developed that could not be more diverse: on the one hand open source and open content approaches are rapidly gaining momentum, also facilitated by maturing Linux-based applications. On the other hand national, regional and international legislation concerning intellectual property rights is becoming more and more strict, and global players are buying substantial portions of cultural heritage for digitisation and commercial exploitation that might eventually endanger the public nature of cultural heritage.

Epistemological issues of global content management will have to be addressed, as well as best practices to be studied in detail in order to develop advanced methods for these complex management tasks. Managing cultural diversity in a dynamic market with rapidly changing consumer interests and preferences, with new technologies to be integrated, also requires a strategy for sustainable teaching and training initiatives (based on knowledge management teaching and training initiatives) in this fascinating field.

In conclusion, it seems that strategies for multilingual learning content development for E-Learning environments require a complex approach to modelling learning processes, didactic knowledge organisation, ontology creation and multilingual resource support.

References

Budin, G. (2005) Strategies for Integrated Multilingual Content Development and Terminological Knowledge Management in Academic E-Learning Environments. In Nistrup-Madsen Bodil (ed.) *Proceedings of the Congress on Terminology and Knowledge Engineering 2005*, Copenhagen, 91–100.

Budin, G. (2006) Theoretische und methodische Grundlagen integrierter Wissens- und Lerntechnologien. In Mettinger, A., C. Zwiauer & P. Oberhuemer (eds.): *E-Learning an der Universität Wien*. Forschung – Entwicklung – Einführung, Waxmann: Münster u.a., 43–56.

Davenport, T. H. & L. Prusak (1998) *Working Knowledge. How Organisations Manage What They Know*. Boston: Harvard Business School Press.

Nonaka, I. and H. Takeuchi (1995) *The Knowledge-Creating Company*. Oxford: Oxford University Press.

Popper, K. (1972) *Objective Knowledge. An Evolutionary Approach*. London: Routledge.

Tiwana, A. (2000) *The Knowledge Management Toolkit. Practical Techniques for Building a Knowledge Management System*. Upper Saddle River: Prentice Hall.

CHAPTER 9

BEYTrans

A Wiki-based environment for helping online volunteer translators

Youcef Bey[1,2], Christian Boitet[2] and Kyo Kageura[1]
[1]Graduate School of Education, The University of Tokyo /
[2]GETALP, LIG laboratory, Université Joseph Fourier

The aim of our research is to design and develop a new online collaborative translation environment suitable for the way online volunteer translators work. In this chapter, we discuss how to exploit collaborative Wiki-based technology for the design of the online collaborative computer-aided translation (CAT) environment BEYTrans, which is currently under development. The system facilitates the management and use of existing language resources and fills the gap between the needs of online volunteer translator communities and existing CAT systems/tools.

1. Introduction

In accordance with the current global exchange of information in various languages, we are witnessing a rapid growth in the activities of online volunteer translators, who individually or collectively make important documents available online in various languages. For instance, the W3C consortium is made up of over 300 volunteer translators who have translated thousands of documents covering more than 40 languages (W3C 2007). The volunteer translators working on the Mozilla project produce documentation in 70 languages (Mozilla 2007). The Global Voices project, which aims at providing translations of blog news and articles, now covers scores of languages, thanks to volunteer translators (Global Voices 2007). In addition to these project-oriented translations, many volunteer translators are individually translating a variety of online documents into a variety of languages (Pax Humana 2007; Tea Not War 2008).

Turning our eyes to the sphere of language resources potentially useful for these translators, the volunteer-based construction of multilingual dictionaries and lexical databases is also growing (Breen 2001; Mangeot 2002; Wiktionary

2007). This extends to less well-represented languages as well (Streiter et al. 2005a, 2005b, 2006). As Streiter states (2006), "seemingly unstructured bottom-up cooperation between knowledgeable volunteers" has provided "new ideas on how scientific cooperation might be organised," which will "influence the field of NLP in general."

Both the translation activities and the NLP-related activities including language resource construction described above are made possible by the Internet, which allows people to meet, communicate and share a variety of multilingual information. It is therefore natural to expect the former activity to be made easier by the latter. For now, however, there is a dearth of online translation-aid tools for volunteer translators which would make use of a wide variety of online language resources (Bey et al. 2006b). It is within this context that we have developed the translation-aid system BEYTrans (Better Environment for Your Translation) to specifically meet the needs of volunteer translation communities. The first version of the system is currently available and used on an experimental basis.

This chapter outlines the design of the BEYTrans system. In Section 2, we briefly introduce the different types of online volunteer translators and their needs. Section 3 postulates the basic desiderata for computer-aided translation (CAT) systems specifically for volunteer translator communities. Sections 4, 5 and 6 explain the collaborative functions in BEYTrans for supporting community-based translations, language resource management, user-oriented functions, and the system interface.

2. Types and needs of online volunteer translators

Online volunteer translators can be divided into two basic types (Bey et al. 2006a, 2006b):

1. Translators involved in closely linked, mission-oriented translation communities: they work in strongly coordinated groups translating clearly defined sets of documents, most typically what can loosely be called technical documentation, such as Linux documentation (Traduc 2007), W3C specifications, and both documentation and software (interface, messages, online help) of open source products. Localisation of W3C consortium documents (W3C 2007) and Mozilla (Arabic Mozilla 2007; French Mozilla 2007) is supported by this kind of volunteer translation community. In many cases, different translators translate different parts of the same document, hence mutual coordination and version control is required for this type of translation. We will call this type of voluntary translation activity "community-based translation." The BEY-

Trans system focuses on the needs of volunteer translators engaged in this type of translation.

2. Subject-oriented translators involved in network communities: they translate online documents such as news, analysis, and reports, and make the translations available on personal or group web pages (Tea Not War 2008; Pax Humana 2007). They tend to work individually, although they often form loose networks with other translators sharing the same interest. In some cases, such as Global Voices (Global Voices 2007), an administrative body takes charge of collecting and organizing the translated texts. In most cases, each translator translates individual documents in their entirety. Unlike community-based translation, therefore, there is little necessity for individual translators to explicitly coordinate with other translators. We will call this type of voluntary translation activity "networked individual translation." A system specifically designed to assist translators doing this type of translation has also recently been developed (Abekawa & Kageura 2007).

While several high-quality commercial CAT systems (DéjaVu 2007; Trados 2007) and some free CAT systems (Omega-T 2007) are available, a survey we carried out showed that volunteer translators tend not to use these tools.[1] This result is in accordance with the results of surveys on British freelance translators' attitudes towards CAT and related tools (Fulford 2001; Fulford & Zafra 2004).

Our survey showed that volunteer translators basically work on their own, using their own paper or electronic dictionaries. Though some of them use basic online dictionaries, they do not make full use of language tools on the server or the wide variety of language resources available online (Bey et al. 2006). In the event that mutual coordination or adjustment is necessary, as often happens in community-based translations, they communicate with each other via mailing lists, forums, Wikis or blogs.

As the available CAT tools are not optimised for use by community-based online volunteer translators, they suffer from several deficiencies from the point of view of these translators.

– The functions for sharing translation and language resources are in most cases not sufficient, a problem that is aggravated by the fact that the tools also lack efficient communication mechanisms through which translators could

1. We interviewed six online volunteer translators (three of them were professional translators who also do online volunteer translation; the other three were non-professionals), and talked with the central figure in the French Linux localisation project. In addition, the third author is a semi-professional translator, who has so far published nine translated books and various articles, and has been involved in volunteer translation work for nearly 10 years in the field of human rights, with special reference to East Timor.

coordinate with other translators, share skills, and solve translation problems (Aguirre et al. 2000);
– Translators cannot efficiently and systematically look up existing translated document pairs within the overall community environment when translating new related documents. In other words, there is no well-developed function for facilitating immediate and well-managed sharing of translation memories (TMs) by communities;
– An online editor that could be used collaboratively by multiple translators in same community is not provided;
– Sometimes the range of file formats that the system can deal with is too limited.

On the other hand, a community-oriented working environment such as Wiki, which some translator communities are using for mutual communication and coordination, lacks functions necessary for translation activities. Among the most important are:

– the lack of a unified management of documents and language resources in the same environment (dictionaries, glossaries, TMs, etc.);
– the lack of language resources functions, which are not provided, whether during translation or for preparation of the translation work, and later during post-edition.

Delving deeper into individual translators' or translation communities' behavior would show that there are many other needs that are not sufficiently met at present, but, for the moment, we limit our work to solving the major problems mentioned above, which are common to many translators and translation communities.

Volunteer translators want solutions to these deficiencies to be provided in a unified way. Furthermore, as they are typically not computer experts and do not have the ability or time to do system-related tasks, it is important to ensure that they do not need to perform technical tasks such as editing HTML and XML tags or modifying script programs. Let us now outline proposals for tackling the major problems described above.

3. Proposals for solving existing deficiencies

Against this backdrop, we have postulated three major points to be pursued in developing BEYTrans as a system aimed at online volunteer translators engaged in community-based translation.

1. *Facilitating Collaborative Translation.* The current movement away from stand-alone systems and toward online environments which facilitate networking is likely to continue (Bowker 2002), thus making it possible for multiple users to share the same TMs, machine translation (MT) systems, and various language resources. Our environment has to consistently unify the collaborative management of documents in various multilingual formats. A common awareness among participating translators of the current state of the translation is important. Our environment is hence designed to allow different translation versions to be retained and controlled as the translation progress. The same requirements are applied to language resources. As a matter of fact, dictionary entries can also be modified by volunteers and become immediately available on the Web. It is also necessary to integrate the functions that facilitate communication among translators, so they can be informed of any modification of the content and can discuss problematic issues. The Wiki architecture supports the basic functions necessary for online collaborative working environments, thus we use Wiki as the basis of BEYTrans. We will elaborate on this in Section 4 below.
2. *Managing Language Resources.* We categorise resources into two separate groups. The first category is language resources, which include dictionaries, glossaries and terminology banks or databases, etc. In order to provide a variety of language resources to meet the specific needs of volunteer translators, existing dictionaries and glossaries are collected and pre-processed for importation into our environment. The second category is TMs, which aim to manage translation units (TUs). TMs can be enhanced by adding new pairs of units during translation or by automatically recycling matching pairs of units from previous translations in the repository of the translation community or from the Web. A system that specifically aims at recycling existing translation document pairs, not "parallel" texts, is currently being developed (Kageura et al. 2007). The technical aspects of language resource management will be discussed in Section 5.
3. *Providing an Integrated Environment and Multilingual Web Editor.* Translators require a uniform environment within which they can access necessary functions easily. The system, therefore, should provide an integrated environment to translators, including a collaborative working environment, language resource management, and an effective interface for making translations, that reflects the actual workflow of translators. It should also be kept in mind that many of the community-based translation projects are concerned with more than two language pairs. For translators to work in a truly collaborative manner within such projects, therefore, it is necessary to provide a translation editor that allows multiple language translators to share TMs and documents for translation. A rich online editor for direct and collaborative translation

should thus be an inherent part of the system. We will elaborate on these aspects of the system in Section 6.

4. Wiki-based functions for an online CAT environment

The technologies that allow collaboration and sharing of knowledge on the Web in general have significantly progressed. Technically speaking, the collaborative functions for translation data management can easily be realised using an existing free collaborative environment such as XWiki, based on the Wiki architecture, which we adopted as the basis of our system. Existing Wiki functions can then easily be extended by exploiting XML and Ajax.

A Wiki is an online collaborative environment. It allows users to freely create and edit web pages which become immediately active on the Web. The pages can be created either in HTML or using simplified tags, called Wiki syntax. This allows the organisation of contributions to be edited in addition to the content itself. As Augar stated (Augar et al. 2004):

> A Wiki is a freely expandable collection of interlinked web pages, a hypertext system for storing and modifying information – a database, where each page is easily edited by any user with a forms-capable web browser client.

Though there are scores of different Wiki implementations, most of them share common basic features. The content of a web document is posted immediately, eliminating the need for distribution. Participants can be notified about new content automatically by e-mail or news list, allowing them to take a position on new modifications. After any modification, the Wiki environment creates a new document and saves a version of the old content. This allows control over the progress of the translation. To control different communities or groups, the space aspect provided in Wikis allows documents to be grouped in specific community projects. A space consists of a set of documents, a group of users, and access rights. For each modification, metadata are attached, so that any user can trace the history of content modifications, including the date of modification and the profile of the users who made the modifications. Access to Wiki is simple and easy; all that is needed by users is a computer with a browser and an Internet connection (Schwartz 2004).

There is an extended Wiki implementation dedicated to online volunteer translators (Translationwiki 2007). Translationwiki makes good use of basic Wiki functions. For instance, in Translationwiki, each translation TU (translation unit) is handled as an independent document. Translators can upload documents in Arabic, Chinese, English and French. After automatic segmentation, the TUs are put up for translation, and are translated separately. The system also contains a versioning module that retains the history of the modifications and allows trans-

lators to check the evolution of a translation and avoid losing content. Translators can easily restore old translations deleted erroneously or by vandals.

Within this context, it was a natural decision for us to use Wiki as the basis of BEYTrans. After analysing several Wiki implementations and Translationwiki, and experimenting with some of the available Wikis, we chose XWiki for developing our system (XWiki 2007). It is a Java-based environment which allows easy integration of existing Java tools for NLP processing.

5. Language resources management

In our environment, the language resources and TMs are pre-processed and managed differently. What we understand under language resources (dictionaries, glossaries, technical terminologies, etc.) are imported as raw textual data and transformed into a structured format. In our work to date, we have imported more than 1.7 million entries in Arabic, English, French, and Japanese in XML format. To be useful, a TM (a set of textual translation units extracted from existing translated documents or during translation) has to match the typology and domain of the documents to be translated. Hence, each translation community has to use its own TM. As we could not find any ready-to-use TMs in such communities, we created a small TM from existing documents translated by the W3C community.

In the following sub-sections, the language resources will be described and the XLD (XML Language Data) format will be introduced. Our management of TM in BEYTrans will also be clearly explained with reference to TMX-C, which has been adapted from the TMX standard (LISA 2007). The description in this section is slightly technical, because at this level the actual treatment of resources and the technical aspects that support them are inseparable.

5.1 Language resources

Existing language resources that we have imported to the structure and that are in actual use include "*Eijiro*" and "*Grand Concise*", two high-quality English-Japanese unidirectional dictionaries widely used by many translators, "*Nichigai*", which covers proper names, "*Medical Scientific Terms*", a medical dictionary included to allow us to check the structure of terminological dictionaries (Bey et al. 2006a), and "Edict", a free Japanese-English dictionary (Table 1).

As our environment is open to all languages, it allows for the importation of other dictionaries, subject to two restrictions: (i) the dictionary must be structured in XLD; (ii) its data must be encoded in UTF-8. For example, the *Arabic Mozilla* (Arabic Mozilla 2007) translation community has a free dictionary created

by volunteers containing around 18,000 entries that we also have imported into our environment.

Table 1. Language resources for the Tea Not War and Arabic Mozilla communities

Reference data	Description	Direction	Entries	Format
Eijiro 86	General English-Japanese dictionary (EDP 2005)	EN→JP	1,576,138	Textual
Edict	Free Japanese-English dictionary	JP→EN	112,898	Textual
Nichigai	Guide for spelling foreign proper names in katakana	JP→EN	112,679	Textual
Medical Scientific Terms	Medical terms	EN→JP	211,165	Textual
Technical terms	English-Arabic technical terms dictionary	EN→AR	18,000	Textual

The language direction of the above resources was not unified and the format of entries was different. To unify the directionality, we transformed them into a unified format. However, some problems like duplication and sorting of entries appeared. We eliminated duplicates by merging and re-sorting entries. As for content, some entries consisted of a set of bi-segments which couldn't be managed as usual dictionary entries (Table 2); that was one reason for using the XLD format.

Table 2. Example of Japanese-English language resources

Reference data	Entry format
Edict	仝 [どうじょう] /(n) "as above" mark/ 々 [くりかえし] /(n) repetition of kanji (sometimes voiced)/
Eijiro	$__ annual membership : __ドルの年会費{ねんかいひ} $__ deposit required for making a bid on: ～に入札{にゅうさつ}するため
Medical Scientific Terms	(id=AGR920000010a)　A horizon A 層

The language resources are specific to each translator community. For example, the *Arabic Mozilla*, *Pax Humana* and *Tea Not War* communities use different language resources. These language resources are often structured in different formats. To deal with this diversity, we developed the XLD XML format, which allows them to be easily managed and imported (Fig. 1).

The XLD format consists of three main parts: (i) description element and attributes for the original linguistic resources and content, (ii) source entry elements, and (iii) target elements that contain expressions for explanation of source entries (Bey et al. 2007).

5.2 Translation memory management

To integrate document management with TM, we need to define efficient data structures that satisfy two requirements: (i) maximum provision of recyclable units and (ii) unified management of translated documents. The first requirement derives from the needs of individual translators, who look for relevant linguistic units in existing translations. The second requirement derives from the needs of the manager of the community in which translators participate or from the community itself. To fulfill it, we found TMX (the LISA translation memory exchange standard) to be suitable (LISA 2007). TMX is an XML format which was developed to simplify the storage/exchange of multilingual TMs (Bey et al. 2006b; Boitet et al. 2005). Annotation is done in our environment in accordance with the TMX standard, and the 3-tier model (Saha 2005).

Our 3-tier model and its various information levels (metadata annotation, sentence and language unit annotations) are illustrated in Fig. 2. The result of segmentation is an annotated document, which is used to automatically consult the TM and the language resources in the Wiki store.

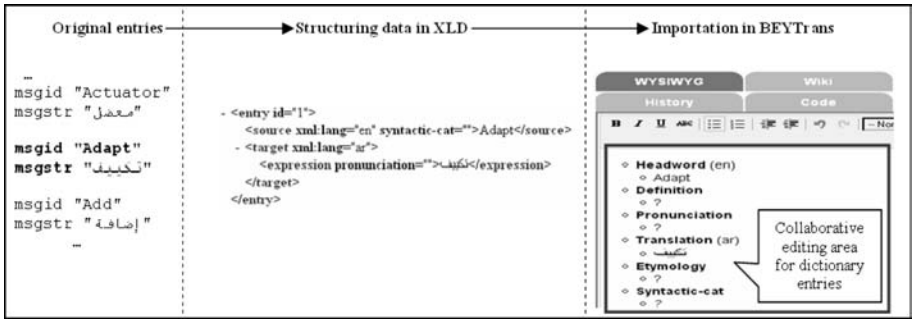

Figure 1. Method for importing dictionaries into BEYTrans

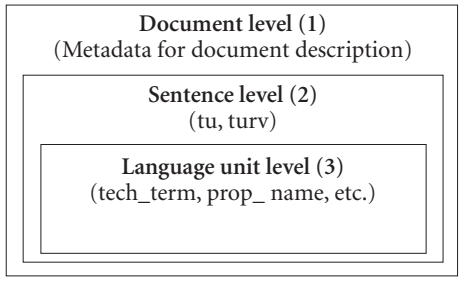

Figure 2. 3-tier model for document segmentation

Taking into consideration the advantages of the 3-tier model and the TMX standard capabilities, we decided to use the TMX standard and extended it for handling collaborative translation.

Our proposed TMX-C format is an adaptation of TMX that deals with three levels during segmentation for constructing the TM format and for supporting collaborative Wiki information.

The segments and XML metadata tags of TMX-C are defined as follows:

Table 3. Metadata annotation tags

Metadata	Description
Domain	Domain of document: technical information, medical, personal, sports, humanitarian, etc.
Original_Community	Original community name
Space	Community space name in the XWiki store
S/T_XWiki_DocName	Document name in XWiki
S/T_XWiki_DocSpace	Space containing the document in XWiki
S/T_XWiki_version	The version generated by XWiki
S/T_XWiki_TU_Order	Order of "TUV" in the document XWiki

Table 4. Translatable and linguistic unit (TU/LU) annotation tags

TU/LU	Description	Imported from
tech_term	Technical term	XLD
Prop_name	Proper name	XLD
Ord_word	Ordinary word	XLD
Quot	Quotation	XLD
Colloc	Collocation	XLD
TU	Translatable unit	TMX
TUV	Translatable unit version	TMX

At the top level, document information is provided, which is not only essential for document management, but also useful for translators to check the context and/or domain of a document. The second and third levels are relative to language units, i.e., sentences at the second level and various language units (quotations, collocations, technical terms, proper names, idioms, etc.) at the third level. These units can be automatically detected using sentence-boundary tools (LingPipe 2007), but translators can manually control these units during translation and editing.

6. BEYTrans: A collaborative environment for helping volunteer translators

6.1 Translator scenario

The environment is open to all volunteers on the Web. Any user/translator can upload web documents and then share the translation work with others by exploiting language resources. Documents to be translated first undergo a process of text extraction that consists of detecting sentence boundaries and identifying TUs (Walker 2001). The language of the source document has to be specified (in the future, we might use a language detector). For each uploaded document, BEYTrans creates a new document, its translation companion (TC), in the TMX-C XML format (TMX for Collaboration), which contains all TUs of its corresponding source document (Fig. 3). The TC is used for collaborative translation: source and target translation segments are stored in it. Before starting translation, the companion document is pre-processed using the local translation memory and available MT systems to get one or more pre-translations or suggestions for the TUs (in the selected target language).

