Third International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III)

Chaired by **Elia YUSTE (University of Zurich)**

A satellite event of

LREC 2006
22 - 28 May
Genoa - ITALY

FIFTH LANGUAGE RESOURCES AND EVALUATION CONFERENCE

28 May 2006
Magazzini del Cotone Conference Centre, GENOA - ITALY

Workshop Website: www.ifi.unizh.ch/cl/yuste/LR4Trans-III
Conference Website: www.lrec-conf.org/lrec2006/

The LR4Trans-III Workshop Programme

09.00 – 09.15	Opening by Workshop Chair
09.15 – 09.45	*Translatability of words denoting emotions: the case of unhappiness in a Japanese-English Parallel Corpus*, by Fumiko **Kondo** [CiTP]*
09.45 – 10.15	*Program Integrated Information ID Tag for Translation Verification Test*, by Naotaka **Kato** and Makoto **Arisawa**
10.15 – 11.00	KEYNOTE 1: *Standards for Content Creation and Globalization*, by Sue Ellen **Wright**
11.00 – 11.30	Coffee break
11.30 – 12.15	KEYNOTE 2: *Corpora for translator education and translation practice - Achievements and challenges*, by Silvia **Bernardini** [CiTP]*
12.15 – 12.45	*Translation as problem solving: uses of comparable corpora*, by Serge **Sharoff** [CiTP]*
12.45 – 13.15	*The Use of Corpora in Translator Training in the African Language Classroom: a Perspective from South Africa*, by Rachélle **Gauton** [CiTP]*
13.15 – 13.30	*Mellange* (Project Flyer Distribution)
13.30 – 14.30	Lunch break
14.30 – 15.15	Poster session: *Standardizing the management and the representation of multilingual data: the MultiLingual Information Framework*, by **Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret and Isabelle Kramer** & *A Platform for the Empirical Analysis of Translation Resources, Tools and their Use*, by David **Day**, Galen **Williamson**, Alex **Yeh**, Keith **Crouch**, Sam **Bayer**, Jennifer **DeCamp**, Angel **Asencio**, Seamus **Clancy** and Flo **Reeder**
15.15 – 16.00	KEYNOTE 3: *Using corpus information to improve MT quality*, by Gregor **Thurmair**
16.00 – 16.30	*The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators*, by Youcef **Bey**, Christian **Boitet** and Kyo **Kageura**

16.30 – 17.00	Coffee break
17.00 – 17.30	*Corpógrafo – Applications*, by Belinda **Maia** and Luís **Sarmento** [CiTP]*
17.30 – 18.45	KEYNOTE 4: *xml:tm - a radical new approach to translating XML based documents*, by Andrzej **Zydron**
18.45 – 19.00	Wrap-up & Closing by Workshop Chair

* The **Third Edition of the Language Resources for Translation Work, Research and Training Workshop** has a workshop special track called Corpora in Translation Practice, CiTP for short, which has been co-directed by Serge **Sharoff** (University of Leeds).

Workshop Chair

Elia Yuste
(Institut für Computerlinguistik der Universität Zürich, Switzerland)

Other Workshop Organisers

Serge Sharoff *(University of Leeds, UK)* → Co-responsible for the Corpora in Translation Practice special workshop track

Detlef Reineke *(Universidad de Las Palmas de Gran Canaria, Spain)* → Advisor in the localisation and translation technology workshop aspects

Paul Rayson *(University of Lancaster, UK)* → Advisor in corpora-related matters

Workshop Programme Scientific Committee

Frank Austermuehl *(University of Auckland, New Zealand)*
Amparo Alcina *(Universitat Jaume I, Castelló, Spain)*
Bogdan Babych *(University of Leeds, UK)*
Toni Badia *(Universitat Pompeu Fabra, Barcelona, Spain)*
Silvia Bernardini *(Università di Bologna a Forlì, Italy)*
Lynne Bowker *(University of Ottawa, Canada)*
Gerhard Budin *(Universität Wien, Spain)*
Gloria Corpas *(Universidad de Málaga, Spain)*
Silvia Hansen-Schirra *(Universität des Saarlandes, Germany)*
Kyo Kageura *(University of Tokyo, Japan)*
Anthony Hartley *(University of Leeds, UK)*
Eduard Hovy *(University of Southern California, USA)*
Hitoshi Isahara *(National Institute of Information and Communications Technology, Japan)*
Natalie Kübler *(Université Paris 7, France)*
Belinda Maia *(Facultade de Letras da Universidade do Porto,Portugal)*
Olga Mudraya *(University of Lancaster, UK)*
Maeve Olohan *(University of Manchester, UK)*
Paul Rayson *(University of Lancaster, UK)*
Detlef Reineke *(Universidad de Las Palmas de Gran Canaria, Spain)*
Celia Rico *(Universidad Europea de Madrid, Spain)*
Reinhard Schäler *(The Localisation Research Centre, University of Limerick,Ireland)*
Serge Sharoff *(University of Leeds, UK)*
Erich Steiner *(Universität des Saarlandes, Germany)*
Krista Varantola *(Tampeen yliopisto, Finland)*
Elia Yuste *(Institut für Computerlinguistik der Universität Zürich, Switzerland)*
Federico Zanettin *(Università per Stranieri di Perugia, Italy)*

Table of Contents

Author Index

Translatability of words denoting emotions: the case of unhappiness in a Japanese-English Parallel Corpus

Fumiko Kondo

University of Birmingham
Birmingham, B29 6FQ, UK
fxk312@bham.ac.uk

Abstract

This work presents a comparative study of an emotion word, *unhappy*, in one Japanese-English Parallel Corpus, the Japanese-English News Article Alignment Data. I aim to find out how the concept of a very basic emotion, unhappiness, differs in English and Japanese. The research methods are corpus-driven by investigating the English word *unhappy*, its Japanese translation equivalents, and their English equivalents in the parallel corpus. The primary interest is to identify the contextual patterns (syntactic and semantic conditions) responsible for the selection of a certain equivalent for a given context. The data in the corpus are too small to allow a real statistical analysis. However, this pilot study is useful in showing tendencies of the concept of unhappiness in English and Japanese and their similarities and differences. The results also show the rich and diverse information which a parallel corpus and translation equivalents can offer.