Translators read the result in the target language, possibly synchronised with the source (the original TUs are replaced by the "best" pre-translations in the original document), and they switch from reading to editing mode to perform translation. The online translation editor has an Excel-like interface where all source TUs are displayed in parallel. The editor allows users to exploit dictionaries

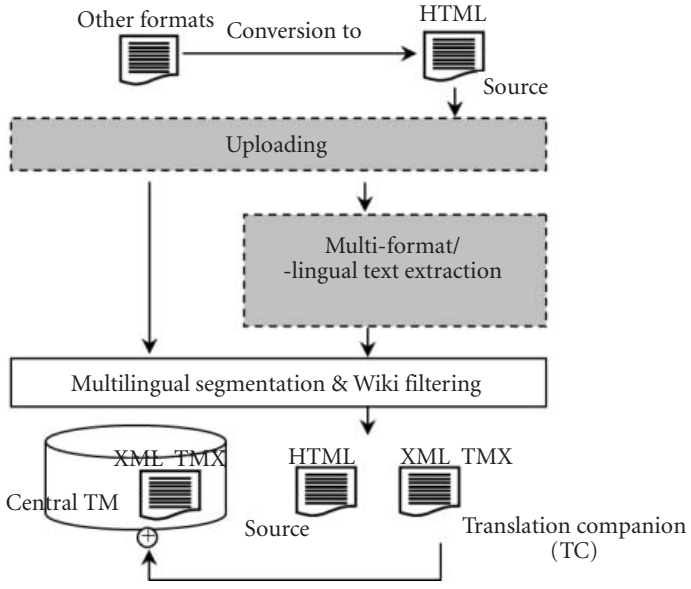

Figure 3. Pre-processing and data importation method

and the TM and is quite flexible. For example, translators can add and remove columns and rows during translation.

During the translation, the editor makes suggestions by computing similarity scores between the currently selected TU and the TUs stored in the Central Translation Memory (CTM), which contains all the previous translations. Translators can, however, adjust the rate of similarity. If a match is detected, translators can select the target segment and insert it in the active editing area. Furthermore, when a source segment is selected, the editor automatically searches existing community dictionaries and displays the translation(s) of words detected in the source segment. After finishing translation, BEYTrans creates a new version of the TC and sends it to the repository. As it is saved in Wiki mode, it is easy to follow the modifications done in previous versions. This enhances the efficiency of collaborative translation, as translators have a simple and concrete mechanism to check modifications and view previous versions.

In the following sub-section, we outline the technical solution for the implementation of BEYTrans. Different questions related to document management, the editor and language resources will then be addressed.

6.2 Translation modules and language help

6.2.1 *Document segmentation*

The uploaded documents are subject to multilingual text extraction and segmentation for TU boundary detection which allows construction of the TC and insertion of TUs in the central TM (Walker et al. 2001). We have exploited the efficiency of LingPipe tools (LingPipe 2007), which deal with sentence-boundary and language unit detection (e.g., named-entity detection, etc.). This tool can be extended to support additional languages and trained resources for more precision, but translators can exploit the Excel-like editor to improve the automatic segmentation and generate new segments manually.

However, the detected TUs are managed in the TC format, which contains all segments of all languages in which the document has been translated. Translators can display all TUs in parallel format or select those of specific language pairs.

The TUs are not only processed at the segmentation stage; it is important also to note that once the translation is finished, new segments (added, deleted or updated) are re-structured in a new version of the translation companion (in TMX-C format). After each translation, the new translation companion of the current document containing the new TUs is sent back to the repository. Accordingly, these segments are also sent to the CTM, in order to make them available to translators during the translation of future documents.

6.2.2 *Multilingual editor*

The BEYTrans online editor allows translators to edit in a rich environment that, among other functions, efficiently manages document formats, and includes language helps to speed up translation and improve quality. A set of translated segments is proposed by the CTM with dictionary suggestions. Translators can also call up several free web MT services to produce pre-translations (e.g., Systran, Reverso and Google), and improve them later by post-editing (Allen 2001).

The editor was developed separately and then integrated into the core of the Wiki. It consists of an Excel-like grid interface (based on free open source software) which displays in parallel format source and target translation units (Dhtmlgrid 2007). If they wish, translators can translate into many languages at the same time by adding new columns.

It is important to note that the editor data is managed as XML data. The textual segments (TUs) stored in the TC and CTM are transformed into an XML format compatible with the Excel-like grid. These segments are extracted using XML XPATH and inserted in a set of XML element "cells" as follow (XPath 2008): <rows> <row id="0"><cell> TU_1 </cell><cell> TU_2 </cell><cell> TU_3 </cell> ... </row> ... </rows>; where TU_i are translation units and the XML element "row" defines a set of TUs to be displayed in the same line (e.g., the source TU and its translation) (Fig. 4).

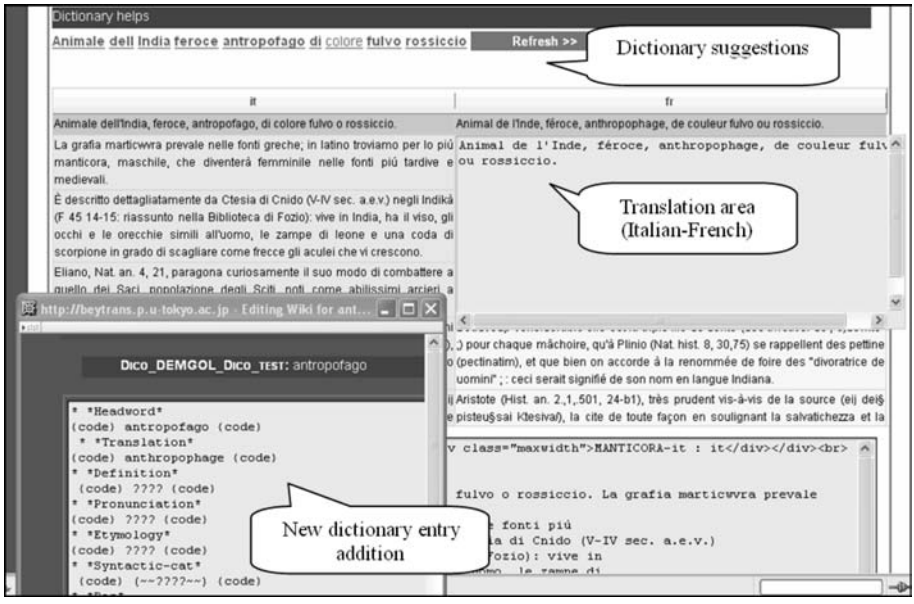

Figure 4. The BEYTrans multilingual editor and online linguistic help

7. Conclusion

In this chapter, we have described the BEYTrans prototype, an environment for helping online volunteer translators to produce high-quality translations of various types of documents. The first version has been implemented, and has been experimentally used by the DEMGOL project. We are also planning to have it tested by the FTEXT multilingualisation project (a Japanese project that makes free high school texts) (FTEXT 2007). Though a detailed analysis of user feedback is yet to be carried out, all in all we have had a very favorable reaction from the DEMGOL participants who have used BEYTrans for their translation work (DEMGOL 2007).

We hope BEYTrans will provide a way for all communities involved in volunteer translation to improve their skills and enhance quality. On the technical side, we have used the Wiki technology to develop collaborative and open editing functions on the Web, and we have integrated the management of translatable units and language resources using various XML annotation systems, sometimes adding original specifications necessary to our system. This work continues, with a view to provide online volunteer translators with important components for producing fast but high-quality translations into many languages, and also to "bootstrap" the system so that translators can manage language resources (dictionaries and TMs) using the same tool. We have already extended BEYTrans so that it is now possible to use it in order to edit and extend to other languages its own translation memories, viewed as a kind a large parallel documents of simple structure.

References

Abekawa, T., and Kageura, K. (2007) A Translation Aid System with a Stratified Lookup Interface. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Poster and Demo Session,* Prague, Czech Republic, 5–8.

Aguirre, E., X. Arregi, X. Artola, A. Díaz De Ilarraza, K. Sarasola & A. Soroa (2000) A Methodology for Building a Translator-oriented Dictionary System. In *Machine Translation,* 15 (4), 295–310.

Arabic Mozilla (2007) http://www.arabeyes.org

Allen, J. (2001) Postediting: An Integrated Part of a Translation Software Program (Reverso Pro 4). In *Language International Magazine*, 13 (2), 26–29.

Augar, N., R. Raitman & W. Zhou (2004) Teaching and Learning Online with Wikis. In *Proceedings of the 21st Australasian Society of Computers In Learning In Tertiary Education Conference (ASCILITE),* Western Australia, Deakin University, Australia, 95–104.

Bey, Y., C. Boitet & K. Kageura (2006a) The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators. In Yuste Rodrigo, E. (ed.) *Proceedings of the 3rd International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III),* 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 49–54.

Bey, Y., K. Kageura & C. Boitet (2006b) Data Management in QRLex, an Online Aid System for Volunteer Translators. In *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, 11(4), 349–376.

Boitet, C., Y. Bey, & K. Kageura (2005) Main Research Issues in Building Web Services for Mutualized, Non-Commercial Translation. In *Proceedings of the 6th Symposium on Natural Language Processing, Human and Computer Processing of Language and Speech (SNLP)*, Chiang Rai, Thailand.

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

Breen, J., W. (2001) A WWW Dictionary and Word Translator: Threat or Aid to Language Acquisition? In *Proceedings of the 6th Annual International Conference of the Japan Association for Language Teaching Computer-Assisted Language Learning Special Interest Group (JALT-CALL 2001)*, Gunma, Japan, 26–27.

DéjaVu (2007) http://www.atril.com/

DEMGOL (2007) http://demgol.units.it/

Dhtmlgrid (2007) http://sourceforge.net/projects/dhtmlgrid/

Global Voices (2007) http://www.globalvoicesonline.org/

French Mozilla (2007) French Mozilla Project, http://frenchmozilla.online.fr/

FTEXT (2007) http://www.ftext.org/

Fulford, H. (2001) Translation Tools: An Exploratory Study for Their Adoption by UK Freelance Translators. In *Machine Translation*, 16 (3), 219–232.

Fulford, H. and J. Granell Zafra (2004) The Uptake of Online Tools and Web-based Language Resources by Freelance Translators: Implications for Translator Training, Professional Development, and Research. In Yuste Rodrigo, E. (ed.) *Proceedings of the 2nd International Workshop on Language Resources for Translation Work, Research and Training (LR4Trans),* Geneva, Switzerland, 37–44.

Kageura, K., T. Abekawa & S. Sekine (2007) QRselect: A User-driven System for Collecting Translation Document Pairs from the Web. In *Proceedings of the 10th International Conference on Asian Digital Libraries,* Hanoi, Vietnam, 131–140.

LingPipe (2007) http://alias-i.com/lingpipe/demo.html/

LISA (2007) http://www.lisa.com

Mangeot, M. (2002) An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language. In *Proceedings of the International Workshop on Standards of Terminology and Language Resources Management,* LREC 2002, Las Palmas, Canary Islands, Spain, 37–44.

Mozilla (2007) http://www.mozilla.com/

Omega-T (2007) http://www.omegat.org/

Pax Humana (2007) http://paxhumana.info/

Queens, F. and U. Recker-Hamm (2005) A Net-Based Toolkit for Collaborative Editing and Publishing of Dictionaries. In *Literary and Linguistic Computing*, 20 (Suppl), 165–175.

Saha, G., K., A. (2005) Novel 3-Tiers XML Schematic Approach for Web Page Translation. In *ACM IT Magazine and Forum*. http://www.acm.org /ubiquity/views/v6i43_saha.html

Schwartz, L. (2004) Educational Wikis: Features and Selection Criteria. In *International Review of Research in Open and Distance Learning*. Athabasca University, Canada's Open University, 5 (1).

Streiter, O. and M. Mathiasser (2005a) XNLRDF, the Open Source Framework for Multilingual Computing. In *Proceedings of the Lesser Used Languages & Computer Linguistics Workshop,* at The European Academy Bozen/Bolzano, Italy, 189–207.

Streiter, O. and M. Stuflesser (2005b) XNLRDF, an Open Source Natural Language Resource Description Framework. In *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation (PACLIC19)*, Taipei, Taiwan, 305–312.

Streiter, O., M. Stuflesser & Q. LanWeng (2006) Models of Cooperation for the Development of NLP Resources: A Comparison of Institutional Coordinated Research and Voluntary Cooperation. In *Proceedings of the 5th Workshop on Minority Languages,* 5th International Conference on Language Resources and Evaluation (LREC 2006), at Magazzini del Cotone Conference Centre, Genoa, Italy, 109–112.

Tea Not War (2008) http://teanotwar.seesaa.net/

Trados (2007) http://www.sdl.com/en/products/products-index/sdl-trados/

Traduc (2007) http://wiki.traduc.org/

Translationwiki (2007) http://www.translationwiki.com/

W3C (2007) http://www.w3.org/

Walker, D. J., D. E. Clements, M. Darwin & W. Amtrup (2001) Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality. In *Proceedings of the VIII Machine Translation Summit,* Santiago de Compostela, Spain, 369–372.

Wiktionary (2007) http://en.wiktionary.org

XPath (2008) http://www.w3.org/TR/xpath

XWiki (2007) http://www.xwiki.com

CHAPTER 10

Standardising the management and the representation of multilingual data

The Multi Lingual Information Framework

Samuel Cruz-Lara, Nadia Bellalem,
Julien Ducret, and Isabelle Kramer
LORIA / INRIA, Nancy University

Due to the critical role that normalisation plays during the translation and localisation processes, we propose here to analyse some standards, as well as the related software tools that are used by professional translators and by several automatic translating services. We will first point out the importance of normalisation within the translation and localisation activities. Next, we will introduce a methodology of standardisation, whose objective is to harmonise the management and the representation of multilingual data. The control of the interoperability between the industrial standards currently used for localisation [XLIFF], translation memory [TMX], or with some recent initiatives such as the internationalisation tag set [ITS], constitutes a major objective for a coherent and global management of multilingual data. The Multi Lingual Information Framework MLIF [ISO AWI 24616] is based on a methodology of standardisation resulting from the ISO (sub-committees TC37/SC3 "Computer Applications for Terminology" and SC4 "Language Resources Management"). MLIF aims at proposing a high-level abstract specification platform for a computer-oriented representation of multilingual data within a large variety of applications such as translation memories, localisation, computer-aided translation, multimedia, or electronic document management. The major benefit of MLIF is interoperability because it allows experts to gather, under the same conceptual model, various tools and representations related to multilingual data. In addition, MLIF should also make it possible to evaluate and to compare these multilingual resources and tools.

1. Introduction

The extremely fast evolution of the technological development in the sector of Communication and Information Technologies, and in particular, in the field of

natural language processing, makes particularly acute the question of standardisation. The issues related to this standardisation are of an industrial, economic and cultural nature. Nowadays, an increasing number of standards are frequently being used within most scientific and technical domains. Translation and localisation activities just cannot remain isolated from this important and novel situation. The advantages of normalisation are currently fully recognised by most professional translators: using standards means working with a high level of quality, performance, and reliability within a very important market that is becoming more and more global and thus more and more challenging. Indeed, the standards combine simplicity and economy by reducing the planning and production costs, and by unifying several kinds of terminology (i.e., validated vocabulary) and several kinds of products. At the national and international levels, standards stimulate cooperation between different communities of trade while ensuring interoperability within information exchanges and reliability of all generated results by using standardised methods and procedures; this is why normative information has became fundamental.

The scope of research and development within the localisation and translation memory process development is very large. Several industrial standards have been developed: TMX, XLIFF, OLIF, etc. However, when we closely examine these different standards or formats by subject field, we find that they have many overlapping features. All the formats aim at being user-friendly, easy-to-learn, and at reusing existing databases or knowledge. All these formats work well in the specific field they are designed for, but they lack the synergy that would make them interoperable when using one type of information in a slightly different context. Modelisation corresponds to the need to describe and compare existing interchange formats in terms of their informational coverage and the conditions of interoperability between these formats, and hence the source data generated in them. One of the issues here is to explain how a uniform way of documenting such data that takes into account the heterogeneity of both their formats and their descriptors.

We also seek to answer the demand for more flexibility in the definition of interchange formats so that any new project may define its own data organisation without losing interoperability with existing standards or practices. Such an attempt should lead to more general principles and methods for analysing existing multilingual databases and mapping them onto any chosen multilingual interchange format.

2. Normalisation: A key issue for translation

The translator is the most important element of the translation process: because of his experience and knowledge he ensures that the translated document is accurate with respect to the original one (Gómez & Pinto 2001). A good translation does not only need linguistic awareness but it also needs a good knowledge of the technical or scientific field of the documents that have to be translated. Most texts are addressed to specialists and ignorance of the specialised expressions can justifiably cause rejection by the reader. In the same way, English technical terms – whose equivalents however exist in the target language – are often used just as they are in French, for example. Obviously, a technical translation is not limited to the data processing or computer science translation. A technical translator must have, in addition to his knowledge, a high-quality set of documents. Even if his level of knowledge is high, he must continuously seek advice from technical documents (i.e., journals, specialised dictionaries, magazines, databases, etc.). Somehow these technical documents correspond to a set of essential tools allowing a translator to analyse the information located on the covered subject. So, a translator must evolve constantly by acquiring new information and new experiences, because this way he will obtain additional linguistic and non-linguistic knowledge related to the domains he has to deal with (Hurtado Albir 1996).

Standards constitute a fundamental tool for translators, because they will provide, as high-level models for technical specifications (i.e., symbols, definitions, codes, figures, methodology, etc.), abundant, exact, and about all, interoperable and reliable information. Unfortunately, several fields – and specially translation – have numerous and often non-compatible standards. This requires a parallel activity of normalisation of these standards in order to ensure, at least, a minimal degree of compatibility. This activity constitutes the main issue of the "Documentation on standards" or the "Structured set of standards" (Pinto 93). As the texts, objects of translation, have a great diversity of subjects, the documentary activities applied to the standards can also guide and direct the translator in the search for standards relative to a given field. The production of standards became sometimes so prolific that it is quite difficult to understand exactly what method or procedure has to be used.

The documentary techniques bring essential assistance. The services of information and dissemination of the national and international organisations of standardisation give easy access to all this information. The standards worked out by the organisations of standardisation (i.e., ISO, W3C, LISA, etc.) are more and more accepted by the translation services and the translators with an aim of guaranteeing a high-level quality in their services and their products. Standardisation thus becomes a synonym of quality for the customers who wish the best possible results. In addition, standards represent also an essential tool for translators,

as they aim at creating normalised terminological and methodological linguistic resources, in order to improve the national and international exchanges as well as cooperation within all possible fields. It is also necessary to point out the important efforts of standardisation of ISO and W3C in the field of information technologies, especially those related to the computerised processing of multilingual data and the Internet.

Standardisation is present during the whole process of translation. So, the access to the normative information represents a stage impossible to circumvent within the activity of translators. Within this task (i.e., standardisation) translators are assisted not only by resource centers charged to disseminate the information of the standards worked out by the national and international organisations, but also by other specialised private agencies (i.e., PRODOC, http://www.prodoc.de) whose main objective is to advise the customers with regards to standards. In the same way, the Internet allows access to thousands of websites (i.e., terminology trade, databases, etc.) that provide important information related to using standards, as well as access to several interesting research projects in progress whose objectives are the development of standards and recommendations in the field of the "industry of the language".

2.1 Translation memories and translators

Among various documentations available for translators, translation memories (TM) occupy a dominating place. Translation memories are built from already translated texts and constitute an assistance for translating repetitive texts, as well as for performing searching operations on terminology databases. However, there does not exist yet a standard for the development and the management of translation memories, but rather some standards related to techniques and methods belonging to the field of document management and indexing. That is the reason why it is necessary to take into account the standards related to human indexing: the ISO 5963: 1985 standard encourages the services of indexing and other resource centers to unify their practices. One can divide translation memories into two great classes: memories built starting from an indexing of complete sentences, and memories built starting from an indexing of all words. The method of textual searching will thus be determined by the technique of indexing having been used.

For the evaluation of the quality of software related to computer-aided translation (CAT), we have the ISO/IEC standard 9126: 1991. This ISO/IEC standard defines the requirements for quality, the methods of evaluation, and the application of the procedures of evaluation. Within this context, it is suitable to point out the work of the Expert Advisory Group on Language Engineering Standards (EAGLES). The activities of this group are related to multilingual electronic dictionaries, multilingual electronic thesaurus, terminology database management

systems, and translation memories. However, the EAGLES group does not aim at producing international standards but rather to present the needs and the requirements of operational applications and to accelerate the process of standardisation in this matter.

There are no specific standards for automatic machine translation either. However, CAT systems were subjected to many evaluations thus making it possible to gradually improve the methodologies used for these evaluations. This is the reason why some of these evaluations can be considered as being *de facto* standards for the future evaluation of CAT technologies.

On its side, the LISA organisation proposes a recommendation called Translation Memory Exchange (TMX) that aims at facilitating the exchange of the data related to translation memories between tools and software CAT systems. Although TM Tools are based on the same basic idea, we must note that for the same sentence each tool proposes rather different ways to implement the required formatting information: on the one hand, formatting is applied to the source and target texts of a translation unit and this formatting is not exported to the corresponding TMX file; on the other hand, formatting is sometimes exported to the TMX file. In the following table (see Table 1), the sample sentence "the **sentence** contains *different* formatting information" is represented in TMX by using several tools (Zerfaß 2005). Some of these tools use external files to store formatting information (i.e., Déjà Vu and SDLX), but all of them use different ways of encoding that information.

Table 1. Comparison of formatting across tools

TRADOS 6.5	DÉJÀ VU	SDLX
<seg> This <ut>{\b </ut>sentence<ut>}</ut> contains <ut>{\i </ut>different<ut>}</ut> <ut>{\ul </ut>formatting information<ut>}</ut>. </seg>	<seg> <ph x="1">{1}</ph>This <ph x="2">{2}</ph> sentence <ph x="3">{3}</ph> contains <ph x="4">{4}</ph>different <ph x="5">{5}</ph><ph x="6">{6}</ph>formatting information <ph x="7">{7}</ph>. </seg>	<seg>This <bpt i="1"x="1"> <1></bpt>sentence <epti="1"></1></ept> contains <bpt i="2"x="2"> <2></bpt>different <epti="2"> </2></ept> <bpt i="3"x="3"><3></bpt> formatting information<epti="3"></ 3></ept>. </seg>

In addition, the segmentation rules used by TM tools are not compatible: each tool applies its own rule to split the text into various segments. In a same sentence some tools consider various separators. For example the semi-colon is

considered as a separator for Déjà Vu, but not for SDLX. Segmentation organises and structures the data. If everyone uses his own rules, the exchange is no more possible; that's why SRX for several years tries to normalise segmentation rules. SRX guidelines are useful to evaluate translation memory qualities and ensure interoperability of multilingual data.

2.2 Standards: Proliferation and necessity

As we have previously mentioned, we have to deal with the growing number of standards. Succession in standardisation is usually a problem (Egyedi & Loeffen 2002). The advantages of improvements are weighed against those of compatibility. This evolution could be easily explained because the priorities in standardisation could change, so the rules for developing standards are revised. Standards could be updated or become obsolete. This is part of the dynamics of standardisation, irrespective of the area of interest.

A number of critical problems in the field of Information and Communication Technology (ICT) occur because many standards have functional equivalents. That is, they address the same problem and offer similar functionalities. Sometimes competition between them leads to "standards wars".

Completeness has been identified as an important design criterion for interchange formats but less attention has been paid to the sequential relations between standards, that is, the way that previous standards (i.e., predecessors) are revised and succeeded by new standards (i.e., successors). Succession in standardisation implies change and renewal. Renewal comes in various shapes: new editions, revisions (i.e., new versions, technical corrigenda, amendments, annexes etc.) and new standards. The successor addresses the same area, and is an improvement on its predecessor. It is designed to succeed and thus take over the predecessor's role. New entrants in the market (standards users) naturally prefer and implement the successor.

Those who standardise the successor may or may not seek compatibility with the predecessor. They usually do, and need to have good reasons not to seek compatibility (e.g., technically impossible or a change in the product). There are many kinds of compatible successors. The most common one is the downward compatible successor, which replaces the more elaborated original standard.

If the successor standard is compatible, compliant technologies should be able to work together with products that interoperated with its predecessor. Such is typically the aim when the successor is a new edition or a minor revision of a standard. Examples are incremental innovations: the improvements made are part of normal problem solving. Dilemmas regarding compatible succession are often of a mixed socio-technical nature (i.e., technical, implementation, esthetic, etc.). A characteristic of dilemmas is that the conflicting arguments are both persuasive.

Within the framework of the management of multilingual content, some standards as TMX – related to translation memories – and XLIFF – related to the activity of localisation – have right now some dedicated software, as well as several resources respecting their respective recommendations. Although they are not at all out of date, these standards however cannot satisfy the needs being born from new information technologies.

Within ISO's TC34/SC4 "Linguistic Resources Management", a group of experts is currently working on the specification of a new standard aiming at, on the one hand, covering the whole functionalities of the above mentioned standards, and on the other hand, satisfying the linguistic enrichment and the interoperability of multilingual data: the "Multi Lingual Information Framework" (MLIF) is currently being developed. MLIF is an ISO's "Approved Work Item" [AWI 24616] from TC37/SC4, working group WG3 "Multilingual Text Representation" (see Table 2).

Table 2. Global comparison of TMX, XLIFF and MILF

	TMX	**XLIFF**	**MLIF**
Related Domains	Translation, Computer Assisted Translation (CAT)	Localization, Computer Assisted Translation (CAT), word processing program, terminology management systems, multilingual dictionary, or even raw machine translation output	Localization, Computer Assisted Translation (CAT) tool, word processing program, terminology management systems, multilingual dictionary, or even raw machine translation output, e-learning, multimedia applications, ...
Global Information (ex : date, author, ...)	Available on the head and on the meaning units.	Global	Available on the head and at the non-terminal levels of the model
Multilingual data	Multilingual	Bilingual	Multilingual
Possibility to use additional linguistic information	No	No	Yes. Terminological data, Lexical data, Morphological data, ...
Segmentation	Textual segments	Blocks, paragraphs, sentences, or phrases	Blocks, paragraphs, sentences, or phrases
Internal or external references	External	External	Internal (ex: anaphoric references, ellipse, ...) External (ex: data bases, terminology, ...)
Missing translation	Ignored	Ignored	Indicated

3. Contribution of standards

As we previously discussed, the life cycle of standards is conditioned by new needs, adaptations to technologies, or new trades. It is important to determine the fields concerned, as well as the concerned people and their work practices. The work practices make it possible to determine the minimum lattice of information to represent, as well as the set of features needed to specify for rendering this information relevant within a given framework. A multilingual software product should aim at supporting document indexing, automatic and/or manual computer-aided translation, information retrieval, subtitle handling for multimedia documents, etc. Dealing with multilingual data is a three steps process: production, maintenance (i.e., update, validation, correction) and consumption (i.e., use). A specific user group and a few specific scenarios correspond to each one of these steps. It is important to draw up a typology of the potential users and scenarios of multilingual data by considering the various points of view: production, maintenance, and consumption of these data. Indeed, we are not just trying to develop a new standard, nor we are aiming at competing with any existing standard. Rather, we are trying to specify a high-level model allowing to represent and to integrate the whole set of actors of the translation and localisation community. This is the reason why the participation of these actors to our work is a fundamental issue in the aim of the creation of a successful new standard.

The development of scenarios considers the possible limits of a multilingual product, thus the adaptations required. Normalisation will also allow the emergence of new needs (e.g., addition of linguistic data like grammatical information). Scenarios help to detect useless or superseded features that may not be necessary to implement within standardised software applications. These scenarios must also be based on well "on work practices" while also making it possible to envisage some possible extensions. Normalisation will facilitate the dissemination (i.e., export multilingual data) as well as the integration of data (i.e., import of multilingual data from external databases).

Providing normalised multilingual products and data must be considered as a way, for a scientific community, to be well known, to be acknowledged.

4. Terminology and methodology of normalisation

In this section, we will introduce our methodology of normalisation as well as the terminology related to our standardisation activities. Like any other technical field, standardisation has its own terminology and its own specific rules. As with the "Terminological Markup Framework" TMF (ISO 16642) in terminology, MLIF will introduce a structural skeleton (metamodel) in combination with

chosen "Data Categories", as a means of ensuring interoperability between several multilingual applications and corpora. Each type of standard structure is described by means of a three-tiered information structure that describes:

– a metamodel, which represents a hierarchy of structural nodes which are relevant for linguistic description;
– several specific information units that can be associated with each structural node of the metamodel;
– several relevant annotations that can be used to qualify some part of the value associated with a given information unit.

4.1 What is a metamodel?

A metamodel does not describe one specific format, but acts as a high level mechanism based on the following elementary notions: structure, information and methodology. The metamodel can be defined as a generic structure shared by all other formats and which decomposes the organisation of a specific standard into basic components. A metamodel should be a generic mechanism for representing content within a specific context. Actually, a metamodel summarises the organisation of data.

The structuring elements of the metamodel are called "components" and they may be "decorated" with information units. A metamodel should also comprise a flexible specification platform for elementary units. This specification platform should be coupled to a reference set of descriptors that should be used to parameterise specific applications dealing with content.

4.2 What is a data category?

A metamodel contains several information units related to a given format, which we refer to as "Data Categories". A selection of data categories can be derived as a subset of a Data Category Registry (DCR) (ISO 12620). The DCR defines a set of data categories accepted by an ISO committee. The overall goal of the DCR is not to impose a specific set of data categories, but rather to ensure that the semantics of these data categories is well defined and understood.

A data category is the generic term that references a concept. There is one and only one identifier for a data category in a DCR. All data categories are represented by a unique set of descriptors. For example, the data category */languageIdentifier/* indicates the name of a language which is described by 2 (ISO 639-1) or 3 (ISO 639-2) digits. A Data Category Selection (DCS) is needed in order to define, in combination with a metamodel, the various constraints that apply to a given domain-specific information structure or interchange format. A DCS

and a metamodel can represent the organisation of an individual application, the organisation of a specific domain.

4.3 Methods and representation

The way to actually implement a standard is to instantiate the metamodel in combination with a set of chosen data categories (DCS). This includes mappings between data categories and the vocabularies used to express them (e.g., as an XML element or a database field). Data category specifications are, firstly used to specify constraints on the implementation of a metamodel instantiation, and secondly to provide the necessary information for implementing filters that convert one instantiation to another. If the specification also contains styles and vocabularies for each data category, the DCS then contributes to the definition of a full XML information model which can either be made explicit through a schema representation (e.g., a W3C XML schema), or by means of filters allowing the production of a "Generic Mapping Tool" (GMT) representation. The architecture of the metamodel, whatever the standard we want to specify, remains unchanged. What are variable are the data categories selected for a specific application. Indeed, the metamodel can be considered in an atomic way, in the sense that starting from a stable core, a multitude of data can be worked out for plural activities and needs.

5. Specifying the Multi Lingual Information Framework

Linguistic structures exist in a wide variety of formats ranging from highly organised data (e.g., translation memory) to loosely structured information. The representation of multilingual data is based on the expression of multiple views representing various levels of linguistic information, usually pointing to primary data (e.g., part-of-speech (POS) tagging) and sometimes to one another (e.g., references, annotations). The following model identifies a class of document structures that could be used to cover a wide range of multilingual formats, and provides a framework that can be applied using XML.

All multilingual standards have a rather similar hierarchical structure but they have, for example, different terms and methods of storing metadata relevant to them. MLIF is being designed in order to provide a generic structure that can establish a basic foundation for all these standards. From this high-level representation we are able to generate, for example, any specific XML-based format: we can thus ensure the interoperability between several standards and their committed applications.

5.1 Description of MLIF[1]

The MLIF metamodel is constituted by the following components:

Multi Lingual Data Collection
Represents a collection of data containing global information and several multilingual units.

Global Information
Represents technical and administrative information applying to the entire data collection. Example: title of the data collection, revision history, . . .

Multi Lingual Component
This component represents a unique multilingual entry.

Mono Lingual Component
Part of a multilingual component containing information related to one language.

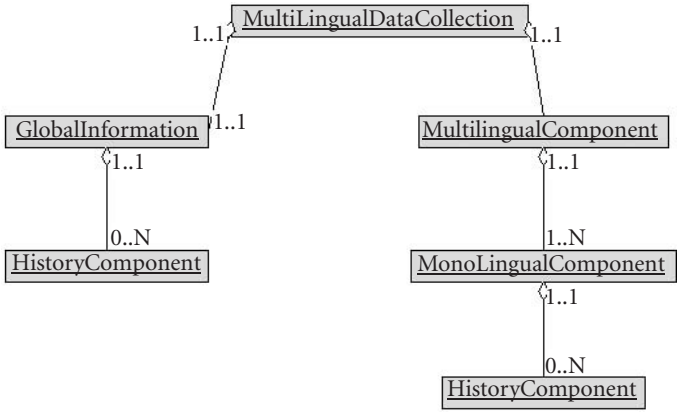

Figure 1. Hierarchical representation of MLIF

History Component
This generic component allows modifications to be traced on the component it is anchored to (i.e., versioning).

In order to provide a larger description of the linguistic content, the MLIF metamodel allows anchoring of other metamodels, such as MAF (morphological

1. This presentation of MLIF is based on the "New Work Item Proposal" (NWIP) submitted to ISO TC37/SC4 "Linguistic Resources Management". This NWIP was approved after an international ballot process in August 2006. MLIF is now an ISO's "Approved Work Item".

description), SynAF (syntactical annotation), TMF (terminological description), or any other metamodel based on ISO 12620: 2003.

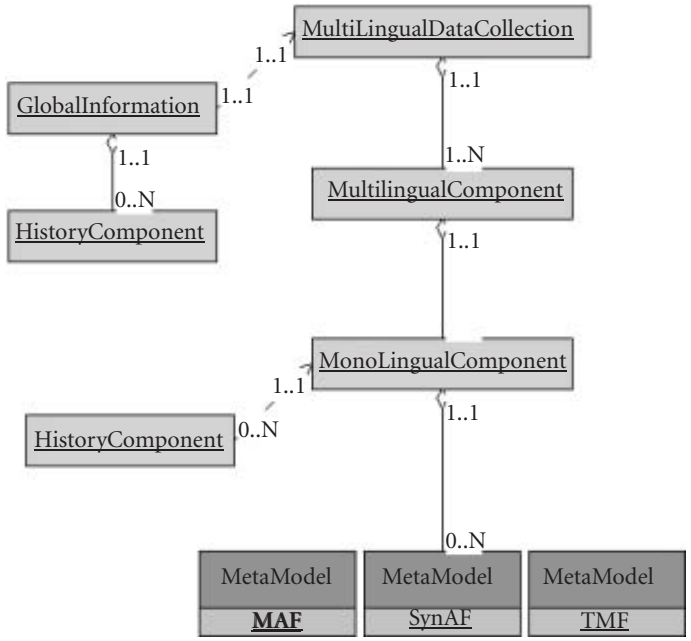

Figure 2. MLIF metamodel

All the different models have very similar hierarchical structure but they have different terms and methods of storing metadata relevant to them in particular. MLIF provides a generic structure that can establish basic foundation for all other models. This model will provide flexibility to make any element or attribute of this format to be defined explicitly or not. If the value is not defined explicitly it will take default value. Most of the models will also define their own elements and attributes those will fit into this using extensibility that is one of the basic requirements of MLIF model.

5.2 Data Categories

Global Information
/source/

– "A complete citation of the bibliographic information pertaining to a document or other resource".
– Reference to a resource from which the present resource is derived.

/sourceType/

- "In multilingual and translation-oriented language resource or terminology management, the kind of text used to document the selection of lexical or terminological equivalents, collocations, and the like".
  - "Both parallel and background texts serve as sources for information used in documenting multilingual terminology entries." (ISO 12620)

/sourceLanguage/

- "In a translation-oriented language resource or terminology database, the language that is taken as the language in which the original text is written".

/projectSubset/

- An identifier assigned to a specific project indicating that it is associated with a term, record or entry.

/subjectField/

- "A field of special knowledge."

Multilingual Component
/identifier/

- A unique name [source:IMDI_Source_Tag]
  - Dublin Core equivalent: DC:Identifier [source:IMDI_Source_Tag]

Monolingual Component
/languageIdentifier/

- A unique identifier in a language resource entry that indicates the name of a language.

/primaryText/

- Linguistic material which is the object of study.

/sourceLanguage/

- "In a translation-oriented language resource or terminology database, the language that is taken as the language in which the original text is written".
  - The identifiers specified in ISO 639 should be used:
    * en = English
    * fr = French
    * es = Spanish (Español)
    * ...

5.3 Introduction to GMT

GMT "Generic Mapping Tool" can be considered as an XML canonical representation of the metamodel. The hierarchical organisation of the metamodel and the qualification of each structural level can be realised in XML by instantiating the abstract structure shown above (see Fig. 3) and associating information units to this structure. The metamodel can be represented by means of a generic element <struct> (for structure) which can recursively express the embedding of the various representation levels of a MLIF instance. Each structural node in the metamodel shall be identified by means of a type attribute associated with the <struct> element. The possible values of the type attribute shall be the identifiers of the levels in the metamodel (i.e., Multilingual Data Collection, Global Information, Multilingual Component, Monolingual Component, Linguistic Element).

Basic information units associated with a structural skeleton can be represented using the <feat> (for feature) element. Compound information units can be represented using the <brack> (for bracket) element, which can itself contain a <feat> element followed by any combination of <feat> elements and <brack> elements. Each information unit must be qualified with a type attribute, which shall take as its value the name of a standard data category (ISO 12620) or that of a user-defined data category.

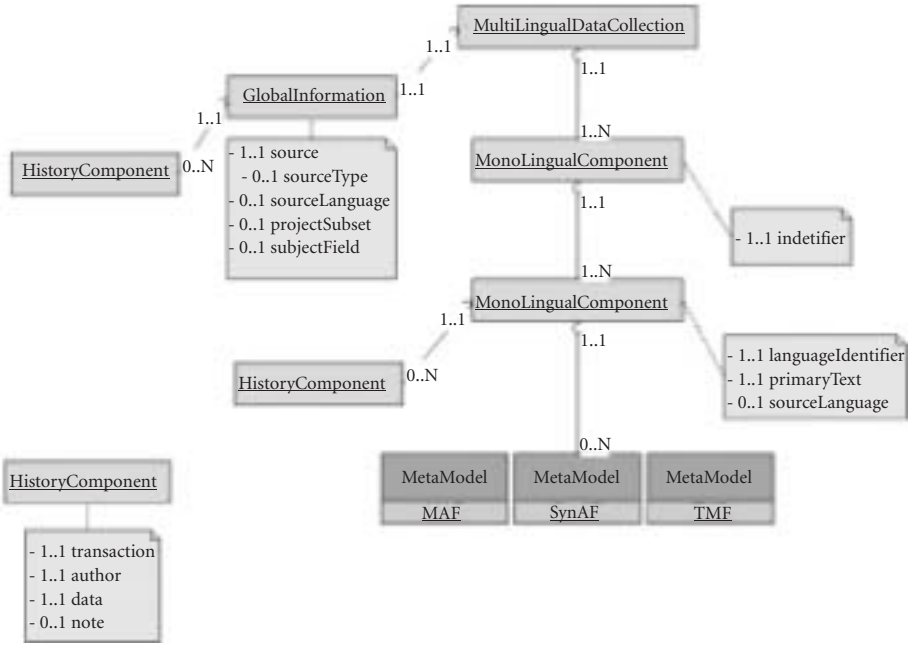

Figure 3. MLIF metamodel with selected "data categories"

5.4 A practical example: MLIF and TMX

Now, we will use a very simple TMX example (see Fig. 4) for the purpose of showing how MLIF can be mapped to other formats. As we discuss further details about MLIF, it will be clear that all features can be identified and mapped through data categories.

In Fig. 4, we found structural elements of TMX : 1 represents the <tmx> root element, 2 the <header> element, 3 represents a <tu> element, 4 and 4' represent respectively the English and French <tuv> element. Next, we will match these structural elements of TMX with the metamodel of MLIF (see Table 3).

Then, we will tag each element descriptor of TMX into 3 types: attribute, element or typed element. All these descriptors will be standardised into a MLIF

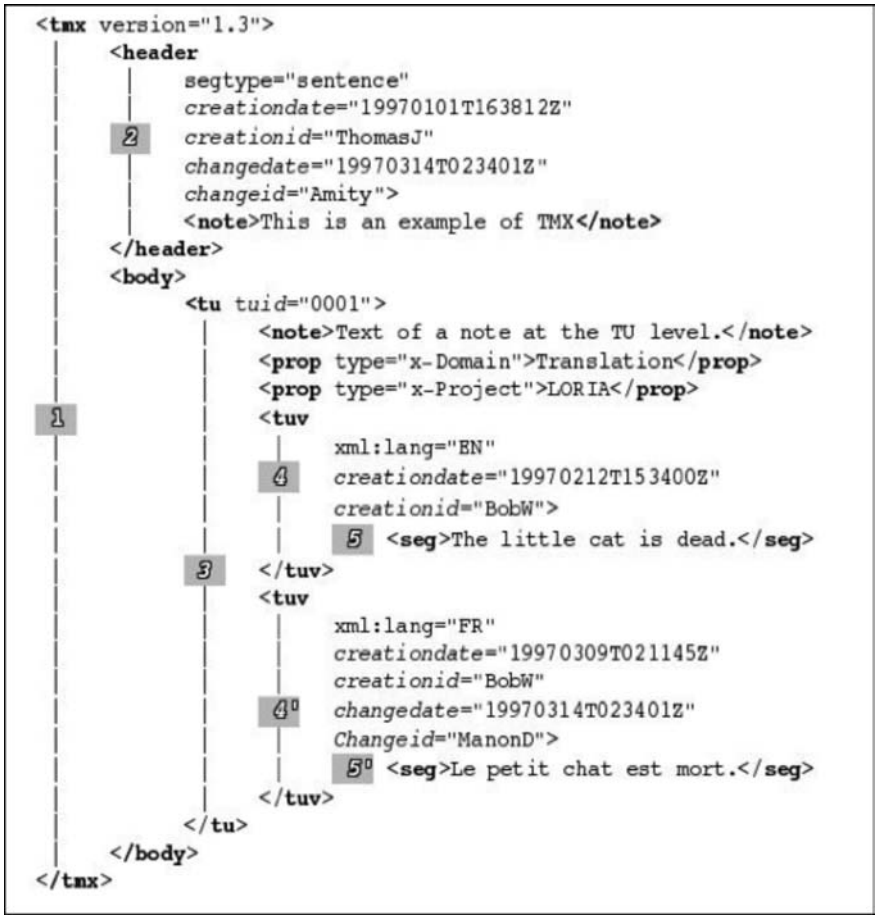

Figure 4. Part of a TMX document

Table 3. Matching TMX with MLIF components

TMX structure	MLIF component
1 <tmx>	Multilingual Data Collection
2 <header>	Global Information
3 <tu>	Multilingual Component
4 <tuv>	Monolingual Component

descriptor element (i.e., a data category, as shown below). For example, the TMX "xml:lang" attribute will be next matched with the data category named /languageIdentifier/, as shown in Table 4.

Table 4. Typing of descriptor elements and matching with data categories

TMX descriptor	Type	Data Categories
<note>	*element*	*/note/*
<prop type="*x-project*">	*typed element*	*/projectSubset/*
xml:lang	*attribute*	/languageIdentifier/
tuid	*attribute*	*/identifier/*
<seg>	*element*	*/primaryText/*

Finally, the mapping of TMX elements into MLIF elements is represented in the following GMT file (see Fig. 5). Note that this GMT file is nothing but a canonical representation of a MLIF document.