1. Introduction

There is an ongoing controversy as to whether emotions are innate and universal or whether they are culture-specific. One of the possible ways to look at this question is to compare words of emotion in different languages, questioning how emotions are expressed by human beings in language. To shed light on the issue, I intend to look at one of the most basic and common emotion words; *unhappy*.

As a method for comparing the concept of unhappiness in English and Japanese, I will look at the way *unhappy* is translated into Japanese. If *unhappy* is always translated into a Japanese translation equivalent, it means that the English concept of *unhappy* overlaps with its Japanese translation equivalent; the concept of *unhappy* is universal between the different English and Japanese language communities. On the other hand, if *unhappy* is translated into several Japanese translation equivalents, it means that the concept of *unhappy* does not entirely overlap with its Japanese counterparts; the concept is culture-specific between English and Japanese and further investigation is required in order to discover how the Japanese translation equivalents are different from each other regarding their expressions and meanings.

There are two main ways to look at how *unhappy* is translated into Japanese: dictionaries and parallel corpora (sometimes called translation corpora). In this paper, I will use a parallel corpus rather than a dictionary, since dictionaries tend to look at the meaning of words in isolation from their contexts, which are crucial for analysing how people express an emotion such as unhappiness. Parallel corpora, on the other hand, give a fuller, more complete picture of translation equivalence: as Salkie (2002) effectively shows, parallel corpora often contain translation equivalents which 'are not mentioned in dictionaries'. Therefore, parallel corpora are a milestone towards facilitating translation. From them we can extract translation equivalents as they are used in a specific context. In this study I will use the Japanese-English News Article Alignment Data (JENAAD), which consists of about 5 million English words and 6 million Japanese morphemes. Although the corpus is not quite sufficient to answer ultimately whether the concepts of unhappiness in English and Japanese are similar or different, this analysis is a significant pilot study that shows how corpus linguistics can aid the investigation of the issue of emotions across cultures.

2. Emotions

It is fairly difficult to define the concept of 'emotion'. Dictionaries do not give us clear indications of the meaning of *emotion*. These are two dictionary definitions of *emotion*: 'a strong feeling such as love or anger, or strong feelings in general' (Cambridge Advanced Learner's Dictionary, 2003) and 'an emotion is a feeling such as happiness, love, hear, anger, or hatred, which can be caused by the situation that you are in or the people you are with' (Collins Cobuild Advanced Learner's English Dictionary, 2003). What these dictionaries are presenting are not definitions in a strict sense but paraphrases using a synonym, *feeling* (cf Wierzbicka, 1999 for the difference between *emotion* and *feeling*). Moreover, the definitions of *feeling* in the dictionaries refer back to *emotion*; Collins Dictionary tells us that 'a *feeling* is an emotion, such as anger or happiness'; and more briefly, the Cambridge one defines *feeling* as 'emotion'. The fact that dictionaries resort to such a narrow circularity undermines the enigmatic character of what emotions or feelings are.

Furthermore, there is another fundamental argument about the attributes of emotions, which Darwin (1904) began to discuss in the nineteenth century with his words 'the young and the old of widely different races, both with man and animals, express the same state of mind', which has not yet been concluded. This is the question whether emotions are innate and universal or whether they are culturally acquired. Universalists argue with Darwin that basic emotions are inborn and everybody understands them. On the other hand, social constructionists maintain that emotions are something acquired by growing up in a particular culture. The two groups represent the opposing ends of a wide spectrum. Additionally, there are other researchers who agree with neither of them and who point out that the issue of innateness is not 'all-or-nothing, but a question of degree' (Evans, 2002).

2.1. Emotions in psychology: Ekman

Ekman, a psychologist, believes in the universality of emotion. Ekman (2004b) argues that the seven basic emotions he examines (sadness, anger, surprise, fear, enjoyment, disgust, and contempt) are innate to all cultures because they have a genetic foundation. He has reached this viewpoint through his experiments in which people in Papua New Guinea and Indonesia listened to stories and then were asked to match these stories with one of seven photographs of facial features expressing the seven basic emotions. As a result, he came to the conclusion that the people in those areas seemed to understand emotion in the same way that English-speaking people do.

This, however, is not as straight forward as it may sound because these experiments may be flawed due to the linguistic methodology used. We are not told how translation equivalence was assured, how the stories were translated and whether the interpreters were qualified. For instance, in the experiments regarding happiness in Papua New guinea, Ekman used the English sentence "His/her friends have come and s/he feels very happy" for the examinees to match with the photograph of a facial expression supposedly signifying happiness (Ekman, 2004a). When this story was told to the people, Ekman (*ibid.*) translated the sentence from English to Pidgin and then used translators from Pidgin to Fore, which is the local language in Papua New Guinea. The problem is that Ekman was not able to check the equivalence of the English, Pidgin and Fore expressions. We cannot be sure that they refer to identical concepts. If in the story for happiness the concept of the word, happy, in English differs from the concept of its translational equivalent of Fore, it would be impossible to conclude that happiness is universal. In that case, this experiment only means that a Papua New Guinean understands the word used for translation for *happy*, but not the English concept of *happy*. It is, therefore, not valid to conclude that people, whose language Ekman himself does not understand, have the same concept of happiness that the English have, unless the issue of translation equivalence is taken into account.

2.1.1. Emotions in Language Studies: Wierzbicka

Wierzbicka, a linguist, believes in the social construction of emotions. Wierzbicka (Harkins and Wierzbicka, 2001) claims that emotions 'vary a great deal across languages and cultures'. She demonstrates the cultural-specificity of emotions by using her notion of 'universal concepts' or 'semantic primitives', such as GOOD, FEEL, I, IF, LIKE, HERE, and so on, which are concepts every language has in common. The following is Wierzbicka's bilingual analysis of sadness, shown below with sadness in English (*sadness*) and in Russian (*pečal'*), with her 'universal concepts' (1999).

sadness
(a) X feels something
(b) sometimes a person thinks:
(c) "I know: something bad happened
(d) I don't want things like this to happen
(e) I can't think now: I will do something because of this
(f) I know that I can't do anything"
(g) because of this, this person feels something bad
(h) X feels something like this

pečal'
(a) X felt something <u>because X thought something</u>
(b) sometimes a person thinks:
(c) "I know: something bad happened
(d) <u>this is bad</u>
(e) I don't want things like this to happen
(f) I can't think now: I will do something because of this
(g) I know that I can't do anything"
(h) <u>because this person thinks this</u>, this person feels something bad
(i) X feels something like this
(j) <u>Because X thought something like this</u>
(k) <u>X thought about it for a long time</u>
(l) <u>X felt something because of this for a long time</u>

Both definitions are made up of only her 'universal concepts', thus she believes, making it easy to compare the English (*sadness*) and in Russian (*pečal'*); the differences between them are clear, as shown above by underlining. Although the cultural specifics of sadness in English and Russian seem to be described clearly at first glance with Wierzbicka's 'universal concepts', this is not as straightforward as it many sound either, since universal concepts themselves are fundamentally problematic.