6. Interoperability, adaptation, and some other important issues

The principles of TMF (ISO 16642) and, by extension those of MLIF, can be translated in the form of formal constraints making it possible to precisely specify the informational cover of a given model and, by doing this, to compare two models with the objective to determine the precise conditions of interoperability. Next, we must be able to translate these constraints into XML-related structures, in order to provide true data models, in particular those which will be used later for the construction of software tools being able to handle and, to exchange multilingual data. This is the stage that we present in this last section, while seeking again to define a general enough framework that would be able to cover a broad variety of applications.

```
<struct type="Multilingual Data Collection">
      <struct type="Global Information">
            <brack>
                  <feat type="transaction">creation</feat>
                  <feat type="date">19970101T163812Z</feat>
                  <feat type="author">ThomasJ</feat>
            </brack>
            <brack>
                  <feat type="transaction">modification</feat>
                  <feat type="date">19970314T023401Z</feat>
                  <feat type="author">Amity</feat>
            </brack>
            <feat type="note"> This is an example of TMX</feat>
      </struct>
      <struct type="Multilingual Component">
            <feat type="identifier">0001</feat>
            <feat type="note"> It's just an example</feat>
            <feat type="subjectField">Translation</feat>
            <feat type="projectSubset">LORIA</feat>
            <struct type="Monolingual Component">
                  <feat type="languageIdentifier">EN</feat>
                  <brack>
                        <feat type="transaction">creation</feat>
                        <feat type="date">19970212T153400Z</feat>
                        <feat type="author">BobW</feat>
                  <brack>
                  <struct type="Lingustic Segment">
                        <feat type="sentence">The little cat is dead.</feat>
                  </struct>
            </struct>
            <struct type="Monolingual Component">
                  <feat type="languageIdentifier">FR</feat>
                  <brack>
                        <feat type="transaction">creation</feat>
                        <feat type="date">19970309T021145Z </feat>
                        <feat type="author">BobW</feat>
                  <brack>
                  <brack>
                        <feat type="transaction">modification</feat>
                        <feat type="date">19970314T023401Z </feat>
                        <feat type="author">ManonD</feat>
                  </brack>
                  <struct type="Lingustic Segment">
                        <feat type="sentence">Le petit chat est mort.</feat>
                  </struct>
            </struct>
      </struct>
</struct>
```

Figure 5. GMT representation

6.1 Multiple utilisations and adaptation

Roughly, as one single format for representing multilingual data is likely always to be regarded either as too complex, or like not answering exactly such or such particular need (Romary et al. 2006). We actually wish to show that it is possible to consider a family of formats, within a sufficiently accessible framework of representation, that many categories of users can easily adapt the groundwork suggested to their own needs. We are thus positioned in the continuity of thoughts

carried out by the actors of standardisation themselves (Romary 2001; Bauman & Flanders 2004) which consider that the proposal of a framework of standardisation is not incompatible with the identification of operations of adaptation of the standards by extension or restriction of a data model. Actually, it is a question of transposing, within the field of the representation of data, the concept of "subsumption" of models. The objective is to arrive to the definition of a true platform of specification of multilingual data which is capable of guaranteeing that the same element of data (i.e., utilisation or reference to data categories) will be represented in an identical way in two different applications and thus to avoid the trap which locks up the standard in a yoke of too specific applications.

The choices that we present here were guided by another important concern, namely the need for foreseeing the potential integration of multilingual data within a broader framework of representation of textual documents. Indeed, we think that multilingual data must not be dissociated from the documents where they are used. Concretely, that means that within multilingual textual documents, we must be able for example, to annotate and to connect all multilingual terms being used, or be able to establish links with the entries of some terminology database.

From this point of view, one can of course mention the official standardisation documents of ISO which integrates, as a mandatory section, the whole set of the terms and definitions used within the body of these documents.

Within another different context (i.e., captions within DVD movies), multilingual textual information may also need to be structured in different ways (i.e., paragraphs, sentences, but also surface annotations) that those related to the field of the translation memories or localisation. It is thus important, since that seems to be possible, to offer a representation of multilingual data which is integrated within a broad framework of representation of textual information.

6.2 Working within the scope of the TEI

The TEI (Text Encoding Initiative: http://www.tei-c.org) is an international initiative which, since 1987, gathers together most of the actors who have to manage great projects of textual data. The TEI covers many applicative domains such as prose, poetry, theatre, manuscripts, and dictionaries and is strongly concerned with multilingualism issues. Today, the TEI offers a platform of specification, ODD (One Document Does it all), which is an ideal framework to implement the approach which we defend here, namely the definition of a family of compatible models.

ODD is a language of data specification which is based on the principle of "literate programming" which combines descriptive elements with formal elements in order to provide, starting from a single source, at the same time: a diagram allow-

ing to control the effective syntax of a document, and a documentation providing to a user the fine semantics of the objects defined in the specification. Without going here into too technical details, we show here the two essential characteristics of ODD, namely the concepts of modules and classes, for providing next, some indications on the specification of objects XML themselves.

The ODD platform makes it possible to organise any documentary structure as a combination of one or more modules joining together a coherent unit of elements and classes. The directives of the TEI propose modules thus making it possible to represent the heading of a document, the common elements (e.g., divisions) to all types of documents, the elements specific to theatre, poetry, etc. A user can thus decide to use the basic modules allowing to represent simple textual data and to associate it to a terminological module, in order to insert descriptions of terms in the body text.

Two principal types of objects are described inside a module, elements and classes. The classes allow to gather elements having a syntactic behaviour or a similar semantics. Thus, all elements giving any morpho-syntactic indication within a dictionary or a terminology database belong to the class "model.morphLike". This way, if one wishes to integrate all these elements within a model of contents, it is enough to refer to the related class. In a complementary way, if one wishes to add a morpho-syntactic descriptor, it is enough to add an element to the class.

For the definition of the models of contents, the TEI is based on elementary fragments of RelaxNg diagrams which are then combined to generate complete RelaxNg diagrams, but also DTD's XML, or W3C XML Schemas.

7. Perspectives

A first implementation of MLIF within multimedia applications has been used within several prototypes developed in the framework of the ITEA Passepartout project (ITEA 04017). Within these prototypes some basic scenarios have been implemented: MLIF has been associated to XMT (eXtended MPEG-4 Textual format) and to SMIL (Synchronised Multimedia Integration Language). Our main objective in this project has been to associate MLIF to multimedia standards (e.g., MPEG-4, MPEG-7, and SMIL) in order to be able, within multimedia products, to represent and to handle multilingual content in an efficient, rigorous and interactive manner (see Fig. 6).

At present, we are also working on the issue of proposing several compatibility-related filters with ODD. Within a more practical framework, we are also developing a PHP multilingual gateway: all multilingual textual information is directly encoded by using MLIF.

Figure 6. Dynamic and Interactive displaying of multilingual subtitles and multilingual textual information

8. Conclusion

In this chapter, we have analysed why normalisation is a key issue within translation and localisation activities. Within this context, we have also shown that it is possible to define, in a coherent way, the various phases of designing a general normalised framework for the handling and representation of multilingual textual data within localisation and translation activities. The MLIF "Multi Lingual Information Framework" ISO is being developed this way. As we have clearly indicated, MLIF must be considered as a unified conceptual representation of multilingual content and is not intended to substitute or to compete with any existing standard. MLIF is being designed with the objective of providing a high-level common conceptual model and a platform allowing interoperability among several translation and localisation standards, and by extension, their committed tools. We think that this platform is a continuum between a truly linguistic work of collecting multilingual data and the development of a data-processing software environment intended to accommodate such data.

MLIF continues to evolve and within the next months an ISO's "Committee Draft" (CD) should be published. This CD will reflect comments and remarks from the MLIF's Experts Committee so the metamodel and related data categories

will certainly be modified. Also, as we have mentioned, our current research tends to prove that the specification of a format of representation such as MLIF can be elegantly associated with a broader normative approach, such as the TEI.

Last but not least, it is important to point out once again that MLIF has been successfully associated to multimedia standards such as XMT and SMIL. In our opinion, text must no longer be considered as the "ugly duckling" of multimedia.

References

Bauman S. & J. Flanders (2004) Odd Customizations. In *Proceedings of Extreme Markup Languages*, Montreal, Canada. 2nd–6th August 2004.

S. Cruz-Lara, S. Gupta & L. Romary (2004) Handling Multilingual content in digital media: The Multilingual Information Framework. In *EWIMT-2004 European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*. London, UK.

Egyedi, T. M. & A. G. A. J. Loeffen (2002) Succession in standardisation: Grafting XML onto SGML. In *Computer Standards & Interfaces*, 24, 279–290.

Gómez, C. & M. Pinto (2001) La normalisation au service du traducteur. In *Meta*, XLVI (3), 564.

Hurtado Albir, A. (ed.) (1996) *La enseñanza de la traducción*, Castellón, Spain: Universidad Jaume I.

ISO 12620: 1999. *Computer applications in terminology – Data categories*, Geneva, Switzerland: International Organisation for Standardisation.

ISO 16642: 2003. *Computer applications in terminology – Terminological markup framework (TMF)*, Geneva, Switzerland: International Organisation for Standardisation.

ISO 5963: 1985. *Documentation – Methods for examining documents, determining their subjects, and selecting indexing terms*, Geneva, Switzerland: International Organisation for Standardisation.

ISO 5964: 1985. *Documentation. Principes directeurs pour l'établissement et le développement de thésaurus bilingues*, Geneva, Switzerland: International Organisation for Standardisation.

ISO 639-1: 2002. *Code for the representation of names of languages – Part 1: Alpha-2 code*, Geneva, Switzerland: International Organisation for Standardisation.

ISO 639-2: 1998. *Code for the representation of names of languages – Part 2: Alpha-3 code*, Geneva, Switzerland: International Organisation for Standardisation.

ISO/IEC 9126: 1991. *Technologies de l'information. Évaluation des produits logiciels. Caractéristiques de qualité et directives d'utilisation.*

ISO WD 24611. *Morphosyntactic Annotation Framework (MAF)*. ISO/TC37/SC4/WG2 WD 24611.

ISO WD 24615. *Syntactical Annotation Framework (SynAF)*. ISO/TC37/SC4/WG2 WD 24615

ITEA "Information Technology for European Advancement". Passepartout project "Exploitation of advanced AV content protocols (MPEG 4/7)" ITEA 04017.

ITS. W3C (2003) *Internationalisation Tag Set (i18n)*. http://www.w3.org/TR/its/

Pinto, M. (1993) *Análisis documental. Fundamentos y procedimientos*, 2ª ed. rev. y aum., Madrid, Spain: Eudema.

Romary, L. (2001) *Un modèle abstrait pour la représentation de terminologies multilingues informatisées*, Cahiers GUTenberg, 39–40, 81–88.

Romary L., S. Salmon-Alt, I. Kramer, & J. Roumier (2006) Gestion de données terminologiques: principes, méthodes, modèles. In *Terminologie et accès à l'information*. Paris, France: Hermes, collection Techniques et traités des sciences et techniques de l'information.

Synchronised Multimedia Integration Language (SMIL 2.0). World Wide Web Consortium. http://www.w3.org/TR/smil20/

SRX (2004) *Segmentation Rules eXchange. SRX 1.0 Specification*. Oscar Recommendation 20 April 2004. http://www.lisa.org/standards/srx/srx.html

Oscar / Lisa (2000) *Translation Memory eXchange (TMX)*, http://www.lisa.org/tmx.

XLIFF (2003) *XML Localisation Interchange File Format*. http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff.

XMT. extended MPEG-4 Textual format. ISO/IEC FCD 14496-11, *Information technology – Coding of audio-visual objects – Part 11: Scene description and application engine*; ISO/IEC 14496-11/Amd 4, XMT & MPEG-J extensions.

Zerfaß, A. (2005) TMX and SRX Exchanging TM Data. In *Proceedings of the LRC-X Conference*, University of Limerick, Ireland. 13th–14th September 2005.

CHAPTER 11

Tagging and tracing Program Integrated Information

Naotaka Kato, and Makoto Arisawa
Translation Services Center, IBM Japan, Ltd. / Faculty of Graduate School of Media and Governance, Keio University

There are two main types of translation involving computer programs. One involves manuals and the other involves Program Integrated Information (PII). This chapter focuses on PII translation. PII translation is substantially different from ordinary text translation. PII is separated out of the programs into externalised text resource files to allow for translation outside the program development laboratory. The contexts of the operations have been discarded. The translators have to translate phrases and words without context in these text resource files. The Translation Verification Test (TVT), which is done with the normal operations of the program, compensates for the lack of context during translation. If the TVT tester finds an inappropriate translation in the GUI (Graphical User Interface), the file it came from and which line in the file is unknown. We have developed a utility program to make it easy to find a source location. The utility adds a short group of ID characters in front of every PII string. We used this systematic approach for CATIA®[1] and found many advantages, such as locating the hard-coded strings that are the biggest problems in program internationalisation. This ID can be inserted independently of program development. We also developed a utility program that helps TVT testers refer to both the original and target PII strings as pairs. This chapter describes the approach in detail. In addition, this chapter presents statistics about PII files. This important statistical information has not been considered in the program internationalisation communities.

1. Introduction

Program internationalisation often requires software developers to translate the strings of programs into nine or more languages. This translation task is not carried out in a software development laboratory but in an organisation that specialises in translation. If the text strings might need to be translated, the

1. CATIA, a CAD/CAM program, is a registered trademark of Dassault Systemes.

development laboratory externalises the strings from the programs into the text resource files called PII (Program Integrated Information) files. The PII file includes the keys and the isolated text strings. The isolated strings are called PII strings. The programs have the keys and use their corresponding strings (Deitsch 2001; IBM 2004; Dr. International 2003; Green 2005). The translators have to translate the PII strings in each file. They often cannot provide good translations because many PII strings are short and lack context. Therefore a verification test phase compensates for gaps in the translations. This verification test is called a TVT (Translation Verification Test). The testers work with the target software and check the validity of the translated PII strings as they are displayed in the GUI in each context.

The current internationalisation process causes difficulties for the translators and the TVT testers. One of the big problems is that TVT tester cannot find the source location of the externalised text found in a GUI message. This chapter addresses this TVT problem and proposes a solution.

Conventionally the testers have used a 'grep' function of the OS or editor program to find the source location in the PII files when they need to fix a translated string. The grep function requires a long time when checking a large number of files. The TVT testers cannot identify the source location if identical strings appear with different keys. One of the goals of our work is to find the locations of such strings in the PII files effectively and efficiently. To achieve this goal, we developed a utility program to make it easy to find the source key of the PII string displayed in the GUI. Another goal of this research is to help TVT testers refer to both the original (English) and target (Japanese) PII strings as pairs. To achieve this goal, we developed a utility program to produce a comprehensive index file showing all of the information about the PII strings in both languages.

We confirmed the effectiveness of the utilities by actually using them for the Japanese TVTs of the CATIA PII translation. We also discovered many useful ways to use our utilities while performing these TVTs. One of the benefits of using these utilities is to expose the hard-coded strings in the tested program. A hard-coded string is called the "granddaddy" of all TVT errors and is the most difficult source string to find (Kehn 2002). The introduction of our utilities reduced the time to find the string locations in the PII files from 40 hours to one or two hours for the CATIA TVTs.

We explain the background of our research by focusing on the PII translation process and a statistical analysis of the characteristics of PII strings in Section 2. In Section 3, we explain the problems that the TVT testers face in conventional TVT. In Section 4, we explain our approach to solving the problems in TVT. In Section 5, we describe the details of our implementation so that programmers may implement our approach in their own systems. Most readers may prefer to skip Section 5. Finally, Section 6 concludes this chapter.

The following terms are used within IBM. The displayed string information in the GUI is called PII (Program Integrated Information) and the Translation Verification Test is called the TVT. There are also strings that are not separated out into external text files. Such strings remain in the tested program and cannot be translated. We call those strings "hard-coded" strings. IBM uses "TranslationManager for Windows" as a tool for PII translation. We call this tool TM for short. TM manages its data in a proprietary format called an IU (Information Unit). A TSC (Translation Services Center) is an organisation that specialises in translations, especially PII and manuals.

2. Background of the research

This section explains the background of the research. Usually translating a word or a short hrase without context is almost impossible because there are ambiguities in word sense. However, many PII strings that consist of a word or two can be translated in PII files without context. Most program internationalisation engineers do not question why PII string can be translated in PII files without context. One of the reasons this question is ignored is that there are no written documents that explain the difficulties of PII string translation. This section explains why translation and validation are so difficult. First, we explain the conventional translation process of PII in Section 2.1. Then, in Section 2.2, we present a statistical analysis of the PII strings. This analysis explicates the problems in the translation and validation of the PII strings. In Section 2.3, we explain how the PII strings can be translated without context. Finally we discuss the related research in Section 2.4.

2.1 The Flow of PII translation and validation

This section shows the flow of PII translation and validation. Fig. 1 shows the flow of the TVT for PII. The TSC handles the parts of the "Translation Folder" and "Translation Tool". There are three parts in Fig. 1, the top part (Original GUI row), the middle part (Translated GUI row), and the bottom part (GUI for TVT row). The bottom part is our new process (step (8)). This new part will be explained in Section 4. The TVT corresponds to steps (4)–(7). The steps (1)–(7) appearing below describe the process flow of the PII translation focusing on the PII files. There are three strings, ABC, XYZ, and HC, in the program. Two of them, ABC and XYZ, are externalised to PII files. Only the string ABC is in the scope of the Japanese translation and the string XYZ is left as English (the original language). HC is a hard-coded string and cannot be translated.

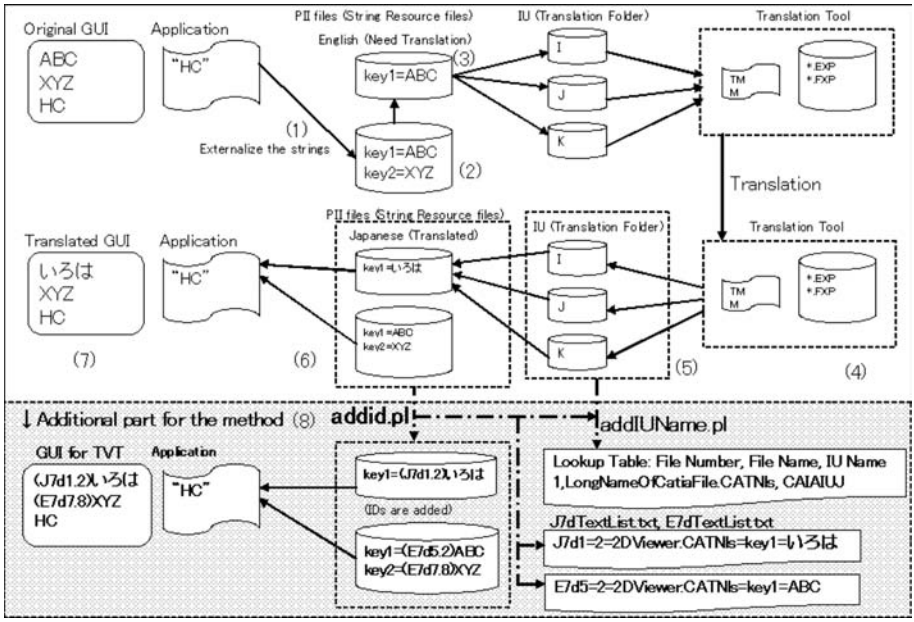

Figure 1. Translation flow of PII

(1) A development laboratory externalises the program's strings into the external text files called PII files.

(2) The PII files consist of keys and their corresponding English strings. The following two lines are examples in the plain text file.
key1=ABC
key2=XYZ

(3) The development laboratory delivers externalised files that require translation to a TSC. The files are grouped into the IU folders.

(4) The TSC translates the PII strings by using TM to import the files from the IU folder. The FXP and EXP files are the internal file formats of TM. The M stands for the memory table of the English and Japanese string pairs.

(5) TM exports all of the IU into plain text files.

(6) The development laboratory receives the IU. The developers copy all of the PII files into their systems.

(7) The TVT is executed on the actual test systems in a laboratory or at a remote site.

(8) Our new process for the ID tags.

If the TVT testers find inappropriate translations, they fix them on the system in step (4), repeat the steps (5) and (6), and confirm the corrected strings in step (7).

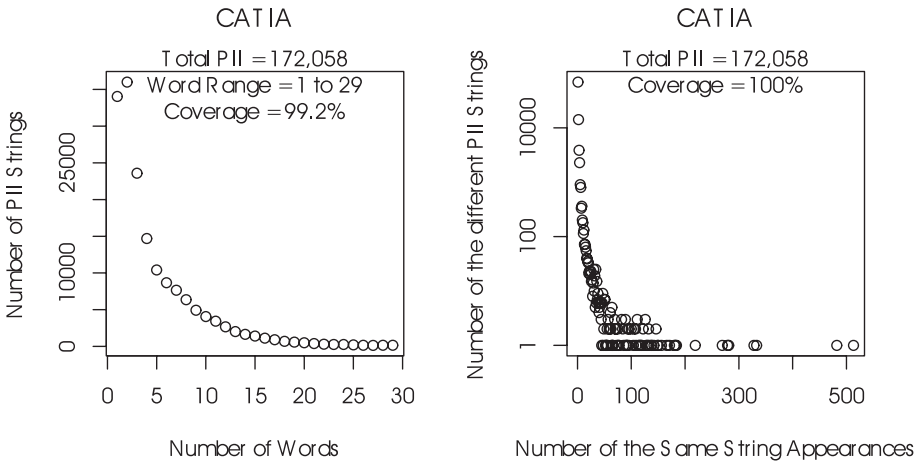

Figure 2. The statistical data for CATIA PII

2.2 Statistical analysis of the PII strings

Usually there are a large number of PII strings in a software system. This section presents statistical analysis of the PII strings. In Section 2.2.1, we show the data for CATIA and Microsoft® Windows®[2] XP SP2. In Section 2.2.2, we extract the statistical characteristics of the PII strings from the data. These characteristics cause the problems that the translators and the TVT testers face.

2.2.1 *The statistical data for CATIA PII and Microsoft PII*

The left side of Fig. 2 shows the distribution of the number of words in each PII string. CATIA Version 5 Release 15 has about 8,500 PII text resource files and about 170,000 PII keys. The horizontal axis is the number of words in each PII string for CATIA. The vertical axis is the number of the PII strings that have that number of words. The figure spans the numbers of strings with less than 30 words. The strings less than 30 words cover 99.2% of the total number of CATIA PII keys. This figure shows that most of the PII strings have only a few words. The average number of words is five words. About 70% of the PII strings have five or fewer words. The figure shows the peak is at two words.

The right side of Fig. 2 shows how many times the same strings appear in the text resource files of the CATIA. The horizontal axis shows the number of string appearances. The vertical logarithmic axis shows the number of different PII strings (type) for the corresponding number of string appearances. If you multiply a point's value on the horizontal axis by the corresponding value on the

2. Microsoft and Windows are registered trademarks of Microsoft Corporation.

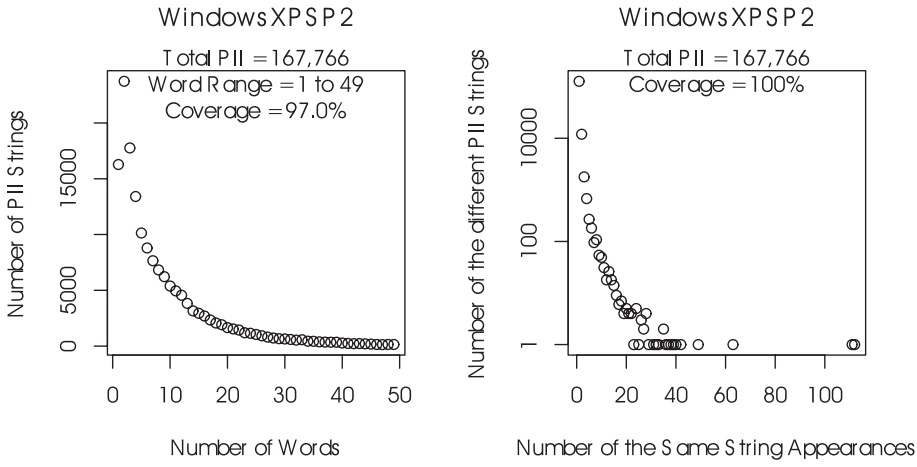

Figure 3. The statistical data for Windows XP (SP2) PII

vertical axis, the result is the total number of the PII strings with that number of appearances.

For example, "Name" appears 333 times. Alternatively, 333 PII keys have the string "Name". The only string that appears 333 times is the string "Name" and therefore that string is plotted at (333, 1). The string "Align Left" appears three times. There are 3,882 unique strings that appear three times and these strings are plotted at (3, 3882) in the right side of Fig. 2. There are 172,058 keys in CATIA PII and there is a point plotted at (1, 69815). This plot means that 69,815 PII keys have only one unique string in the PII files, whereas all of the other strings in the other keys are not unique in the PII files. Therefore those other keys, about 100,000 keys, cannot be identified uniquely when such a string appears in the GUI.

The statistical results for the Microsoft Windows XP (SP2) PII are shown in Fig. 3. We see almost the same characteristics as for CATIA. We used the "Microsoft Glossary" data found on the Internet (Microsoft 2005) and analysed that data. Microsoft calls a collection of "PII strings" a Glossary. We checked 122 applications and OS files for the Microsoft PII. We found that all of the applications have similar characteristics.

2.2.2 *Statistical characteristics of the PII strings and difficulties in the translation of the PII strings*

Our statistical observations give us two facts. First, most PII strings are short. Consequently translators have to translate short phrases without context. This task is so difficult that many PII strings receive inappropriate translations. Second, the same PII strings are often repeated in PII files. If grep is used to find the source location of those repeated PII strings, there are many candidates for the source

locations. The same facts are also true for Japanese PII. Therefore the TVT testers cannot uniquely locate the source text of the inappropriate translations found in the GUI. Our approach addresses these problems.

2.3 Translation without context

Translating PII strings in a PII file is different from translating sentences in a manual. The PII strings in a PII file do not have contexts, while the sentences in a manual do have contexts. Most of the PII strings are not complete sentences and the length of a PII string is very short, as shown in Section 2.2.

There are two helpful strategies the PII strings can be translated without context even though many PII strings are less than 3 words. One strategy is utilizing implicit information coming from neighbouring PII strings in a file. The other strategy is referring to the previous translations from a previous version of the software.

The neighbourhood information and the translation for a previous version can help the translators to imagine the context of the PII strings. Many PII strings have relationships to the neighboring lines. Skilled translators can reconstruct the contexts of the PII strings from the neighborhood information and produce correct translations.

Most PII translations PII files are for files with previous versions. PII string translation is done with reference to the PII strings of the previous version. The previous version gives the translators many hints, even though it also lacks context. Translators often use the same translation when there is a pair of an unchanged original and a previously translated target string.

This translation method works in most cases. However, this method depends on the translators' imagination and occasionally the translators' guesses are incorrect. Therefore the PII translation has problems and needs a translation verification test (TVT) to correct the gaps and errors. This involves looking at the PII strings in context in the GUI. The TVT is an important part of the translation process.

2.4 Related research on tracking PII strings

One of the goals of this research is to find the source location of the PII strings appearing in the GUI. To do this, we developed a utility program. The program adds a short group of ID (identification) characters in front of every PII string. The tested programs in the TVT display the ID as part of the string displayed in the GUI. This tag is called the PII ID tag or ID tag.

The Mock Translator (Muhanna 2003) and the IBM invention disclosure in (IBM 2003) are related research. The Mock Translator allows program developers

to test the program for PII translation. This tool can check whether the program displays the various fonts of supported languages correctly. It does not support any functions for translators.

The invention disclosure (IBM 2003) adds a file name and a directory name for the PII in front of the PII strings similar to the ones used in our approach. Our PII ID tag uses a configurable ID, but the disclosure uses the original names for the ID. The names of the files (including the directory paths) can easily exceed 50 characters, but such long strings cannot be handled properly by typical GUIs, and therefore cannot be used for TVT. For example, the average length of a source file name for CATIA V5 Release 13 was 34 characters. Such file names by themselves cause many problems in a GUI, since the GUI was specifically designed to display short strings in those locations.

The system in the disclosure (IBM 2003) assigns the file names and keys only to certain PII strings, whereas our approach assigns ID systematically to all of the PII, and to both the original and target language files. This systematic approach is an important point of our technique for benefiting from the PII ID tags.

In the linguistic research field, our research is related to the word sense disambiguation (Ide & Veronis 1998). However, there is no research about the word sense disambiguation for PII strings.

3. The main problems that TVT testers face

In this section we describe the main problems that TVT testers face. In Section 3.1, we explain the problems in identifying the source location of a PII string. In Section 3.2, we explain the problems in referring to the original language.

3.1 Problems in identifying the source location

The TVT testers face difficulties in tracking the PII strings. When the TVT testers verify the translated PII strings according to the execution scenario of the software, they cannot identify where the strings are located in the PII files during the TVT. The displayed strings in the GUI have no information about where the strings came from, whether from an external PII file in the original language, from an external PII file in the translated language, or from hard-coded strings in the program itself.

Our research focuses on and solves this problem faced by the TVT testers. Please refer to Fig. 1 again. If a TVT tester finds that the translated 'いろは' (Japanese) is wrong and should be fixed, the tester needs to find the key for the string in a PII file. Then the TVT tester must find the source location of the string that needs to be checked. In the past, the TVT tester has used a grep function of the OS or editor

program to find the source locations of the PII. A TVT tester cannot know whether or not the XYZ string is out of the translation's scope or whether or not the HC string is a hard-coded string. The TVT tester faces difficulties when grep is used. Grep requires a long time to scan the files when there are many PII files. There are about 8,500 files in for CATIA and scanning takes twenty to thirty seconds. Identical strings appear with different keys in various files. Also, grep cannot find a string if the displayed string is actually formed by concatenation in the GUI. Grep is also unable to find or identify the hard-coded strings.

3.2 Problems in referring to the original language

Translation reviewers for a manual can refer to the original and target languages simultaneously. However, a TVT tester for a PII file cannot refer to the both languages simultaneously if the tester is using a single system for testing. When a tester needs to refer to the original language, they need to switch applications from a Japanese version to an English version. If two systems are used for English and Japanese, they need to operate both systems in the same way. The TVT testers need to check hundreds of PII strings in pairs, so the efficiency of referring to the original language greatly affects the total effort for testing. When the translation was from an English source to Japanese, the testers most frequently need to identify the key of a PII string in a Japanese PII file and then search in the English file with that key. Our approach also addresses this problem.

4. Our approach to solving the problems of TVT

In this section we explain our approach towards solving the problems. We briefly explain the ID tag approach and the PII data representation in Section 4.1 and Section 4.2, respectively. Section 5 presents more detailed information for the use of programmers who want to replicate our approach. Then we describe the merits and the applicability of our approach in Sections 4.3 and 4.4, respectively.

4.1 ID tag

Our approach to solving the problems is to create additional PII files that have formatted ID strings in front of the PII strings. These IDs are systematic tags used to track externalised strings. The TVT testers can find these tags in the GUI. If the PII file name and key name are inserted instead of the proposed ID, the average length of the combined names would be about 60 characters for CATIA.

The bottom part of Fig. 1 (GUI for TVT row, Step (8)) shows our additional process to create the PII files for the TVT. We use our Perl program named addid.pl

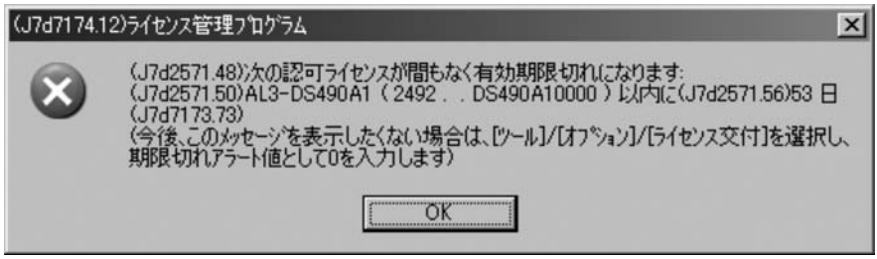

Figure 4. A GUI with the PII ID tags

for both the English files and Japanese files. The program generates additional PII files and a mapping file of the file numbers and the file names. It also generates a comprehensive index file for the language (such as J7dTextList.txt). The mapping file is used by the addIUName.pl program to create a mapping table of the file numbers, the file names, and the IU names. For example, if the original and target PII files include the lines "key1=Link Manager" and "key1=ﾘﾝｸ ﾏﾈｰｼﾞｬｰ" respectively, then the generated files include "key1=(E7d25.22)Link Manager" and "(J7d25.22)ﾘﾝｸ ﾏﾈｰｼﾞｬｰ", respectively. 'E7d' means English PII, Release17, PII ID tag build 'd', and 'J' refers to the Japanese PII. The E7d and J7d are prefixes for the ID tags and are controlled as argument strings for addid.pl. The '25' in this example means the 25th file of the PII files. The '22' means that the key is located in the 22nd line of the 25th file. The ID becomes a part of the PII string, so this approach can work for any programs that have externalised strings.

If a string in the GUI is not associated with an ID tag, then it means that the string was not externalized, but is a hard-coded string. Fig. 4 and Fig. 5 are examples of the GUI with the IDs. We confirmed that the problems mentioned in the previous section were solved by using the IDs. This approach is now used by the IBM TSCs of other countries for the CATIA TVT. We will explain the details of our implementation in Section 5.

4.2 PII data representation

A TVT tester can easily find the source locations of the PII strings by referring to the ID displayed in the GUI. To simplify the ID references for testers, we prepared another utility program to generate a comprehensive index file that lists the IDs, strings, file names, and keys for both languages (See Section 5.5). The comprehensive index file is a single text file. This single text file has all of the PII strings both in English and in Japanese for one software system. An example paragraph in the text file is shown below:

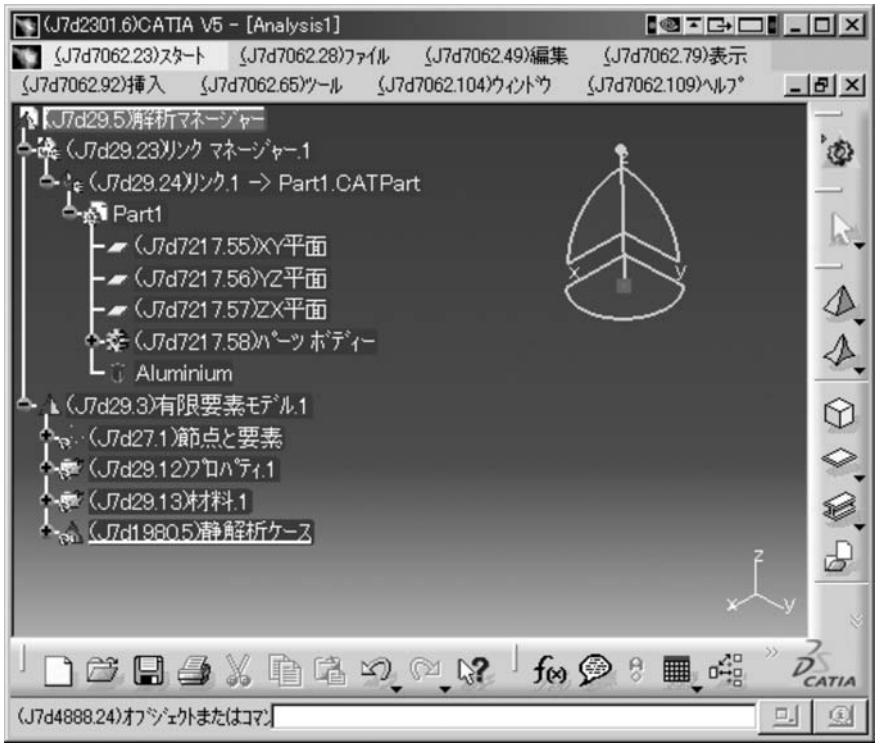

Figure 5. The PII ID tags in CATIA

E7d6088, 36, "Curve Creation", CATStCLA.CATNls, SmartCurves.Title
J7d4891, 36, "曲線を作成"
<Followed by a blank line in the actual text file.>

There is a pair of these lines for each PII key. We can also use the file in various ways for PII maintenance. An ordinary text editor can display the comprehensive index file. The size of this comprehensive index file for CATIA is about 30 Mbytes, but a powerful editor such as K2Editor (Koyabu 2008) can load the file in a few seconds. Extending the editing functions is easy with the macro functions of such an editor. For example, the jump function of the editor can be used to jump to the original PII file from the comprehensive index file (see Section 5.5).

4.3 Merits

Our approach allows a TVT tester to identify the source location of a PII string quickly. In CATIA TVT, the total time for locating the source was reduced to a few hours from 40 hours. Our approach can also separate the problems by using a structural prefix for the ID tags. Our ID approach can identify the hard-coded

strings. English text (using Latin characters) appearing in a Japanese GUI can be identified to determine whether it is a translated string or an untranslated string.

The comprehensive index file is not only useful for looking up the original English strings, but also helps translators and TVT testers to easily view all of the original and translated PII strings in pairs. With this comprehensive index file it is possible to check the consistency of the translations. This comprehensive index file is also useful in the maintenance phase for the PII strings.

No specification is needed for this comprehensive index file because the format of the contents is straightforward. There are many ways to use the comprehensive index file.

The following are additional merits of our systematic approach:

- The ID prefix encourages a user to use an appropriate architecture.
- TVT testers do not need to have knowledge of the tested program itself.
- The approach can work for any programs with externalised strings.
- The ID has important merits beyond replacing 'grep' searching. It can uniquely identify the source location of a string appearing in the GUI and show whether or not the string is concatenated.
- An ordinary text editor can view the comprehensive index file. A database system such as DB2®[3] is not needed to view the index data.
- The comprehensive index file allows us to find specific language pairs by using the editor's search function with regular expressions.

4.4 Applicability to other systems

Our approach can be applied to any development systems that externalize the PII strings. It can support XML format as well, because it only modifies the PII strings instead of the PII file format.

We applied our approach to a Java application and it worked well. In the Java application, we confirmed that we could easily switch PII files between the PII with IDs and the PII without IDs by utilizing the Java '-Duser' start option. The naming convention of the properties file can be exploited for this purpose. We discuss the details of the Java application in Section 5.6.

5. Implementation details for programmers

This section is for programmers who wish to implement our approach or to investigate its functions. General readers may prefer to skip this section. We use CATIA

3. IBM and DB2 are registered trademarks of International Business Machines Corporation.

to explain the detailed implementation of the ID tags, except for Section 5.6, which is about a Java application (which also shows that our approach is not limited to CATIA). Other systems can apply our approach. In Section 5.1, we show the PII text resource files supported by our tools. In Sections 5.2 and 5.3, we show the PII ID tag format and the prefix architecture of an ID tag, respectively. Then we explain the ID tag generation program in Section 5.4. The comprehensive index file for TVT testers is explained in detail in Section 5.5. Finally, in Section 5.6, we explain how our approach is used with Java.

5.1 Supported PII text resource files

This section explains the PII text resource files that we support in CATIA. Each PII file includes many lines in the following format:

key=string

The left side of the equal sign is called a PII key. The right side of the equal sign is called a PII string. The set of a key, equal sign, and string is called a PII entry. The Java properties file for internationalisation is an example of this standard format. CATIA uses this standard format, but it has some special characteristics. It requires quotation marks around strings and it uses a semicolon ";" to separate the lines. The file has file type extension CATNls and is called a Cat NLS file. The following are examples format of some lines of the original language.

```
...
key1="Link Manager1";
key2="Link Manager2";
key3="Link Manager3";
...
```

Usually the key string is quite long. The following example is slightly longer than the average length:

NoFilteringOnVisuMode.Title="Filtering Management";

When we translate an original language file into a target language file, the PII strings, the right sides of the lines, are translated into the target language. If we translate the examples of these Link-Manager lines, the translated PII file should include the following lines.

```
...
key1="リンクマネージャー1";
key2="リンクマネージャー2";
key3="リンクマネージャー3";
...
```

The original English CATNls files are located in the following directory.

B15\intel_a\resource\msgcatalog

The B15 folder is located in the installation directory of the program. The translated Japanese PII files would then be located in the following directory.

B15\intel_a\resource\msgcatalog\Japanese

The PII files with the ID tags will replace the files in these two directories. About eight thousand PII files (CATNls files) for each language have to be replaced. It is necessary to replace the original English files so as to identify the hard-coded strings. If we replace both the original and the target PII files and the application then displays in the GUI any strings without ID tags, it means that those strings did not come from any PII files, and therefore shows that the string is hard-coded. Such strings reside in the program code, not in the PII files.

After the replacement of the PII files, the example keys, key1, key2, and key3, now include ID tags, as shown below.

English:

key1="(E7d245.7)Link Manager1";
key2="(E7d245.8)Link Manager2";
key3="(E7d245.9)Link Manager3";

Japanese:

key1="(J7d240.7)リンクマネージャー1";
key2="(J7d240.8)リンクマネージャー2";
key3="(J7d240.9)リンクマネージャー3";

We will explain the prefix structure in Section 5.3. In the above examples of tags in parentheses, the "E" means English, "7" means Release 17, "d" means the fourth build of the ID tag, "245" means file number 245 in the English PII files, "240" means the file number 240 in the Japanese PII files, and the "7", "8", and "9" after the period signify the line numbers in the PII files. The file number of Japanese is smaller than that of English. This is because large software may have a scope of translation. Some part of large software may remain in original English language. As shown in Fig. 1, the tagged files come from the system PII files, not from the files from Translation Services Center. Some English files may not be passed to Translation Services Center if those are out of a translation scope.

5.2 PII ID tag format

Fig. 5 shows a GUI display from CATIA showing the PII with ID tags. The ID tag (J7d7062.23) appears in front of the string "スタート". The "J7d" is a prefix with an

important role for the ID tag. We will explain the architecture of this prefix more precisely in this section. The prefix J7d signifies Japanese, Release 17, d build for this ID tag. Here is the complete format of an ID tag:

(PrefixFilenumber.Linenumber)

Here are the semantics of the elements:

"(": Start of the ID tag
")": End of the ID tag
"Prefix": Prefix
"Filenumber": The file number of the PII file that includes the string. The file number is assigned within each language directory.
".": Separator between the file number and the line number

The parentheses not only separate the ID tag from the PII strings, but also make it possible to recognise concatenated PII strings in the GUI. We will explain the "Prefix" more fully in the next section. To keep the ID tags short, we did not insert any character to separate the "Prefix" from "Filenumber". We used decimal numbers for the file numbers (Filenumber) and the line numbers (Linenumber). The average length of the PII file names for CATIA is 28 characters and the average key length for CATIA is 34 characters. The GUI elements often do not have space for very long names for the PII strings. The length of an ID tag needs to be short to limit the space that it occupies in the GUI. We considered using hexadecimal numbers to shorten the length of the ID tags. The hex numbers would reduce the length when the file numbers are large. However, this called for a separator between the prefix and the file number and actually increased the average length of the ID tag. The readability of the ID tag was also greatly decreased as hexadecimal. Therefore, we used ordinary decimal numbers.

We prepared the comprehensive index files to help the TVT testers find specific PII keys in the PII files with or without ID tags. The format of the ID tag allows a TVT tester to search for the location of the ID in the comprehensive index file. The comprehensive index file uses a plain text format, and it is a list of all of the PII for both the original and target languages. We will describe the comprehensive index file more fully in Section 5.5.

5.3 Prefix architecture of ID Tags

A TVT tester can assign any ASCII strings to the prefixes when an ID tag is used in TVT. However, the TVT tester must conform to the prefix architecture. The prefix architecture is an important feature of the ID tag. Without it, the value of the ID tags would be reduced.

The main function of the prefix is grouping. The group is not only language group, but also levels of strings including the version/build numbers etc. The grouping function is used to identify the language sources of the PII files. The function identifies whether or not the string was located in the original language (English) PII or in the target language (Japanese) PII. The displayed strings do not always identify the language source unless the ID tag is used, because the target language may include the original language without a change. For example, if an English identifier such as "E7d240.7" is included in a prefix and is shown in a Japanese GUI, this means that no Japanese PII key was found by the programs. The program uses an English PII string instead of a Japanese PII string if it cannot find the Japanese PII key. One possible cause of such errors is that development laboratory might have failed to send the corresponding English PII file to the Translation Services Center.

The grouping function may include additional information identifying the levels of the PII files such as the version or release numbers of the PII files and ID tag build numbers for the files. We can have many levels of PII files within each version or release, which calls for tracking the ID tag build numbers.