First of all, Wierzbicka believes that 'universal concepts' are 'language-independent', neutral concepts (Harkins and Wierzbicka, 2001). However, those words are obviously English words. It is not possible to determine the borderline between what the English word *feels* means and what the universal concept FEELS means. It is highly unlikely that we can define and treat her sixty 'universal concepts' in a language-independent way. Another significant drawback of Wierzbicka's analysis is that she endeavours to describe the meaning of *sadness* itself as if it always has the same meaning in any context. However, words rarely occur in isolation. Normally they are embedded in their context; the meaning of words is influenced by the words with which they co-occur. This causes an enormous gap between Wierzbicka's analysis of sadness and the usage of this word in reality. Here are some examples from the Bank of English corpus (selected from 3326 citations):

> 'Whatever you are, I am there with you'. There was a sweet *sadness* about all this.
>
> But he was greatly surprised to find, when he looked behind the irritation he felt at having been dragged into this, a curious *sadness* where he would have expected anger to be.

Unlike in Wierzbicka's definition, the *sadness* in the citations above is not described as an entirely negative feeling, but it is used rather as a welcome feeling, as indicated by the adjectives *sweet* or *curious*. Real language data show that word meanings are not always the same; the meaning highly depends on context. There is little point in trying to grasp the meaning of *sadness* in isolation, as Wierzbicka does. Neither Ekman nor Wierzbicka, provide clear answers to the question how universal or language-specific concepts denoting emotions actually are.

Whether emotions are universal or culture-specific is not only an anthropological issue but also a linguistic one, since it raises the question of whether emotion words can be translated into different languages or not. In this study, I would like to clarify those two issues, universality and translatability, by examining the concept of unhappiness

in English and Japanese, and taking account of its context. This should produce more accurate descriptions of meaning and concepts.

3. Methodology

3.1.1. The study of translation equivalence

This study will deal with the English word *unhappy*, its Japanese translation equivalents, and their English translation correspondences. The most traditional and easily accessible method used to identify translation equivalents is to look them up in bilingual dictionaries. In fact, these dictionaries are originally designed for bilingual learners to find translation equivalents of a given word. If we translate into our own native language, bilingual dictionaries, which are comprehensive and frequently updated, are often good enough for finding the proper translation equivalents. However, if we want to translate into a non-native language which we speak imperfectly, bilingual dictionaries normally do not provide sufficient information for choosing the proper equivalent. For this task, they are not necessarily the best tools. There are several reasons for their limitations: For instance, they are always restricted by 'consideration[s] of space' (Johansson, 1998). They are never complete and they are subject to the 'lexicographers competence' (Teubert, 1996). However, the most important drawback of bilingual dictionaries arises from the ambiguity of single words.

Ambiguity in bilingual dictionaries is unavoidable as long as the entries are those of single words. Single words, in isolation, can mean many things; Weigand (2004: 14) maintains that 'isolated words are often considered to be polysemous'. Lexicographers endeavour to describe their meanings by assuming different senses but they fail to give precise instructions about how to determine the word sense in question. We normally find, for a given word in a bilingual dictionary, a set of possible translation equivalents depending on the senses assigned to it by the lexicographer. These equivalents are often presented without sufficient instruction as to which of them should be used in a given context. Thus, users end up being at a loss, not knowing how to choose among them. In order for bilingual lexicography to solve this problem, one has to accept the fact that single words are not necessarily units of meaning and it is pointless to look at the meaning of a single word in isolation to try to find its translation equivalent. As long as bilingual dictionaries are organised on the single word principle, they will not be sufficiently reliable to let the users select the correct translation equivalent.

A new approach to identify translation equivalents could solve this issue of ambiguity: the use of the parallel corpus—i.e., a corpus which 'consists of original and translated texts' (Danielsson and Mahlberg, 2003). This source overcomes the disadvantages of bilingual dictionaries. Parallel corpora take away the ambiguity of single words in isolation. In the parallel corpus, words are embedded in their contexts; this normally resolves any ambiguity. Sinclair (2004) points out that 'meaning is created, not over each single word, but over several words together'. As long as the context is taken into account, the issue of ambiguity will not arise. Thus, the parallel corpus is an excellent resource which enables us to find the proper translation equivalent in the target language for a source language expression.

3.1.2. The JENAAD

The JENAAD, the largest Japanese-English parallel corpus, will be used as the best analytical tool for the comparative and contrastive study of words denoting unhappiness in English and Japanese. It consists of about 5 million English words and 6 million Japanese morphemes, and covers broadsheet newspapers from 1989-2001 (Japanese articles and their English translations).

Two points must be stated about this resource: First, this corpus contains Japanese original articles and their English translations. There is always the argument that translations do not really mirror how the target language is, i.e. a translation is not necessarily an accurate representation of the target language. However, as Teubert (1996) mentions, translations in parallel corpora are the only resources that enable us to investigate lexical analysis across languages.

Secondly, this Japanese-English Parallel Newspaper corpus is a 'unidirectional' parallel corpus, i.e. a corpus which consists of translations 'in one direction only from language A to language B', not a 'bidirectional' parallel corpus, i.e. a corpus which consists of translations 'in both directions from language A to language B and from language B to language A' (Altenberg and Granger, 2002). However, I will use this parallel corpus for both directions. In order to validate the reversibility of parallel corpora for this study, I previously carried out experiments which showed that for the purpose of studying translation equivalence in parallel corpora (unlike bilingual dictionaries) it is not crucial whether their texts have been translated from one language to another or the other way round (Kondo, 2004).