The prefixes must not include the characters used to separate the elements in the ID tag format, such as parentheses or the period. The last letter of the prefix must not be a numeric character. If the last letter was a number, then the users could not see where the file numbers begin.

The current prefix format for CATIA is as follows:

PII language (One character): e.g., E is used for English, J is used for Japanese
Release/Version (One numeric character): e.g., 5 means release 15. The use of ID tags began after release 10, so we ignored the "1" to shorten the ID tags.
PII ID tag build number (One lowercase alphabetic character): e.g., "a" means the first build.

The addid.pl program included a function to replace the separator "." with another character according to a table. The table is plain text and has pairs of text strings. Each pair maps a key with its file name to a character such as ":", "_", ".". The mapping table file is prepared separately before the addid.pl is executed. If there is no mapping table file, the default separator "." is used. For example, if a PII string for a key is being updated from the previous version, then the key has mapping information for a colon. If not, it uses the default mapping to a period. Therefore, if the ID tag is "(J7d240:7)", it shows that the string is being updated from the previous version. Greater attention needs to be paid to such translations. If a key was created in the latest version, then the key has mapping information to the underscore and the ID tag becomes "(J7d240_7)". Then the TVT tester must check whether or not the translation is appropriate.

5.4 ID tag generation program

We developed a Perl program to create a separate set of files with ID tags for each set of PII files. Perl was selected in 2003 because of the maturity of its regular expressions and its Unicode support. As of 2007, other programming languages could be used to prepare such a program. We do not change the original PII files, but only create additional files to replace the original PII files. The additional files have no effect on the tested program and its PII files. These files are only used for the TVT tests. The addid.pl supports the CATNls files and the Java properties files. The addid.pl program handles one directory of files at a time. A typical directory is a set of English PII files or Japanese PII files. All of the CATIA PII files for a language are located in one directory. The addid.pl program does not support the nested directory structure of PII files, so the program must be run separately for each subdirectory. Care is needed if the PII files have a nested directory structure.

The addid.pl program has the following parameters. The default values are shown in the parentheses.

-l Directory name of the PII files (Japanese)
-f A filter for the file names that are processed (*)
-p Prefix (Null)
-e End of the ID tag (no)

If the -e option (Position of the ID tag) is set to "yes", then the ID tag is put at the end of each PII string instead of at the beginning (though this option is rarely used).

If we use J7d as a prefix, we will copy the Japanese directory of CATIA PII files into the new J7d directory for the generated files. The command "addid.pl -l J7d -f *.CATNls -p J7d" generates the following output.

J7dWithID: An output directory of files with ID tags

This directory contains all of the copied files with the PII strings modified by adding the ID tags.

J7dNumList.txt: A map file from the ID file numbers to the PII file names.

This contains a CSV text mapping of the ID file numbers to the PII file names.

J7dTextList.txt: A text file that has all of the PII

This text file contains all of the sets of PII (pairs of a key and a string) for all of the files in the J7d directory. We call this comprehensive index file the "Basic Form" in Section 5.5.

We prepare a J7dIU directory that contains the IU directories copied and separated for TM (TranslationManager) use. If addIUName.pl is executed with

J7d as an argument, it outputs the following two files. This J7dIU directory is a Translation-Manager-specific requirement if you are using that program.

J7dNumListWithFolder.txt: J7dNumList.txt with IU directory name

The IU directory name for each PII file is added to the PII file name. The following is a part of the content in the J7dNumListWithFolder.txt file. The file number 5 is mapped to the ACLEditor.CATNls file, which is included in the PLM61AAP004 TM folder.

```
...
5,ACLEditor.CATNls,PLM61AAP004
6,ACOClaCreationWbenchHeader.CATNls,PLM61AAP001
7,AECBAnalysisMsgCat.CATNls,PLM61AAP006
8,AECBPropertiesMsgCat.CATNls,PLM61AAP006
...
```

J7dNumListNotInFolder.txt: A list of files that do not exist in the IUs

This file is a list of files that do not exist in the IUs. These files exist in the actual program. If a PII file was translated in the previous version or release and the English PII file was not sent to Translation Services Center, then that PII file will not exist in the IU directory of the TM.

We use translation assistance tools such as Trados®[4] and TranslationManager to maintain the translation memories. A translation memory is a set of segmented original and target language pairs. TVT testers cannot update the PII files directly. They must update the assistance tool they are using. This operation is not directly related to the ID tag, but how to integrate the ID tag functions into the translation assistance tools is a future research topic.

5.5 Comprehensive index file

This section explains the comprehensive index file. If a TVT tester finds an inappropriate translation, the corresponding English PII must be located. The comprehensive index file helps to refer to the corresponding English string by using the ID tag. Before using the ID tag, the tester needed to look for the inappropriate Japanese by using the grep function. If the search string appears only once in the PII files, then the key would be identified easily and the tester could refer to the corresponding English file and key. If more than one matching Japanese string appeared in the PII files, the tester had to spend time specifying the PII key location by editing the Japanese PII files. It could take a long time to find the corresponding English. The comprehensive index file and ID tags solved this problem. Ten min-

4. TRADOS is the registered trademark of SDL.

utes of work searching for a string was reduced to ten seconds of computerised searching.

The comprehensive index file is a text file that contains all of the PII data. The comprehensive index file can be viewed with an ordinary text editor. There are four types of text data, "Basic Form", "One-Line Form", "Reference Form", and "Reference Form with Tag Jump". The following are actual examples from CATIA.

[**Basic Form**]
E7d1=2=2DViewer.CATNls=Visibility.Title="Visibility"

[**One-Line Form**]
E7d1=2=2DViewer.CATNls=Visibility.Title=
"Visibility"="表示/非表示"=J7d1=2

[**Reference From**]
E7d1, 2,"Visibility", 2DViewer.CATNls, Visibility.Title
J7d1, 2, "表示/非表示"
<One blank line>

[**Reference Form with Jump**]
E7d1, 2, "Visibility"
J7d1, 2, "表示/非表示"
.\E7d\2DViewer.CATNls 2:Visibility.Title
.\J7d\2DViewer.CATNls 2:Visibility.Title
<One blank line>

We call a comprehensive index file such as J7dTextList.txt the "Basic Form". The Basic Form file is created by addid.pl when the PII files with ID tags are generated (see Fig. 1, (8)). There is one blank line after each entry in the Reference Form and in the Reference Form with Jump. There are no blank lines in the Basic Form or in the One-Line Form. The above types of lines for each form are repeated for all of the PII data. The Reference Form is mainly used in the TVT and in the PII maintenance phase. The One-Line Form is convenient when searching for language pairs and can be handled with a spreadsheet such as Excel. Excel alone cannot be used for CATIA because of the large amount of PII data, since the number of lines of data is over Excel's limit of 64k lines. If the comprehensive index file is used for Java properties files, there are no quotation marks in the forms. The comprehensive index files after the Basic Form are created from pairs of Basic Form files such as E7dTextList.txt and J7dTextList.txt. The Reference Form with Jump adds two lines to the Reference Form. Many text editors have a function to open another file by hitting a function key (e.g., F10) depending on the location of the editor's cursor. The additional two lines help to open the PII files where the PII key exists. Use of the Reference Form with Jump can easily check whether or not the actual PII file has additional information for the PII keys, such as comments.

Using a Reference Form file, a TVT tester can easily find the language pair by searching for the ID tag. The people who maintain the PII files can also use the comprehensive index file to investigate the PII language pairs when there is a customer comment or complaint. Users of the comprehensive index file need to refer not only to a specific PII string, but also to many PII strings around the specific PII strings for the paired languages. Observing the strings around the specific PII string helps to clarify the meaning of a specific PII string. Viewing the pairs around a specific PII string is often most important feature of this comprehensive index file. Observing only a specific pair of PII strings is often useless. Translators cannot translate a language without a certain amount of context. The pairs around a specific PII give a certain amount of context though they are often still insufficient. Another important point is that very quick search is possible within a text editor, but the grep approach is not as fast.

5.6 Java implementation for switching PII files

The same approach used for CATIA can be applied to Java applications. The only difference is that the file name portion of comprehensive index file includes the directory names because the PII properties files of Java are usually located in nested directories.

The ID tag technique is independent of the language. However, if there is a language selection mechanism in a program, there is no need to replace the PII files with the PII files with ID tag. CATIA did have such a mechanism, but it is too specific to CATIA. Therefore we will explain the Java approach for switching between the PII files with and without ID tags.

Java applications can specify the language, country, and variant using the -Duser option when the program is started. There is no need to replace PII files with PII files with ID tags. Major languages including Japanese, German, and French usually do not use the country or the variant. We used the naming convention of the Java properties file for TVT.

For example, if a Java application program supports, Japanese (ja), Japan (JP), and Kansai (kansai), then the supported PII files would be the following.

(A) filename.properties
(B) filename_ja.properties
(C) filename_ja_JP.properties
(D) filename_ja_JP_kansai.properties

A Java application program searches in the properties files to substitute strings for PII keys in the order of (D), (C), (B), and (A). The application program displays in its GUI the first matching PII string found in that search.

Assume that the original language is English and the target language is Japanese. We can define a virtual country 00 and a virtual variant TVT. The virtual country uses English with ID tags. The virtual variant uses Japanese with ID tags. If the abcd.properties file is an English PII file and the abcd_ja.properties file is a Japanese PII file, the full set of properties files will be:

(A) abcd.properties
(B) abcd_ja.properties
(C) abcd_ja_00.properties
(D) abcd_ja_00_TVT.properties

The file (A) is an English PII file. File (B) is a Japanese PII file. File (C) is an English PII with ID tags. File (D) is a Japanese PII with ID tags. Following are examples of a key1 line in each file.

(A) key1=Link Manager1
(B) key1=ﾘﾝｸﾏﾈｰｼﾞｬｰ1
(C) (E7d28.22)Link Manager1
(D) (J7d24.22)ﾘﾝｸﾏﾈｰｼﾞｬｰ1

If this application starts without setting -Duser in a Japanese OS environment, it uses the file (B) Japanese PII strings. If the application cannot find a key in File (B), it displays the data from the File (A) English PII strings. If this application starts with the following -Duser options,

-Duser.language=ja
-Duser.country=00
-Duser.variant=TVT

then the application program selects the virtual country 00 and the variant TVT. It displays the File (D) Japanese PII strings with ID tags. If it cannot find a key in File (D), it displays the data in File (C), the English PII strings with ID tags.

In this way, we do not need to replace the PII files to shift between the pure PII files and the PII files with ID tags for testing. This Java approach is not the standard approach to switch the PII files for testing. Programming languages need to offer switching functions that help the TVT testers. We hope that this article helps computer scientists understand what functions of localisation are required for programming languages and programming environments.

6. Conclusion

There are two major problems in PII translation. One is the PII translation problem itself, and the other one is the verification problem. This chapter has focused on the verification problem.

TVT testers cannot identify the source locations of the PII strings that are shown in a GUI. We systematically inserted a distinctive and compact ID in front of every PII string for all of the PII files, without internal knowledge about the tested target programs. By using the modified PII files with the unmodified executable programs, TVT testers without deep knowledge of the program were able to quickly and easily find the exact sources of the PII strings. One of the useful and important features of the ID includes recognizing the hard-coded strings in the tested program.

We also developed a comprehensive index file to help the TVT testers refer to all of the PII information in one file. This file is not only used to identify the source location of the PII strings but also used to refer to the original and translated strings in pairs. In the problem analysis, we showed statistical information about the PII strings. This information has not been clearly recognised by the program internationalisation communities.

References

Deitsch, A. & D. Czarnecki (2001) *Java Internationalisation*. O'Reilly.

Dr. International (2003) *Developing International Software*. 2nd ed. Microsoft Press.

Green, D. (2005) *Trail: Internationalisation*. http://java.sun.com/docs/books/tutorial/i18n/index.html

IBM (2003) *Debugging method for message translation of software product*. http://www.priorartdatabase.com/IPCOM/000018811/

IBM (2004) *Designing Internationalised Products*. National Language Design Guide, vol. 1, 5th ed.

Ide, N. and J. Veronis (1998) Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. In *Computational Linguistics* 24(1), 1-40.

Koyabu, K. (2008) *K2Software's page*. http://k2top.jpn.org/

Kehn, D. (2002) *Testing your internationalised Eclipse plug-in*. http://www-106.ibm.com/developerworks/opensource/library/os-i18n2/

Microsoft (2005) *Microsoft Glossary ftp site*. ftp://ftp.microsoft.com/developr/msdn/newup/Glossary/

Muhanna, E. (2003) *Mock Translator: A tool to generate national language assets*. IBM Technical Report: TR-74.193-46-d, CASCON 2003 Software Testing Workshop.

CHAPTER 12

Linguistic resources and localisation

Reinhard Schäler
Localisation Research Centre (LRC), University of Limerick

Traditional mainstream localisation processes, tools and technologies supporting a business approach whose principal purpose it is to achieve a short-term return on a minimum investment have reached their limits. In order to respond effectively to today's localisation challenges disruptive approaches are needed and an overhaul of current practices is required. This contribution makes a solid case for localisation as a long-term investment that is backed up by case studies and advocates the innovative use of language technologies and language resources.

1. Introduction

What is the next "big thing" in localisation? How can an ever-increasing amount of digital content be made available to customers all over the world simultaneously in an ever-increasing number of languages? Many experts believe that traditional mainstream localisation processes, tools and technologies have already reached their limits and that a radical overhaul of how localisation is approached today is needed to respond adequately to the localisation challenges of tomorrow.

Tools that interact only with their custom linguistic resources; technologies that only address specific aspects of the localisation process; lack of interoperability between language resources, tools and technologies; process automation that locks clients into their vendors' framework; standards that focus more on the needs of developers than on those of the users, standards that are not undergoing stringent peer reviews, that are rarely implemented, "closed" and do not have a proper policy on ownership (IPR); desktop-based, standalone translation and quality assurance (QA) tools; cascading supply chains – all of this will soon have to become just a distant memory for those who want to survive and grow their localisation business into the future.

It is the requirement to deliver multilingual and cross-cultural digital content to its clients in a timely and cost effective fashion that makes dramatic changes in the localisation industry necessary. Language technologies, combined with process automation and based on solid standards, offer solutions to the industry's constant

demand for higher throughput at lower cost. Yet, to-date, no concerted efforts have been undertaken to create a robust infrastructure for the localisation industry comprising language data and tools coupled to process automation and based on widely accepted standards.

When did you try the last time to break your existing localisation process? What was the last "wow!" experience you had as a localiser? Different professionals would answer these questions in slightly different ways, but most would agree that it has been a long time since mainstream localisation "wisdom" was seriously challenged.

In this chapter, we will attempt to do just that.

1.1 Important definitions

Two terms usually associated with localisation are *internationalisation* and *globalisation*. We define *internationalisation* as the process of designing (or modifying) software so as to isolate the linguistically and culturally dependent parts of an application, as well as the development of a system that allows linguistic and cultural adaptation supporting users working in different languages and cultures. By contrast, we define *globalisation* in this context as a business strategy (not so much as an activity) addressing the issues associated with taking a product to the global market which also includes world-wide marketing, sales and support.

Probably the most difficult distinction to make, however, is that between localisation and *translation*. Not just localisers, translators as well adapt products (text) linguistically and culturally so that they can be understood in different *locales*. However, translation does not necessarily deal with digital material whereas localisation is *always* happening in the digital world. This has a number of implications in a number of different areas.

Firstly, the material localisers deal with is multimodal, i.e., the files they localise can contain text, graphics, audio and video applied to a large variety of services or products, from websites to desktop applications, video games and courseware.

Secondly, the digital nature of the material determines the process (analysis, pre-processing, translation automation, testing, engineering), the tools and technologies, the release and distribution, as well as a number of very specific challenges which are all quite different from those encountered in traditional translation. These include:

- File formats (huge variety, ever growing number)
- Encoding, fonts, rendering (dependent on standards; sometimes difficult to implement; not always available)
- Input methods (keyboard, mouse, scanner, speech)

– User interface space restrictions (size, structure, display quality, memory)
– Context (or lack thereof) and visual translation environment

Language experts working in localisation as *translators* cannot restrict themselves to translation *per se*; in fact, translation very often takes up only a fraction of their working day, the rest they spend on file management, translation memory and terminology database maintenance, coordination between large groups of translators, as well as linguistic testing of the target material.

2. Localisation

The rationale behind current localisation efforts has remained the same since the early 1980s when the first localisation projects were undertaken: it is the drive to increase sales. When the sales and marketing experts of large US-based IT developers in the 1980s looked for opportunities to grow sales outside of their native US-market (which was by then considered to be largely saturated), they targeted Europe as their next major market. This was the moment that the localisation industry was born. Although people living in large European economies, such as Germany, France, Italy and Spain, had a need for and the means to buy expensive computing hardware and software, they could not use them in English. Word processors, spreadsheets and, soon after, presentation software had to be translated and adapted for these new users.

2.1 Increase ROI

The formula to achieve an increase in short-term return on investment (*ROI)* was simple: adapt an already developed product superficially to the requirements of foreign markets, with a minimum effort, and then sell it into these new markets for a similar price as the original product.

Multigual digital publishers soon realised that they could increase their profit margins even more if they implemented small changes in the way their original product was developed, so that the cost of adaptation (or *localisation*) could be reduced even further. Products were made more easily localisable, functionality and content were separated and globally acceptable content was used wherever possible, based on the lowest common denominator (LCD) of cultural acceptability. The main aim was to offer a world-ready out-of-the-box product. Developers started to use recognisable colours, symbols, sound and signs, because the lower the adaptation effort – the higher the potential earnings. The overall aim was (and still is) to reduce localisation to translation.

While localisation was made easier and became less expensive, translation became one of the biggest single cost in localisation and, therefore, the first and most important target for automation efforts. *Leveraging* was the order of the day, localisers were asked to re-use as much as possible from previous translations, to process as much as possible and to translate as little as possible. Changes to the source text were limited to an absolute minimum to eliminate a domino effect: just one change in the source would trigger off adaptation efforts in all of the different language versions.

2.2 Key phases

We identify three key phases in localisation since it emerged as an industry in the mid 1980s.

Table 1. Key phases of localisation

Phase	Period	Characteristic
I	1985–1995	Initial unstructured
II	1995–2005	Structured
III	2005–	Virtual

The *first* phase was characterised by *ad hoc* solutions to what were then perceived to be *ad hoc* problems. Even senior people in the industry were proud to say that the biggest attraction of localisation for them was the constant change, that no two projects were ever the same. It was not unusual then to hear managers say, for example, that checking what aspects of a previously translated version of a project could be reused would be more expensive than paying the translators to simply get on with the translations of the new version.

The *second* phase of localisation saw a certain degree of maturity emerge. Localisation projects no longer had to rely on *heroes* living on take-away pizzas and on engineers that fixed problems as they occurred. Organisations such as the Localisation Industry Standards Association (LISA) and the Localisation Research Centre (LRC) made case studies and best–practice recommendations available, while – for the first time – two distinct types of third-party tools and technologies revolutionarised the way localisation was done: user interface (UI) localisation platforms, such as Catalyst and Passolo, and translation memory systems, such as TRADOS and the IBM Translation Manager.

The *third* phase of localisation is still evolving and will mainly be characterised by a move from desktop-based to web-based localisation environments responding to the needs of distributed localisation teams working on the management, the engineering, testing and translation of digital material into close to one

hundred languages simultaneously and on demand. First evidence of this move is the astonishing growth rate of companies offering corresponding services and technologies, among them SDL, Lionbridge, and across. In this scenario, easy access to relevant linguistic resources, covering language data, tools and standards, is paramount for both developers and users.

2.3 Localisation success

The success of this strategy, to use localisation in order to increase revenues and profits, is unquestionable. For example, for many Fortune 500 firms, non-US revenue, or *xenorevenue,* accounts for 20 to more than 50% of their global income, according to Common Sense Advisory (DePalma & Beninatto 2002). This fact alone makes it easy to see the value in catering for buyers in global markets with localised products and services in their language. Compared to the benefits of gaining market share and customer loyalty localisation expenditures are minuscule, 2.5% and lower of non-US revenue.

One of the planet's largest companies, Microsoft, now generates more than 60% of its revenues from international operations, more than US$ 5 billion per year. It manages more than 1,000 localisation projects (product/language) per year. In Ireland alone and in just one year (2001), it created revenues of US$ 1.9 billion (Balmer 2002).

Although it is difficult to provide an exact estimate of the volume of the localisation market, there have been some efforts to capture market volumes, of which the best known are those prepared by Common Sense Advisory and the European Union of Associations of Translation Companies.

In 2005 Common Sense Advisory estimated that the market for outsourced language services was US$ 8.8 billion worldwide and growing at 7.5% per year to over US$ 9 billion for 2006, and to US$ 12.5 billion by 2010. These calculations were based on the aggregate revenues of the several thousand companies active in the business, the many freelancers, and on an approximation of the revenue generated by international and ethnic marketing agencies, boutiques, system integrators, consultants, printers, and other service providers who facilitate translation and localisation. In 2006 demand grew at 15 to 20% per year, driven by national regulations, website and product localisation and consumer need for more information in their own language. At the same time, industry growth has been holding steady at below 10% (DePalma & Beninatto 2006).

According to the European Union of Associations of Translation Companies (EUATC), the total market in 2005 was worth US$ 11.7 billion worldwide, with Europe being the world leader in the translation industry. The EUATC believes that the market will grow to US$ 12.5 billion in 2006 and be worth US$ 15.8 billion in 2010 (Boucau 2006).