3.1.3. Methodology

First, I will look at the English concordance lines of *unhappy* and their aligned Japanese concordance lines in the JENAAD in order to find out whether *unhappy* is translated into a single Japanese translation equivalent or not. If not, the next task is to examine how different the Japanese translation equivalents are by observing the contextual patterns. Here, my primary interests are to identify the syntactic patterns and semantic preference—defined by Sinclair (2000) 'the co-occurrence of words with semantic choices'—responsible for the selection of a certain translation for a given context.

Also, as looking at the English word *unhappy* and its Japanese translation equivalents is not enough to obtain a better understanding of how the concepts of unhappiness differ in English and Japanese, I will examine the Japanese translation equivalents and their English translation equivalents in terms of the contextual patterns (syntactic and semantic patterns) as well. In analysing *unhappy* with this corpus, I will use ParaConc (parallel text concordance software; Barlow, 2004) and employ a feature of the programme known as HotWords in order to retrieve the top-10 translation alternatives that occur most frequently in the translated concordance lines according to their association rates.

4. Investigation of the Japanese translation equivalents of *unhappy*

The word *unhappy* appears in the JENAAD 44 times; 20 of them were translated into a Japanese translation equivalent *fukouna* (adj) and 5 of them were translated into another Japanese translation equivalent *fuman* (n). Below are some citations.

> Such a term serves only to recall *unhappy* times in history when Christians suppressed Muslims.
>
> The committee said that citizens in both Japan and China were *unhappy* about the meeting.

Unhappy in the first example was translated into *fukouna* (adj); while *unhappy* in the second example was translated into *fuman* (n). What is the difference between those two? How did the translators make the choice between *fukouna* and *fuman*? In order to clarify this, I will look carefully at (1) the contextual patterns of the concordance lines in which *unhappy* was translated into *fukouna* and (2) the contextual patterns of the concordance lines in which *unhappy* was translated into *fuman*.

4.1. *Fukouna* (adj)

Unhappy in the following concordance lines were translated into *fukouna* in the JENAAD.

…s After the short, *unhappy* administrations of M..
…to help close this *unhappy* chapter of the regio…
… conclusion to the *unhappy* confrontation. In …

The dominant syntactic pattern is *unhappy*+N, which is used in 19 out of the 20 lines (95%). The nouns are *administrations*, *chapter*, and *confrontation*. All of the nouns are abstract nouns denoting a situation having a certain period.

In terms of semantic preferences of the concordance lines in which *unhappy* is translated into *fukouna*, there are three main features, shown as below.

> Given the growing sentiment in favor of a settlement, this opportunity to put an end to 40 years of *unhappy* confrontation should not be missed.
>
> Japan and Korea share an "*unhappy* past".

One repeated semantic preference is the co-occurrence with something denoting an end. This feature appears in 11 citations out of 20 (55%). Some of them are verbs (e.g. *close*) while the others are nouns (e.g. *conclusion* and *end*). Another semantic pattern is *unhappy* co-occurring with an expression denoting the past (e.g. *past* and *history*). This feature is seen in 7 lines out of the 20 (35%). Finally, the other feature is the co-occurrence with elements describing a period such as *40 years* and *earlier this century*. This tendency is exhibited in 7 lines out of the 20 (35%).

Thus we can see in which contexts *unhappy* is translated into *fukouna*. If *unhappy* appears (1) in *unhappy*+N (for situations lasting a certain period) and (2) with the expressions denoting an end, the past, or a period of time, it is likely to be translated into *fukouna*.

4.2. *Fuman* (n)

The second most frequent translation equivalent of *unhappy* in the JENAAD is *fuman* (n). *Fuman* has a completely different use from *fukouna*, as shown below. Here are some concordance lines in which *fuman* is used.

..industrial circles are *unhappy* because the conten...
..ricans are particularly *unhappy* when it comes to th..
..ighter. We are still *unhappy* with the draft,, whi..

These *unhappy* correspond to one dominant syntactic pattern, Person+BE+*unhappy*+Reason, occurring this way 5 times out of 5 (100%). Although the frequencies are not high, *unhappy* in this syntactic pattern is always translated into *fuman*, and never into *fukouna*. The expressions denoting 'Person' in the above are *financial and industrial circles*, *Americans*, and *we*, respectively. The expressions denoting 'Reason' are *because the contents of the expected package remain unclear*, *when it comes to the acid test of collective security*, and *with the draft*, respectively.

In terms of the semantic preferences of the concordance lines in which *unhappy* was translated into *fuman*, there are three dominant tendencies.

> The survey also found that a record 90 percent of pollees said they were *unhappy* with the current state of the nation's politics, when those …
>
> However, financial and industrial circles are *unhappy* because the contents of the expected package remain unclear.

First of all, unlike *fukouna* which is associated with the past, *fuman* co-occurs with something denoting an issue which is controversial in the present. This feature appears in 4 lines out of 5 (80%) in the JENAAD. The expressions for this preference, in the above examples, are *the current state of the nation's politics* and *the contents of the expected package* and. Each of them is an ongoing issue which has not yet come to an end.

Second, *unhappy* translated into *fuman* co-occurs with an expression denoting a group of people. This feature appears in 4 lines out of 5 (80%). In the above examples, the words for this feature are *90 percent of pollees* and *financial and industrial circles*, referring to a group of people, not individuals.

Finally, *fuman* co-occurs with an expression denoting a continuation. This feature appears in 2 lines out of 5 (40%). This semantic preference appears, as *remain*, shown above.

We now again understand in which context *unhappy* is translated into *fuman*. If *unhappy* appears with (1) Person +BE+*unhappy*+Reason and (2) expressions denoting an ongoing controversial issue, a group of people, and a continuation, it is likely to be translated into *fuman*.

Thus, my current analysis demonstrates that *fukouna* (adj) and *fuman* (n) have their own distinctive and complementary contextual information and meaning. Both syntactically and semantically *fukouna* and *fuman* are quite complementary. Each of them focuses on different aspects of what *unhappy* means. Japanese makes a clear distinction between these two emotion words, one is an 'unhappy' feeling referring to people or to situations with an end, the past, and a period (*fukouna*), and other 'unhappy' feeling refers to people with a certain reason, an ongoing issue, or a continuation (*fuman*), focusing

attention on distinctions for which English does not have names.