2.4 Vectors of growth

The demand for localisation is growing; customers want services and products in their language. According to the findings of a new market study undertaken by Wordbank (2005) into the impact of language on the consumer's purchasing behaviour, more than eight out of ten consumers expect global companies to sell to them in their own language and seven out of ten will not buy a product if they cannot understand the packaging. Other key findings of the study, entitled "Are you talkin' to me?" include:

- 98% of those with no knowledge of English want to be communicated with in their own language as do three quarters of those who speak fluent English as a foreign language.
- When faced with a choice of buying two similar products, 73% of consumers are more likely to purchase the one that is supported by product information in their own language.
- Consumers are negatively influenced by poor translation – 61% are reluctant to purchase a product if the information has been badly translated into their own language.
- More than seven out of ten (71%) respondents are more likely to purchase the same brand again if the after-sales care is in their mother tongue.
- The top three products and services that consumers need to be communicated with in their own language are: banking and financial services (86%), pharmaceutical and beauty products (78%), and consumer electronics (73%).
- These are closely followed by business equipment (71%), home entertainment (71%), and computer hardware and software (71%).
- Of the 39 languages covered in the survey, product communication in their own language is most important to Portuguese speakers. They are closely followed by Spanish and German speakers.
- The older the respondent, the more they want to be communicated with in their own language.

The growing demand by consumers for localised products translates into what we have identified as the nine vectors of scalability and growth that we have associated with the three phases of the industry's development as defined earlier.

Moving into phase 3, localisers face demands for a dramatically improved throughput without an increase in overall cost or time from both internal and external customers. While release cycles for a digital product could have been as high as 18 months in the early days, the move today is toward a continuous release of digital content, including updates and upgrades of applications distributed over the Internet. The pressure to simultaneously ship ("simship") all language versions of this content, i.e. to make it available for download at the same time as the

Table 2. Nine vectors of scalability and growth associated with the Three Phases of localisation industry development

No.	Vector	Phase 1	Phase 2	Phase 3
1	Geography and languages	Europe	Asia	Global
2	Standards	Trial &Error	Proprietary	Open
3	Content	Manuals/UI	General technical	Any content
4	Rationale	Return on Investment	Investment	Rights-based
5	Medium of delivery	Documents / Boxed products	CD-ROM	Online / Pure internet-based
6	Culture	Symbols	Rights	Values
7	Delta	6-9 months	"SimShip" (within quarter)	True SimShip
8	Cost per language	High	Medium	Low
9	Release cycles (months)	9–18	3–6	Continuous

original version, is tremendous: *deltas*, the time between the release of the original and a localised version of the same digital content, are moving towards zero. The number of languages in which digital content is being made available was initially restricted to just French, Italian, German and Spanish, the so-called FIGS languages, but has since grown dramatically.

On 29 January 2007, Microsoft's Chief Executive Officer, Steve Ballmer, announced at the Windows Vista and the Microsoft 2007 Office System Worldwide Availability Celebration at The Windows Vista Theatre, Times Square, New York, N.Y., that Microsoft VISTA will be available in over 70 countries, starting out in 19 languages and becoming available in over 99 languages by the end of this calendar year. He said, "Afterward you can go think whether you can name 99 individual languages, but that's sort of the extent and the cover and the reach that we'll have with the Windows Vista product" (Ballmer 2007).

Of course, it would never be possible to tackle this extraordinarily high number of language versions using the same processes and paying the same price companies were prepared to pay in the old FIGS-only days. The cost per word processed in a localisation environment had to be dramatically reduced. One way of achieving this cost reduction is described in a report by Forrester Consulting for SDL in 2007, entitled "The Total Economic Impact™ of SDL Global Information Management" (Forrester 2007). Based on this report, SDL estimates the cost of "missed opportunity" due to poor localisation by global business at $4.7 bn and recommends a highly automated and well managed Global Information Management (GIM) approach to localisation (SDL 2008).

Yet, there is a believe by industry experts that even today only 90% of what could be localised, *can* be localised given the still relatively high cost and high

level of dependency on human localisers – a dependency that clearly has to be reduced and be substituted by a higher degree of automation using an appropriate linguistic infrastructure for localisation.

2.5 Case study

The following case study will illustrate how large digital publishers have already begun to use web-based, automated and open standard-based environments to achieve a localisation throughput that would have been unachievable using traditional desktop-based localisation environments.

Tony Jewtushenko, then Tools Manager with Oracle's Worldwide Translation Group in Ireland, presented the following example of the use of language resources in localisation at the LRC's 2003 Annual Localisation Conference.

The project constraints Oracle was dealing with were as follows:

- Projects included four million words in software strings
- These strings were stored in 13,000 localisable files
- The company was aiming for the simultaneous release of their products in 30 languages
- Projects were handles by a localisation group in Dublin collaborating with a 5,000 people world-wide distributed development team

The objectives of the development project were to achieve a 24/7 around the clock and 100% automated process with no exceptions. Translations were to happen in parallel with the development of the original product and immediately following code check-in. The company aimed at "translation on demand" bringing an end to the "big project" model.

The solution Oracle's tools development team came up with was "the translation factory". It was capable of:

- Handling 100,000 language check-ins per month
- Achieving a two million files throughput per month with an average process time of 45 seconds per file and a 98% word leverage rate
- Allowing for the simship of 30 language versions at the same time as the release of the original version

The company achieved a positive return on their investment into the factory within one year and reduced the number of release engineers from twenty to just two resulting in US $20 million saving per year.

It is important to keep in mind that Oracle is one of the world's leading digital publishers and runs one of the most sophisticated internationalisation and localisation operations in the world. Oracle is also centrally involved in the development of two key standards under the umbrella of OASIS, the XML-based Localisation Exchange Format (XLIFF) and the Translation Web Services Group (TWS), which,

combined, have the potential to fundamentally change not just the way localisation is done by Oracle, but by every digital publisher bringing its contents to the global market.

However, it seems to be more than sensible to transfer the lessons learned by Oracle and adapt their large company approach to a point where it can be used to tackle the localisation challenges faced by small and medium sised enterprises (SMEs) and, ultimately, by consumers.

This can only be achieved with sophisticated process automation supported by standards and integrated process and language technologies.

3. Language technologies and automation

Automation requires process environments that have not always been perceived to be possible in localisation where *change is the only constant*, as Teddy Bengtsson, then Localisation Director at Oracle, put it in an interview with the LRC. Projects that are always different from previous ones do not lend themselves to automation. Highly creative pioneers, capable of dealing with problems as they arise, are generally not overly enthusiastic about the development of reproducible standard processes. Custom made tools and technologies, fine tuned to highly specific customer dependent requirements are unlikely to be interoperable.

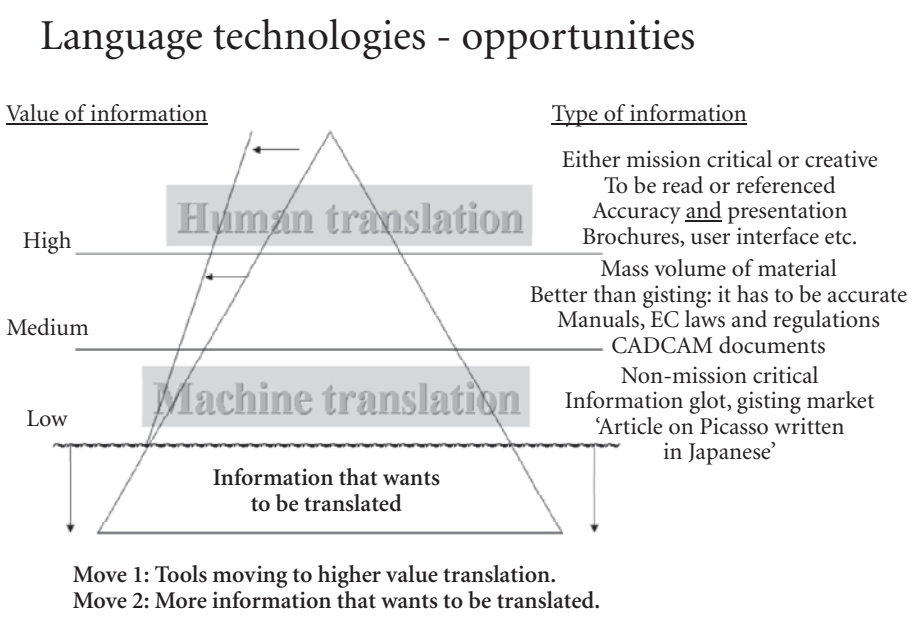

Figure 1. Language technologies – opportunities

Yet, as we have seen developments have taken place in recent years that have demonstrated the validity of standard approaches to localisation problems. This has lead to a readjustment of the way localisation is being perceived and to a significant increase in efforts around standardisation. Today, the most progressive and advanced localisation operations highlight and take advantage of commonalities in the localisation process, as demonstrated in the earlier case study.

Among these commonalities are:

- Frequent updates
  Only in exceptional cases is digital material localised for the first time; most of the localisation effort is spent on updates to previously localised material.
- Short product cycles
  Digital material has an extremely short shelf live, product cycles of months or even years are virtually unknown; timely delivery of localised versions is therefore a crucial business requirement.
- Composition of material
  Typically, the material to be localised is highly repetitive in itself and especially across different versions (updates); it can be composed of:
  - Non-translatables (5%) – text fragments that should not be translated, such as company and product names
  - Unknown (10%) – new text not known from previous versions
  - Known (15%) – text that has only been modified slightly, such as part numbers, small linguistic corrections/modifications
  - Unchanged (70%) – text that has been carried over unchanged from a previous version
- Consistency requirements
  Most projects are being completed by large teams of translators who achieve consistency between their translations by implementing stringent terminology guidelines and by using translation memory technologies. Consistency is required within a translation, between different versions, between different products of the same publisher, and across application types and operating systems.

We have identified two fundamental problems that are causing tremendous difficulties for localisation automation efforts. These are:

1. Mark-up and formatting of source material (finding a needle in the haystack)

How can the material to be localised be *identified* among tens of thousands of files, unambiguously *marked-up* and *formatted* in a generally accepted standard so that it can be processed by any tool or technology complying with this standard? There is a believe that the main barrier to cheaper, faster and better localisation is not

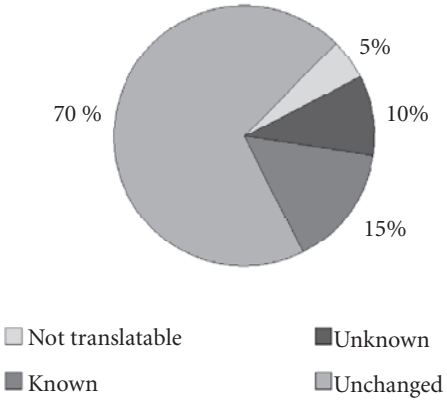

Figure 2. Typical localisation project – make up of contents (clock-wise values / left-to-right legenda)

the lack of (linguistic) tools and technologies but the multitude of file formats and processes in use which, it should be noted, still often prevents the full sharing of linguistic resources and assets between different types of tools.

2. Complexity of localisation process (too many cooks not only spoil the broth – they also make it very expensive)

How can material be localised with a minimum of manual intervention? How can repetitive tasks be avoided? Transactions are not sufficiently automated and, if they are, not in a generally accepted standardised way. Clients continue to pay high rates to multi-language vendors (MLVs). Generally, there is little or no choice between different (language) technologies.

As we have seen, language technologies have already been used successfully in proprietary environments. They will be used by the wider, general multilingual digital content industry only when the fundamental markup and formatting problem in relation to the source material has been solved. Many experts agree that *XLIFF: XML-based Localisation Interchange File Format* for the identification and exchange of material to be localised, goes a long way towards solving this problem (XLIFF 2007).

Only when vendors and clients can use the language technologies of their choice within a connected web services environment that automates the most important transactions, only when this almost endlessly repetitive and incredibly expensive cascading supply chain has been cut back to basics, only then will the market for language technologies open up, be able to grow even more and become accessible and affordable even to companies that currently cannot consider localising (at least certain aspects of) their services and products. *Trans-WS: Translation*

Vendor Web Services is being developed by a group of companies to address this challenge (Trans-WS 2007).

There are a number of key considerations that have to be taken into account for the development of frameworks that address these fundamental problems. They need to be:

- (Preferably) vendor independent
  Any solution on offer should preferably be vendor independent; while a tendency can be observed that place tools and technology solutions increasingly in the hands of vendors, e.g., SDL and SDLX/TRADOS/Passolo/Idiom, Lionbridge and Freeway, there are also a number of vendor-independent solution providers who are capturing a portion of the market, such as across/Nero (across).
- Based on open standards
  Proprietary solutions offered by most solution providers currently make it difficult for users switching between applications, sharing and reusing resources referred to by these applications, leading to what is referred to as "lock-in"; to protect users' investment in these resources (or assets), they need to be based on open standards.
- Interoperable
  Different technologies and tools address problems arising at different touch points within the localisation process; while standards such as XLIFF and Trans-WS described earlier address interoperability issues, they do not resolve them completely. Any technology solution must be thoroughly tested for interoperability by independent and credible authorities.
- Strong on IPR
  Discussions around international services and trade in the age of the information and knowledge society are dominated by questions of ownership of the relevant intellectual property rights (IPR). Standards and technology solutions, especially those developed by consortia, need a clear and unambiguous IPR policy that will shield users from later, unexpected claims for royalties.
- Responsive, reliable, stable
  As any other business, any solution offered in this space needs to come from a responsive, reliable and stable source that has a clear commitment for many years to come.

Finally, any new framework needs to be affordable and easily accessible especially to SMEs and even individual localisers.

One such framework, IGNITE, a collaboration between the European Union and five consortium partners, was developed between 2005 and 2007 based on a linguistic infrastructure for localisation that covered language data, tools and standards. It resulted in a *demonstrator* of an automated open standards-based lo-

calisation environment, complying with many of the requirements and boundaries outlined here so far.

4. Linguistic infrastructure for localisation 2.0

The term "Localisation 2.0", first coined by Lionbridge's CEO Rory Cowan in 2005, is a reference to Web 2.0 and is an umbrella term referring to "next generation" automated electronic content localisation workflow. Although industry analysts Common Sense Advisory mock the L10N 2.0 term as a meaningless buzzword to anyone but localisation industry insiders (Global Watchtower 2007) the repeated use of this term in the printed media and even by Common Sense Advisory signifies recognition of a new evolutionary era in the localisation industry. The Localisation 2.0 era is based on a number of significant technical and business developments in the industry:

– Growing maturity and industry adoption of localisation related XML content Industry Standards such as W3C's ITS (Internationalisation Tag Set), OASIS' XLIFF, Translation Web Services and DITA, and LISA OSCAR's TMX and TBX. These standards enable seamless unification of enterprise content authoring and management within the localisation process. The net effect of using these standards is an increase in content reuse (original source as well as localised target content), a reduction in time to market for localised content, and elimination of most of the manual labour associated with localisation (e.g. desktop publishing (DTP), file management, manual testing and translation).
– Globalisation Management Systems (GMS) availability as hosted Software as a Service (SaaS) enables wider market access and awareness of optimised enterprise localisation automation workflow systems. GMS' in the past have been inaccessible due to their very high entry price point as well as very high running costs. The "pay-in-advance" model for GMS systems is now being replaced with the SaaS "pay-as-you-go" model. Since 2006, the industry has seen major LSP's and tools vendors SDL, Lionbridge and others offer much lower cost hosted GMS solutions. Lionbridge is offering free user registration for its GMS Freeway 2.0, and has already registered over 10,000 users. Industry reports suggest fast growing demand for lower cost SaaS GMS solutions will continue.
– Availability and commercial adoption of Free / Libre Open Source (FLOSS) tools and technology continues to grow. Robust and highly efficient Open Source operating systems such as Linux and enterprise systems components and tools such as MySQL database and Hibernate framework, and Eclipse development environment have forever changed the revenue model for soft-

ware products from payment. The Commercial Open Source model has been proven a successful business model in other industry sectors such as CRM and ERP. At present, there is no end-to-end FLOSS GMS solution available to the market. *]project open[* comes close, but it is geared to managing localisation services providers' business administration and invoicing requirements. *]project open[* lacks CMS integration capabilities and is not based on XML based web services, content management or most localisation industry standards (TMX excluded). Finally, although not Open Source, Lionbridge's Freeway 2.0's free user registration provides evidence that customers of localisation services are seeking robust but inexpensive (or possibly free) technology solutions to address content localisation challenges.

The challenges of the Localisation 2.0 paradigm require the localisation industry to become one of the early adaptors of language technologies. Yet, up to very recently, few concerted efforts had been undertaken to create a robust infrastructure for the localisation industry comprising language data, tools and standards.

In 2005, five European organisations, all stakeholders in different aspects of localisation, formed the IGNITE consortium to work out a model for Localisation 2.0. The consortium was coordinated by the Localisation Research Centre (LRC) at the University of Limerick and had four industrial partners:

- Archetypon, a Greek-based IT developer and localisation service provider
- Pass Engineering, a leading German-based localisation tools and technology developer
- VeriTest, a division of Lionbridge, based in Ireland and world leader in standards compliance testing
- Vivendi Games, one of the world's leading video games developers based in Ireland

IGNITE proposes a radical new approach to localisation that aims to disrupt the way localisation is done today. Its approach is based on open standards, tried, tested and further developed using linguistic resources developed within the project. Its implementation, finalised in early 2007 at the Localisation Research Centre (LRC), is a prototype. It is not ready to be used (yet) in a live production environment, but it is capable of demonstrating a real version of localisation 2.0 that has the potential to shake up the industry as we know it.

We will now introduce the major components of IGNITE developed by the consortium over a two year period, namely:

- IGNITE Information Repository
- Standards and Tools Certification
- Localisation Memory
- Localisation Factory

Each of these form part of the IGNITE system that demonstrates the functional viability of the approach taken by the consortium.

4.1 IGNITE Information Repository

The IGNITE Information Repository is an infrastructure that allows the collection, storage and maintenance of a wide variety of multimodal, multilingual and cross cultural language data for use in the localisation industry and makes it available to industry professionals through a state-of-the-art web portal. Data held in the repository includes digital linguistic resources, terminology databases and translation memories, while external links provide information on linguistic tools and technologies, standards, guidelines and best practices in localisation.

It currently holds more than 1,000 files across 62 file types; information on approximately 180 international standard bodies; links to dozens of relevant publications and organisations as well as to third party online linguistic resources (terminology, glossaries, dictionaries); descriptions and information on close to 50 relevant tools; more than 2,200 reviewed entries in the professional directory.

The IGNITE Information Repository is a unique resource with an unmatched functionality that is freely available to the localisation community. Similar repositories are built by a consortium under the umbrella of the Translation Automation User Society (TAUS). However, access to these repositories is limited to its subscribers (TAUS 2004). Fully localised into French, German, Italian, Spanish and Greek, the IGNITE Information Repository provides the perfect technical infrastructure for the collection, maintenance and publication of information that is essential for localisation professionals. In addition to being a useful resource for localisation professionals, it is also a significant element in what will become a robust and effective linguistic infrastructure for localisation.

4.2 Standards and tools certification

Standards are a prerequisite to automation and interoperability. Yet, there is much confusion about the effectiveness of standards.

IGNITE developed a genesis of standards applied to localisation tools that demonstrates how different factors influence their usefulness and their potential uptake, from the "early days" through to today.

- **Individual Effort** Issues are dealt with in a fire fighting mode
- **In-house Tools Support** Individual tasks are automated for specific aspects of the localisation process
- **Third Party Tools** Specialised tools companies start to offer solutions for aspects of the process

– **Best of Breed** Tools and processes start to be combined, e.g., terminology tools and translation memories
– **"Ad-Hoc" Standards** Groups start to promote their own approaches as standards, e.g., TMX and TBX
– **Open Standards** Large sections of the industry join to develop open standards that undergo a rigorous review process, have a watertight policy on IPR and require a substantial uptake as a condition for their recognition as a standard, e.g., XLIFF and Trans-WS

IGNITE designed processes and frameworks for the certification of tools and technologies to encourage the uptake of standards.

4.3 Localisation memory

An IGNITE *fast track* project that concentrated on the XML-based Localisation Interchange File Format (XLIFF) and on the Translation web services (Trans-WS) standard work carried out under the umbrella of OASIS (www.oasis-open.org) lead to the development of the "Localisation Memory", an independent, open standard-based, extendible container of all information that is deemed to be relevant for localisation. Using the IGNITE Factory approach, this localisation mem-

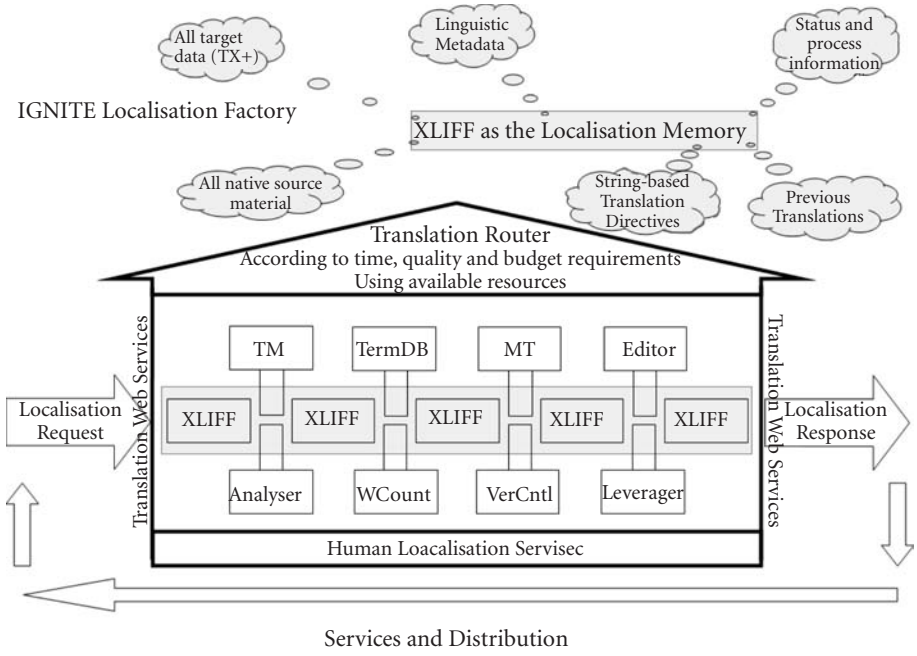

Figure 3. The IGNITE Localisation Factory

ory can potentially be used from design to customer release and serve as a single, easily maintainable and accessible resource for localisers that is independent of the tools and technologies used by them. The localisation memory is the backbone of the IGNITE localisation factory and it is transparent to localisers who can continue to use their preferred localisation technology resources.