5. Detailed investigation of the English translation equivalents

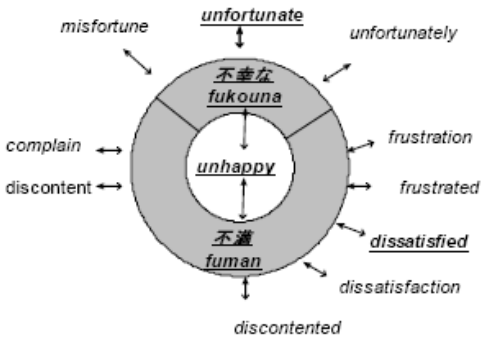


Figure 1: Translation equivalents of *unhappy*

Figure 1 shows the intricate network of the translation equivalents of *unhappy* in the parallel corpus, with the English word *unhappy* in the centre circle and its Japanese translation equivalents in the outer circle. The English equivalents are shown outside the circle and *unhappy* in the centre circle itself. In the previous section, I analysed the Japanese words equivalents to the English word *unhappy*, *fukouna* and *fuman*. In order to understand the diversity contained in the English word for unhappiness, I will now look into the English equivalents of these two Japanese words (shown outside of the circle in Figure 1). Since space does not permit a detailed discussion of all eleven English equivalents, I will focus on just three English equivalents underlined in Figure 1: *unfortunate*, *dissatisfied*, and *unhappy*.

5.1. *Unfortunate* (adj)

The dominant syntactic pattern of the concordance lines in which *fukouna* was translated into *unfortunate* is *unfortunate*+N, appearing 21 out of 27 lines (78%). There are two types of nouns: nouns denoting a situation with a certain period in the past (e.g., *history* and *past*) and nouns denoting results of something happening earlier (e.g., *departure* and *results*). Another dominant syntactic pattern is *it*/*this*+BE+*unfortunatee*+(*for*)+*that*/*to*/*if*, appearing 6 out of 27 lines (22%).

…will feel that the *unfortunate* past has been..
..disposal. It will be *unfortunate* for the Japan-U.S.
…and polices bring *unfortunate* results. This…

The English equivalent of *fukouna* (adj), *unfortunate*, has four semantic preferences: it is associated with expressions denoting (1) representatives of a country (41%), (2) the past (41%), (3) a period (33%), and (4) an effort (26%). The words reflecting these features in the following example are (1) *the Emperor*, (2) *past* and *history*, (3) *wartime* and *from 1931 to 1945*, and (4) *strive*, respectively.

At the same time, we must strive to resolve the issue of our *unfortunate* wartime past vis-a-vis Asian nations.
"At the outset of the conference at the Imperial Palace, the Emperor said, "In modern ages, there was an *unfortunate* history," referring to the war between the two countries from 1931 to 1945.

By comparing the syntactic and semantic features of *unfortunate* with those of *unhappy*, the underlying reasons why the translators chose either *unfortunate* or *unhappy* are identified. Both *unfortunate* and *unhappy* share a syntactic pattern, ～+N (N=situation with a certain period in past), and two semantic preferences: co-occurrence with something to denote the past, and a period. This means that both *unfortunate* and *unhappy* can be correctly used in these patterns as a translation of *fukouna*. On the other hand, both *unfortunate* and *unhappy* have exclusive patterns. The distinctive syntactic patterns of *unfortunate* are ～+N (N=results of something happening earlier) or *it*/*this*+BE+～+(*for*+N)+*that*/*to*/*if* and its semantic preferences co-occur with something which denotes an effort as well as a representative of a country. The distinctive semantic pattern of *unhappy* is its co-occurrence with an expression denoting an end. These idiosyncratic patterns indicate the right translation equivalent, either *unfortunate* or *unhappy*, when *fukouna* is translated.

5.2. *Dissatisfied* (adj)

The noticeable syntactic pattern of the concordance lines in which *fuman* is translated into *dissatisfied* is Person+BE+*dissatisfied*+Reason. It occurs in 26 out of the total 42 lines (62%) in the JENAAD.

…rity. Soldiers are *dissatisfied* because their sala…
..employees are also *dissatisfied* with a decrease in ..
..ent of pollees were *dissatisfied* with the new line…

All the expressions denoting 'Person' refer to social groups (e.g. *soldiers*, *employees*, and *middle class*), not individuals. The expressions implying 'Reason' appear with *why*, *at*, *because*, and *with*.

Dissatisfied, the translation equivalent of *fuman* (n), has three main semantic preferences: co-occurrence with the words denoting (1) an ongoing issue (86%) (2) some group of people (69%), (3) a survey (31%).

Results showed two out of three adults are *dissatisfied* with school education, the second worst outcome in the past 10 annual surveys on the same subject.

The expressions for these semantic features in the above examples are (1) *school education*, (2) *two out of three adults*, and (3) *surveys*, respectively.

In comparing the syntactic and semantic features of *dissatisfied* with those of *unhappy*, the reasons underlying the choice which translators make between *dissatisfied* or *unhappy* are identified. Both *dissatisfied* and *unhappy* share a syntactic pattern, Person+BE+～+Reason, and a semantic preference, i.e., the co-occurrence with something that denotes an ongoing issue. This overlapping distribution means that both *dissatisfied* and *unhappy* can be correctly used in these patterns as a translation of *fuman*. On the other hand, *dissatisfied* has an exclusive semantic preference: it co-occurs with an expression denoting a survey. If *fuman* appears in contexts of surveys, it is highly likely to be translated into *dissatisfied*, not *unhappy*; while *fuman* appears in a different kind of

situation, it is likely to be translated into *unhappy*, not *dissatisfied*.

6. Implications

Table 1 shows all the syntactic patterns and semantic preferences of both the English and Japanese words for 'unhappy': *unhappy*, *fukouna*, *fuman*, *unfortunate*, and *dissatisfied*. The + marks highlighted by a grey background mean that the word has the preference.

Fuman	*Fukouna*		*Unfortunate*	*Unhappy*	*Dissatisfied*
	+	~+N N=results of something happening earlier	+		
	+	*it/this*+BE+~+*(for*+N)+*that/to/if*	+		
	+	~+N N=situations with a certain period in past	+	+	
+		Person+BE+~+Reason		+	+
	+	an effort	+		
	+	representatives of a country	+		
	+	the past	+	+	
	+	a period	+	+	
	+	an end		+	
+		a continuation		+	
+		an ongoing issues		+	+
+		some group of people		+	+
+		a survey			+

Table 1: The concept of unhappiness

An important finding is that, as shown above, neither of the three English equivalents, *unhappy*, *unfortunate*, and *dissatisfied*, is identical to either of the two Japanese equivalents, *fukouna* and *fuman*. No pairs have completely the same distribution. This indicates that the English verbalise the feeling 'unhappy' in a different way from the Japanese.

For instance, one of the significant differences is concerned with two semantic preferences: end and continuation. Japanese makes a distinction between feeling 'unhappy' associated with end and continuation, giving different names to each of these feelings, *fukouna* and *fuman* respectively; while English does not make a clear distinction here. *Unhappy* is associated with both end and continuation. Similarly, another interesting difference is concerned with a different pair of semantic preferences: past and present, i.e., ongoing issues, as well. These findings indicate that even this very fundamental feeling, 'unhappy', does not mean the same to English and Japanese native speakers. Japanese speakers recognise feeling 'unhappy' in a different way from English speakers.

I have investigated the concept of unhappiness in English and Japanese in the JENAAD by looking at how the word *unhappy* is used to correspond to Japanese and how *fukouna* and *fuman* correspond to English. The results show that even a fundamental feeling, unhappiness, is culture-specific and verbalised differently in each culture. It must be stated that all of the findings of this study are only drawn from this very limited special corpus; in particular, the size and genre are very restricted. Although most of findings are supported by analysing data from the Bank of English with 25 randomly selected concordance lines of *unfortunate*, *unhappy*, and *dissatisfied* (Kondo 2004), larger parallel corpora focusing on fiction are needed to validate these preliminary findings.

This analysis clearly shows how corpus linguistics can contribute to the ongoing controversy regarding the status and origins of emotions. By looking at translation equivalents in a parallel corpus, the difference and similarities in concepts such as unhappiness between two languages can be revealed. This is a practical and necessary way of clarifying our view of how people conceptualise emotions.

7. References

Altenberg, B., Granger, S. (2002). Recent trends in cross-linguistic lexical studies. In B. Altenberg, S. Granger (Eds.), *Lexis in Contrast*. Amsterdam: John Benjamin.

Barlow, M. (2000). Parallel texts in language teaching. In S. Botley, T. MacEnery, A. Wilson (Eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, pp. 106-115.

Cambridge Advanced Learner's Dictionary (2003). First Edition. Cambridge: Cambridge University Press.

Collins Cobuild Advanced Learner's English Dictionary (2003). Fourth Edition. Glasgow: HarperCollins.

Danielsson, P., Mahlberg, M. (2003). There is more to knowing a language than knowing its words. *English for specific purposes world: Web-based Journal*.

Darwin, C. (1904). *The Expression of the Emotions in Man and Animals (Popular Edition): edited by Francis Darwin*. London: John Murray.

Ekman, P. (2004a). *Emotions Revealed: Understanding Faces and Feelings*. London: Phoenix.

Ekman, P. (2004b). Emotions Revealed: Recognising facial expressions. *StudentBMJ*, 12: pp. 141-142.

Evans, D. (2002). *Emotion: The Science of Sentiment*. New York: Oxford University Press.

Kondo, F (2004). Words denoting emotions. Unpublished manuscript.

Harkins, J., Wierzbicka, A. (Eds.) (2001). *Emotions in Crosslinguistic Perspective*. Berlin: Mouton de Gruyter.

Johansson, S. (1998). On the role of corpora in cross-linguistic research. In S. Johansson, S. Oksefjell (Eds.) *Corpora and cross-linguistic research: theory, method and case studies*. Amsterdam: Rodopi, pp. 3-24.

Salkie, R (2002). Two types of translation equivalence. In B. Altenberg, S. Granger (Eds.), *Lexis in Contrast*. Amsterdam: John Benjamin.

Sinclair, J. (2000). Lexical Grammar. Naujoji Metodologija. Retrieved December 10, 2004, from http://donelaitis.vdu.lt/publikacijos/sinclair.pdf.

Sinclair, J. (2004). In Praise of the dictionary. Unpublished manuscript.

Teubert, W. (1996). Comparable or Parallel Corpora?. In *International Journal of Lexicography*, 9: pp. 238-276.

Utiyama, M., Isahara, H. (2003). Reliable Measures for Aligning Japanese-English News Articles and sentences. *ACL-2003*: pp.72-79.

Weigand, E. (2004). *Emotion in Dialogic Interaction*. Amsterdam: John Benjamins.

Wierzbicka, A. (1999). *Emotions Across Languages and Cultures: Diversity and Universals*. Paris: Cambridge University Press.

8. Acknowledgements

I would like to thank Prof. Wolfgang Teubert for his expert knowledge and assistance. Also, I greatly appreciate the National Institute of Information and Communication Technology (Japan) and the University of Birmingham (UK) for permission to use and quote from their corpora.

Program Integrated Information ID Tag for Translation Verification Test

Naotaka Kato[1], Makoto Arisawa[2]

National Language Support, IBM Japan, Ltd.[1]
Faculty of Graduate School of Media and Governance, Keio University[2]
1623-14 Shimo-tsuruma, Yamato-shi, KANAGAWA, Japan[1]
5322 Endo, Fujisawa-shi, KANAGAWA, Japan[2]
katosan@jp.ibm.com[1], arith@sfc.keio.ac.jp[2]

Abstract

There are two types of translation for computer programs. One is for manuals and the other is for Program Integrated Information (PII). This paper focuses on PII translation. PII translation is substantially different from ordinary text translation. PII is separated from the programs themselves into externalized text resource files to allow for translation outside the program development laboratory. The contexts of the programs' operations are discarded. The translators have to translate phrases and words without context in the text resource files. The Translation Verification Test (TVT), which is done with the actual program operations, 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 of the file is unknown. We have developed a utility program to make it easy to find the source locations. The utility adds a short group of ID characters in front of every PII string. We used this systematic approach for CATIA (a CAD/CAM program from Dassault Systems) and found many advantages, such as locating hard-coded strings that are the biggest problem in program internationalization. This ID can be inserted independently of program development. This paper describes the approach in detail. In addition, this paper presents statistics about PII files. This important statistical information has not been considered in the program internationalization community.