4.4 The IGNITE Localisation Factory

IGNITE produced a demonstrator of a configurable, modularized, and extendible automated localisation environment based on open standards and using an XLIFF-based localisation memory as a backbone, while using Trans-WS rules for its interfaces with clients and service providers.

The IGNITE Factory uses configurable converters for non-XLIFF file formats at its entrance and exit points and already supports a large variety of file formats. Leveraging and editor components have already been implemented.

5. Conclusions

The localisation industry emerged in the mid 1980s as a function of international sales departments who had to provide *localised* versions of typical office applications to their expanding international customer base. Over the following decade, an *enterprise localisation* model evolved that was supported by a growing number of desktop-based tools for translators and engineers working on localisation projects that were limited to what now seems like a small range of applications, platforms, languages and media. Localisation projects were scheduled with defined *deltas*, where localised versions were released in tiers, i.e., groups of language versions bundled for a specific release date according to the importance of the market they covered. Ten years later, in 2005, Lionbridge's CEO Rory Cowan announced the arrival of Localisation 2.0. This model is largely driven by *consumer localisation* requirements that are less structured, often *ad hoc* and much more varied in terms of the number of languages and the type of digital media covered. A multi-million euro Irish government supported research initiative, *Next Generation Localisation*, was announced in 2007 to carry out the fundamental and applied research underpinning the design, development, implementation and evaluation of the blueprints for the Next Generation Localisation Factory and addressing the challenges described above (Science Foundation Ireland 2007).

Traditional, often stand-alone localisation approaches are no longer adequate in the networked world of immediate localisation demands. Large, experienced multinational publishers and service providers with adequate in-house development capacities have already begun to move towards Localisation 2.0 by investing

heavily in bespoke automated and web-based localisation solutions. We have described one such model in our case study.

Today's challenge is twofold:

1. The industry, supported by targeted publicly funded research must continue to build an adequate linguistic infrastructure for localisation, covering language data, tools and standards;
2. Bespoke solutions must be moved into a space where they become usable, accessible and affordable for the 75% of localisers made up of entrepreneurs and small and medium sised enterprise (SMEs) (Boucau 2006).

The IGNITE project, co-funded by the European Union and its five consortium members, went some way in demonstrating what an adequate response to the challenges of Localisation 2.0 could look like. Following the development of the IGNITE Localisation Factory and Repository, a laboratory-based localisation environment was set up to evaluate their impact and performance on typical localisation projects. The business process and the requirements of a "real-life" Localisation Service Provider (LSP), Archetypon S.A., provided an ideal test case to measure the performance and the impact of the IGNITE approach. Archetypon's performance evaluation demonstrated that the IGNITE Factory and the support of XLIFF and Trans-WS offered by it has the potential to supply real and comprehensive solutions to a series of localisation challenges encountered today by localisers, among them: interoperability between tools, support for the overall localisation workflow and the necessity of localisation tool developers to support a growing number of different formats in addition to a large number of proprietary intermediate formats.

The current IGNITE framework will be further developed following the advice of leading international experts and in cooperation with supporting projects, such as the "Localization4all" project at the University of Limerick (Localization4all 2008). We have already embarked on a dual approach which includes university-based research and development as well as commercial exploitation for some of IGNITE's core components. There has been significant interest in IGNITE's potential to provide an adequate response to the challenges of Localisation 2.0. One of our intentions is to integrate the IGNITE components with additional open source localisation technologies in an *IGNITE Open Source Localisation Distribution* that will not just become available to large multinationals but also to thousands of localisers in so-called "emerging markets". This distribution will allow them to gain access to state-of-the-art localisation technology and will, therefore, significantly increase their capacity to not just offer cut-price localisation services to multinational clients, but – much more significantly – also to start producing localised versions of their native content.

Acknowledgements

The author would like to acknowledge the support of the European Union's eContent programme for the IGNITE project (EDC-22275), the support of the consortium partners (Archetypon, PASS, Vivendi Games and Veritest) to the successful completion of the project and the input from Product Innovators' Director of Research and Development, Tony Jewtushenko, assisted by Anastasia Violatou of the Athens University of Economics & Business to the completion of IGNITE's exploitation plan and, subsequently, aspects of this chapter.

References

Ballmer, S. (2002) Microsoft reports substantial growth. In: *The Irish Times*, 21 October 2002.

Ballmer, S. (2007) *Windows Vista and the 2007 Microsoft Office System Worldwide Availability Celebration*. Remarks by Bill Gates, Chairman, and Steve Ballmer, CEO, Microsoft Corporation. Windows Vista and the Microsoft 2007 Office System Worldwide Availability Celebration. The Windows Vista Theatre – Times Square. New York, N.Y. 29th January 2007. http://www.microsoft.com/presspass/exec/billg/speeches/2007/01-29VistaOfficeLaunch.mspx

Boucau, F. (2006) *The European Translation Markets Updated Facts and Figures 2006–2010*. European Union of Associations of Translation Companies (EUATC) http://66.102.9.104/search?q=cache:wb6K6Cyyd9EJ:www.euatc.org/conferences/pdfs/2006/Boucau_FactsAndFigures.pdf+EUATC+translation+market&hl=en&ct=clnk&cd=2&gl=ie

DePalma, D. and R. Beninatto (2002) *Beggars at the Globalisation Banquet*. Common Sense Advisory. http://www.commonsenseadvisory.com/research/report_view.php?id=1#, http://www.commonsenseadvisory.com/pdf/BeggarsPR.pdf

DePalma, D. and R. Beninatto (2006) *Ranking of Top 20 Translation Companies for 2005*. Common Sense Advisory. http://www.commonsenseadvisory.com/members/res_cgi.php/060301_QT_top_20.php#table2#table2

Forrester Consulting (2007) *The Total Economic Impact of Global Information Management.* Prepared for SDL. Cambridge, Massachussetts, USA. (A copy of the report can be downloaded from http://www.sdl.com/en/globalisation-knowledge-centre/research_results/tei-report.asp after registration.)

Global Watchtower (2007) *Lionbridge announces 2006 Results*. http://www.commonsenseadvisory.com/news/global_watchtower_one.php?wat_id=364

IGNITE (2007) *Linguistic Infrastructure for Localisation: Language Data, Tools and Standards*. www.igniteweb.org

Localization4all (2008) *Localization4all* – A project of the Localisation Research Centre (LRC) at the University of Limerick, Ireland. http://www.localization4all.com/

SDL (2008) *$4.7 bn: cost of 'missed opportunity' due to poor localisation by global business.* Independent commissioned study finds global businesses lose substantial market share when localisation is not up-to-scratch. Press Release (6th March 2008). SDL Maidenhead, United Kingdom. http://www.sdl.com/en/events/news-PR/4.7bn-dollar-cost-of-missed-opportunity-due-to-poor-localization-by-global-business.asp

Science Foundation Ireland. (2007) *Next Generation Localisation*. http://www.sfi.ie/content/content.asp?section_id=674&language_id=1

TAUS (2004) *TAUS established at Localisation World – November 2004.* http://www.translationautomation.com/joomla/index.php?view=article&catid=45%3Anews_archive&id=83%3Ataus-established-at-localization-world-november-2004&option=com_content&Itemid=69

Translation Webservices (2007) *Automating the translation and localisation process as a Web service.* http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=trans-ws

Wordbank (2005) *Are You Talkin' to Me? An international market study into the impact of language on a consumer's purchasing behaviour*. http://babelport.com/articles/pdf/areyoutalkingtome.pdf

XLIFF (2007) *The XML-based Localisation Interchange File Format. Advancing a multilingual data exchange standard for localisation*. http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff

Index

Benjamins Translation Library

A complete list of titles in this series can be found on *www.benjamins.com*

83 **TORIKAI, Kumiko:** Voices of the Invisible Presence. Diplomatic interpreters in post-World War II Japan. vii, 191 pp. + index. *Expected February 2009*

82 **BEEBY, Allison, Patricia RODRÍGUEZ INÉS and Pilar SÁNCHEZ-GIJÓN (eds.):** Corpus Use and Translating. Corpus use for learning to translate and learning corpus use to translate. x, 151 pp. + index. *Expected January 2009*

81 **MILTON, John and Paul BANDIA (eds.):** Agents of Translation. vi, 329 pp. + index. *Expected January 2009*

79 **YUSTE RODRIGO, Elia (ed.):** Topics in Language Resources for Translation and Localisation. 2008. xii, 220 pp.

78 **CHIARO, Delia, Christine HEISS and Chiara BUCARIA (eds.):** Between Text and Image. Updating research in screen translation. 2008. x, 292 pp.

77 **DÍAZ CINTAS, Jorge (ed.):** The Didactics of Audiovisual Translation. 2008. xii, 263 pp. (incl. CD-Rom).

76 **VALERO-GARCÉS, Carmen and Anne MARTIN (eds.):** Crossing Borders in Community Interpreting. Definitions and dilemmas. 2008. xii, 291 pp.

75 **PYM, Anthony, Miriam SHLESINGER and Daniel SIMEONI (eds.):** Beyond Descriptive Translation Studies. Investigations in homage to Gideon Toury. 2008. xii, 417 pp.

74 **WOLF, Michaela and Alexandra FUKARI (eds.):** Constructing a Sociology of Translation. 2007. vi, 226 pp.

73 **GOUADEC, Daniel:** Translation as a Profession. 2007. xvi, 396 pp.

72 **GAMBIER, Yves, Miriam SHLESINGER and Radegundis STOLZE (eds.):** Doubts and Directions in Translation Studies. Selected contributions from the EST Congress, Lisbon 2004. 2007. xii, 362 pp. **[EST Subseries 4]**

71 **ST-PIERRE, Paul and Prafulla C. KAR (eds.):** In Translation – Reflections, Refractions, Transformations. 2007. xvi, 313 pp.

70 **WADENSJÖ, Cecilia, Birgitta ENGLUND DIMITROVA and Anna-Lena NILSSON (eds.):** The Critical Link 4. Professionalisation of interpreting in the community. Selected papers from the 4th International Conference on Interpreting in Legal, Health and Social Service Settings, Stockholm, Sweden, 20-23 May 2004. 2007. x, 314 pp.

69 **DELABASTITA, Dirk, Lieven D'HULST and Reine MEYLAERTS (eds.):** Functional Approaches to Culture and Translation. Selected papers by José Lambert. 2006. xxviii, 226 pp.

68 **DUARTE, João Ferreira, Alexandra ASSIS ROSA and Teresa SERUYA (eds.):** Translation Studies at the Interface of Disciplines. 2006. vi, 207 pp.

67 **PYM, Anthony, Miriam SHLESINGER and Zuzana JETTMAROVÁ (eds.):** Sociocultural Aspects of Translating and Interpreting. 2006. viii, 255 pp.

66 **SNELL-HORNBY, Mary:** The Turns of Translation Studies. New paradigms or shifting viewpoints? 2006. xi, 205 pp.

65 **DOHERTY, Monika:** Structural Propensities. Translating nominal word groups from English into German. 2006. xxii, 196 pp.

64 **ENGLUND DIMITROVA, Birgitta:** Expertise and Explicitation in the Translation Process. 2005. xx, 295 pp.

63 **JANZEN, Terry (ed.):** Topics in Signed Language Interpreting. Theory and practice. 2005. xii, 362 pp.

62 **POKORN, Nike K.:** Challenging the Traditional Axioms. Translation into a non-mother tongue. 2005. xii, 166 pp. **[EST Subseries 3]**

61 **HUNG, Eva (ed.):** Translation and Cultural Change. Studies in history, norms and image-projection. 2005. xvi, 195 pp.

60 **TENNENT, Martha (ed.):** Training for the New Millennium. Pedagogies for translation and interpreting. 2005. xxvi, 276 pp.

59 **MALMKJÆR, Kirsten (ed.):** Translation in Undergraduate Degree Programmes. 2004. vi, 202 pp.

58 **BRANCHADELL, Albert and Lovell Margaret WEST (eds.):** Less Translated Languages. 2005. viii, 416 pp.

57 **CHERNOV, Ghelly V.:** Inference and Anticipation in Simultaneous Interpreting. A probability-prediction model. Edited with a critical foreword by Robin Setton and Adelina Hild. 2004. xxx, 268 pp. **[EST Subseries 2]**

56 **ORERO, Pilar (ed.):** Topics in Audiovisual Translation. 2004. xiv, 227 pp.

55 **ANGELELLI, Claudia V.:** Revisiting the Interpreter's Role. A study of conference, court, and medical interpreters in Canada, Mexico, and the United States. 2004. xvi, 127 pp.

54 **GONZÁLEZ DAVIES, Maria:** Multiple Voices in the Translation Classroom. Activities, tasks and projects. 2004. x, 262 pp.

53 **DIRIKER, Ebru:** De-/Re-Contextualizing Conference Interpreting. Interpreters in the Ivory Tower? 2004. x, 223 pp.

52 **HALE, Sandra:** The Discourse of Court Interpreting. Discourse practices of the law, the witness and the interpreter. 2004. xviii, 267 pp.

51 **CHAN, Leo Tak-hung:** Twentieth-Century Chinese Translation Theory. Modes, issues and debates. 2004. xvi, 277 pp.

50 **HANSEN, Gyde, Kirsten MALMKJÆR and Daniel GILE (eds.):** Claims, Changes and Challenges in Translation Studies. Selected contributions from the EST Congress, Copenhagen 2001. 2004. xiv, 320 pp. **[EST Subseries 1]**

49 **PYM, Anthony:** The Moving Text. Localization, translation, and distribution. 2004. xviii, 223 pp.

48 **MAURANEN, Anna and Pekka KUJAMÄKI (eds.):** Translation Universals. Do they exist? 2004. vi, 224 pp.

47 **SAWYER, David B.:** Fundamental Aspects of Interpreter Education. Curriculum and Assessment. 2004. xviii, 312 pp.

46 **BRUNETTE, Louise, Georges BASTIN, Isabelle HEMLIN and Heather CLARKE (eds.):** The Critical Link 3. Interpreters in the Community. Selected papers from the Third International Conference on Interpreting in Legal, Health and Social Service Settings, Montréal, Quebec, Canada 22–26 May 2001. 2003. xii, 359 pp.

45 **ALVES, Fabio (ed.):** Triangulating Translation. Perspectives in process oriented research. 2003. x, 165 pp.

44 **SINGERMAN, Robert:** Jewish Translation History. A bibliography of bibliographies and studies. With an introductory essay by Gideon Toury. 2002. xxxvi, 420 pp.

43 **GARZONE, Giuliana and Maurizio VIEZZI (eds.):** Interpreting in the 21st Century. Challenges and opportunities. 2002. x, 337 pp.

42 **HUNG, Eva (ed.):** Teaching Translation and Interpreting 4. Building bridges. 2002. xii, 243 pp.

41 **NIDA, Eugene A.:** Contexts in Translating. 2002. x, 127 pp.

40 **ENGLUND DIMITROVA, Birgitta and Kenneth HYLTENSTAM (eds.):** Language Processing and Simultaneous Interpreting. Interdisciplinary perspectives. 2000. xvi, 164 pp.

39 **CHESTERMAN, Andrew, Natividad GALLARDO SAN SALVADOR and Yves GAMBIER (eds.):** Translation in Context. Selected papers from the EST Congress, Granada 1998. 2000. x, 393 pp.

38 **SCHÄFFNER, Christina and Beverly ADAB (eds.):** Developing Translation Competence. 2000. xvi, 244 pp.

37 **TIRKKONEN-CONDIT, Sonja and Riitta JÄÄSKELÄINEN (eds.):** Tapping and Mapping the Processes of Translation and Interpreting. Outlooks on empirical research. 2000. x, 176 pp.

36 **SCHMID, Monika S.:** Translating the Elusive. Marked word order and subjectivity in English-German translation. 1999. xii, 174 pp.

35 **SOMERS, Harold (ed.):** Computers and Translation. A translator's guide. 2003. xvi, 351 pp.

34 **GAMBIER, Yves and Henrik GOTTLIEB (eds.):** (Multi) Media Translation. Concepts, practices, and research. 2001. xx, 300 pp.

33 **GILE, Daniel, Helle V. DAM, Friedel DUBSLAFF, Bodil MARTINSEN and Anne SCHJOLDAGER (eds.):** Getting Started in Interpreting Research. Methodological reflections, personal accounts and advice for beginners. 2001. xiv, 255 pp.

32 **BEEBY, Allison, Doris ENSINGER and Marisa PRESAS (eds.):** Investigating Translation. Selected papers from the 4th International Congress on Translation, Barcelona, 1998. 2000. xiv, 296 pp.

31 **ROBERTS, Roda P., Silvana E. CARR, Diana ABRAHAM and Aideen DUFOUR (eds.):** The Critical Link 2: Interpreters in the Community. Selected papers from the Second International Conference on Interpreting in legal, health and social service settings, Vancouver, BC, Canada, 19–23 May 1998. 2000. vii, 316 pp.

30 **DOLLERUP, Cay:** Tales and Translation. The Grimm Tales from Pan-Germanic narratives to shared international fairytales. 1999. xiv, 384 pp.

29 **WILSS, Wolfram:** Translation and Interpreting in the 20th Century. Focus on German. 1999. xiii, 256 pp.

28 **SETTON, Robin:** Simultaneous Interpretation. A cognitive-pragmatic analysis. 1999. xvi, 397 pp.

27 **BEYLARD-OZEROFF, Ann, Jana KRÁLOVÁ and Barbara MOSER-MERCER (eds.):** Translators' Strategies and Creativity. Selected Papers from the 9th International Conference on Translation and Interpreting, Prague, September 1995. In honor of Jiří Levý and Anton Popovič. 1998. xiv, 230 pp.

26 **TROSBORG, Anna (ed.):** Text Typology and Translation. 1997. xvi, 342 pp.

25 **POLLARD, David E. (ed.):** Translation and Creation. Readings of Western Literature in Early Modern China, 1840–1918. 1998. vi, 336 pp.

24 **ORERO, Pilar and Juan C. SAGER (eds.):** The Translator's Dialogue. Giovanni Pontiero. 1997. xiv, 252 pp.

23 **GAMBIER, Yves, Daniel GILE and Christopher TAYLOR (eds.):** Conference Interpreting: Current Trends in Research. Proceedings of the International Conference on Interpreting: What do we know and how? 1997. iv, 246 pp.

22 **CHESTERMAN, Andrew:** Memes of Translation. The spread of ideas in translation theory. 1997. vii, 219 pp.

21 **BUSH, Peter and Kirsten MALMKJÆR (eds.):** Rimbaud's Rainbow. Literary translation in higher education. 1998. x, 200 pp.

20 **SNELL-HORNBY, Mary, Zuzana JETTMAROVÁ and Klaus KAINDL (eds.):** Translation as Intercultural Communication. Selected papers from the EST Congress, Prague 1995. 1997. x, 354 pp.

19 **CARR, Silvana E., Roda P. ROBERTS, Aideen DUFOUR and Dini STEYN (eds.):** The Critical Link: Interpreters in the Community. Papers from the 1st international conference on interpreting in legal, health and social service settings, Geneva Park, Canada, 1–4 June 1995. 1997. viii, 322 pp.

18 **SOMERS, Harold (ed.):** Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager. 1996. xii, 250 pp.

17 **POYATOS, Fernando (ed.):** Nonverbal Communication and Translation. New perspectives and challenges in literature, interpretation and the media. 1997. xii, 361 pp.

16 **DOLLERUP, Cay and Vibeke APPEL (eds.):** Teaching Translation and Interpreting 3. New Horizons. Papers from the Third Language International Conference, Elsinore, Denmark, 1995. 1996. viii, 338 pp.

15 **WILSS, Wolfram:** Knowledge and Skills in Translator Behavior. 1996. xiii, 259 pp.

14 **MELBY, Alan K. and Terry WARNER:** The Possibility of Language. A discussion of the nature of language, with implications for human and machine translation. 1995. xxvi, 276 pp.

13 **DELISLE, Jean and Judith WOODSWORTH (eds.):** Translators through History. 1995. xvi, 346 pp.

12 **BERGENHOLTZ, Henning and Sven TARP (eds.):** Manual of Specialised Lexicography. The preparation of specialised dictionaries. 1995. 256 pp.

11 **VINAY, Jean-Paul and Jean DARBELNET:** Comparative Stylistics of French and English. A methodology for translation. Translated and edited by Juan C. Sager and M.-J. Hamel. 1995. xx, 359 pp.

10 **KUSSMAUL, Paul:** Training the Translator. 1995. x, 178 pp.

9 **REY, Alain:** Essays on Terminology. Translated by Juan C. Sager. With an introduction by Bruno de Bessé. 1995. xiv, 223 pp.

8 **GILE, Daniel:** Basic Concepts and Models for Interpreter and Translator Training. 1995. xvi, 278 pp.

7 **BEAUGRANDE, Robert de, Abdullah SHUNNAQ and Mohamed Helmy HELIEL (eds.):** Language, Discourse and Translation in the West and Middle East. 1994. xii, 256 pp.

6 **EDWARDS, Alicia B.:** The Practice of Court Interpreting. 1995. xiii, 192 pp.

5 **DOLLERUP, Cay and Annette LINDEGAARD (eds.):** Teaching Translation and Interpreting 2. Insights, aims and visions. Papers from the Second Language International Conference Elsinore, 1993. 1994. viii, 358 pp.

4 **TOURY, Gideon:** Descriptive Translation Studies – and beyond. 1995. viii, 312 pp.

3 **LAMBERT, Sylvie and Barbara MOSER-MERCER (eds.):** Bridging the Gap. Empirical research in simultaneous interpretation. 1994. 362 pp.

2 **SNELL-HORNBY, Mary, Franz PÖCHHACKER and Klaus KAINDL (eds.):** Translation Studies: An Interdiscipline. Selected papers from the Translation Studies Congress, Vienna, 1992. 1994. xii, 438 pp.

1 **SAGER, Juan C.:** Language Engineering and Translation. Consequences of automation. 1994. xx, 345 pp.