Keyword: PII, translation verification test, string externalization, localization, internationalization, ID, TVT, GUI

1. Introduction

Program internationalization 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 organization that specializes in translation. If the text strings might need to be translated, the development laboratory externalizes the strings from the programs into PII files. The text resource file includes the keys and the isolated text strings. The programs have the keys and use their corresponding strings (Deitsch, 2001; IBM, 2004; Dr. International, 2003; Green, 2005). The current internationalllization process causes difficulties for the translators and the TVT testers. The translators have to translate short phrases without contexts. The TVT tester cannot find the source location of the externalized text found in a GUI message. This paper addresses this TVT problem.

The TVT testers and program development team members test the translations of the PII files. If there are errors in the translated strings, the testers need to fix them in each text file. Conventionally the testers have used a 'grep' function of the OS or editor program to find the source location in the PII files. The TVT testers face difficulties in finding the source location. For example, the TVT testers cannot identify the source location if identical strings appear with different keys. The goal of this research 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. The developed 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 ID is called the PII ID. For example, if the English string is 'Angular', the string displayed in the GUI might be '(E36.20)Angular' where the '(E36.20)' is the ID of the English string. We confirmed the effectiveness of the utility by actually using it for the Japanese TVT of the CATIA PII translation. We also discovered many useful features of the ID in this test. One of the results of using IDs is to reveal 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). It is intrinsically difficult for a TVT tester to know whether or not a string is hard-coded. Introduction of the ID reduced the time to find the string locations in the PII files from 30 hours to one or two hours during the CATIA TVTs. In Section 2, we describe the related research and the background of our research. In Sections 3 and 4, we present the details of our technique to address the TVT problems. In Section 5, we present statistics explaining why the approach works so well.

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 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 "Translation Manager" 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 Service Center) is an organization that specializes in translations, especially PII and manuals.

2. Background of the research

2.1. Related Research

The Mock Translator (Muhanna, 2003) and the IBM invention disclosure in Reference (IBM, 2003) are related research. The Mock Translator allows program developers

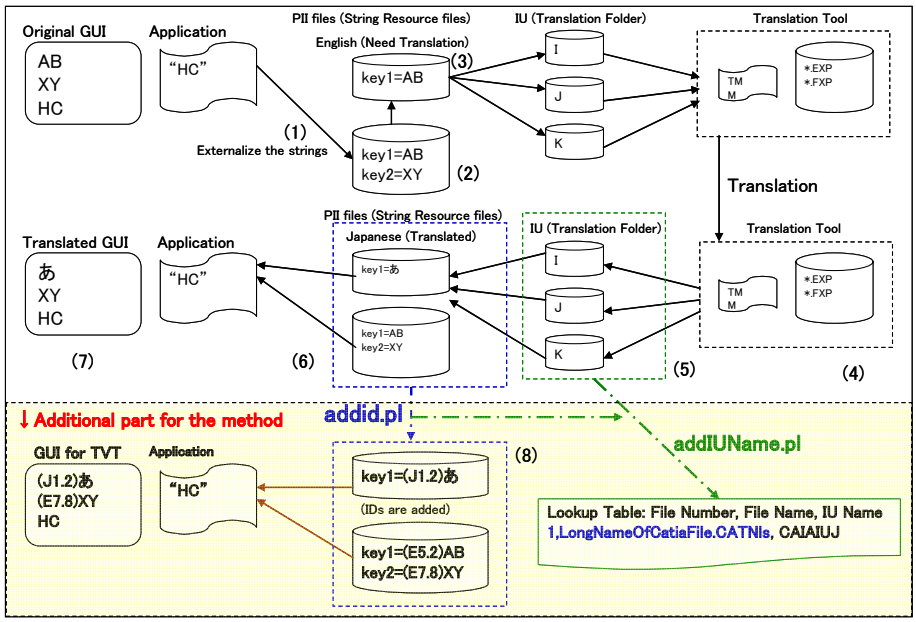


Figure 1. Translation flow of PII

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 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 a file name by itself is already too long for ordinary GUIs. The system in the disclosure assigns the file names 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. The systematic approach is an important point of our technique for benefiting from the PII ID tag.

In the linguistic research field, our research is related to the word sense disambiguation. However, there is no research about the word sense disambiguation for PII (Ide & Veronis, 1998).

2.2. The Flow of the PII Translation and Validation

Figure 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 Figure 1, the top part (Original GUI row), the middle part (Translated GUI row), and the bottom part (GUI for TVT row). The part under the dashed line is our new process. 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, AB, XY, and HC, in the program. Two of them, AB and XY, are externalized to PII files. Only the string AB is in the scope of the Japanese translation and the string XY is left as English (the original language). HC is a hard-coded string and cannot be translated.

(1) A development laboratory externalizes 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=AB
    key2=XY
(3) The development laboratory delivers externalized 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 English 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 a remote site.

If the TVT testers find inappropriate translations, they fix them on the system in Step (4) and repeat the Steps (5)-(6) and confirm the corrected strings in Step (7).

3. The problems that the TVT testers faced

The TVT testers faced difficulties in tracking the PII strings. When the TVT testers verify the translated PII strings according to the execution scenario of the tested program, 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.

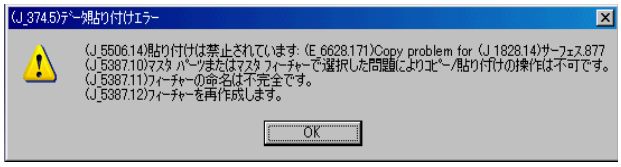


Figure 2. A GUI with the PII ID tag

Figure 3. The PII tag without prefix

Our ID tags focuses on and solves this problem faced by the TVT testers. Please refer to Figure 1 again. If a TVT tester finds that the translated 'あ' (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 XY 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 8,000 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. 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.

4. Our approach to solving the problems of TVT

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 externalized strings. The TVT testers can find these tags in the GUI. The part under the dashed line of Figure 1 shows our additional process to create the PII files for the TVT. We use the Perl program addid.pl 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. The mapping file is used by the addIUName.pl program and is used 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 マネージャー" respectively, then the generated files include "key1=(E25.22)Link Manager" and "(J25.22)リンク マネージャー", respectively. 'E' means English PII and 'J' means Japanese PII. The E and J are prefixes of the ID and are determined as argument strings of the addid.pl. The '25' in this example means the 25th file of the PII files. The '22' means that the key is located on 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 externalized strings. A TVT tester can easily find the source locations of the PII string 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 single text file that lists IDs, strings, file names, and keys for both languages. An example paragraph in the text file is shown below.

```
E5b6088, 36, "Curve Creation", CATStCLA.CATNls, SmartCurves.Title
J5b4891, 36, "曲線を作成"
```

There is a pair of these lines for each PII key. We also use the file in various ways for PII maintenance. If a string in the GUI does not have an ID, it means that the string was not externalized, but is a hard-coded string. Figure 2 and Figure 3 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 also applied our approach to a Java application and it worked well. In a Java application, we confirmed that we could easily switch PII files between PII with ID and PII without ID by utilizing the Java '-Duser' start option. The naming convention of the properties file was utilized to enable the function. We will discuss the details for a Java application in a future paper.

The following are the merits of our systematic approach:

- The ID prefix permits 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 externalized strings.
- The ID has important merits beyond replacing 'grep' searching. It can uniquely identify the source location of a string appearing in GUI whether or not the string is concatenated. Our ID approach can identify the hard-coded strings. English text (using Latin characters) appearing in a Japanese GUI can be identified as to whether or not it is a translated string or a non-translated string. The prefix architecture can clarify these differences.

5. PII Statistics

This section presents the statistical data about the PII files. We show data for CATIA PII and Microsoft Windows XP SP2 PII. The statistical data shows two facts. First, most PII strings are short. Consequently translators have to translate short phrases without context. Secondly, the strings of PII are repeated often in PII files. Therefore the TVT testers cannot uniquely locate the source text of the inappropriate translations found in the GUI.

5.1. The Data for CATIA PII

The left side of the Figure 4 shows the distribution of the number of words in each PII string. CATIA has about

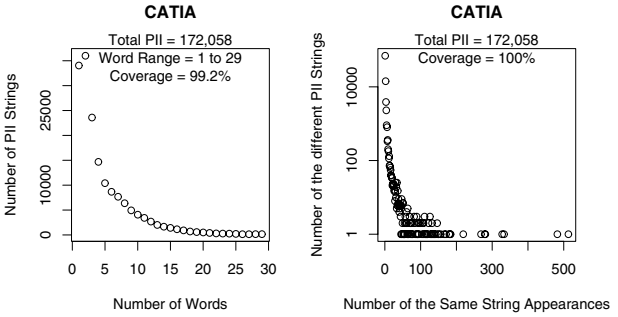


Figure 4. The statistical data for CATIA PII

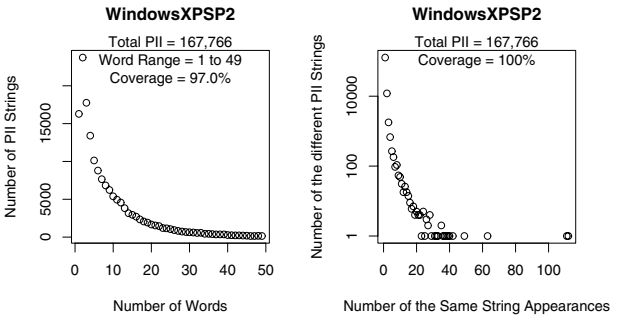


Figure 5. The statistical data for WindowsXP (SP2) PII

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 shows the number 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 two words.

The right side of Figure 4 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 vertical axis, the result is the total number of the PII strings with that number of appearances. For example, "Name" appears 333 times. Alternatively, the 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 (type) that appear three times and these strings are plotted at (3, 3882) in the right side of Figure 4. 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 the string in the other keys are not unique in the PII files. Therefore those other keys, about 100,000 keys, cannot be identified uniquely from a string in the GUI.

5.2. The Data for Windows XP (SP2) PII

The statistical results for Microsoft Windows XP(SP2) PII are shown in Figure 5. We see almost the same characteristics as for CATIA. We used the "Microsoft Glossary" data found on the Internet (Microsoft, 2005) and analyzed 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.

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 for PII translations. This paper focuses on the verification problem. The TVT testers could not identify the source locations of the PII strings when the TVT testers executed the test scenarios and found inappropriate translations. We systematically inserted a useful and compact ID in front of every PII string for all of the PII files, independently of 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. Lastly, we showed statistical information about PII. This information has not been clearly recognized by the program internationalization community.

7. References

Deitsch, A. and Czarnecki, D. (2001). *Java Internationalization.* O'Reilly.

Dr. International. (2003). *Developing International Software 2nd ed.* Microsoft Press.

Green, D. (2005). *Trail: Internationalization.* 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 Internationalized Products. *National Language Design Guide Volume 1, Fifth Edition*.

Ide, N. and Veronis, J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics.* 24(1), pp. 1-40.

Kehn, D. (2002). *Testing your internationalized 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*.

Acknowledgements

We would like to thank Tetsuya Nasukawa for helpful discussions and thank Shannon Jacobs for correcting our English.

Standards for Content Creation and Globalization

Sue Ellen Wright

Institute for Applied Linguistics
Kent State University, Kent, Ohio, USA
sellenwright@gmail.com

Abstract

Language industry experts have established a Wiki-based forum to facilitate the discussion of language-related standards across the boundaries of specific communities of practice and of the multiple bodies developing these standards. Kent State University and Mitre have created a standards portal dedicated to providing a comprehensive view of the many-faceted aspects of the field. The purpose of these efforts is to facilitate communication among active parties, to promote cooperation, and to reduce the potential for duplicate effort. Open discussion should also enable the participation of interested individuals who are not members of standardizing bodies. This paper introduces the structure of the current information portal and briefly outlines the variety of thematic areas covered by the standards with reference to the groups involved in elaborating them. The paper concludes by proposing a broader categorization of the language-related standards by incorporating knowledge organization standards into the framework of a revised standards portal.

1. Creating a Standards Portal

1.1. Creating a Global View

Activities involving technical writing, translation, and content management in publication, both in hardcopy and on the World Wide Web, have developed into a wide-ranging industry, while at the same time the number of standards governing these activities has burgeoned. This trend has been accompanied by a proliferation of the various standards, professional, and commercially oriented bodies involved in writing language-related standards. At a "summit"-style conference held in Berlin in December of 2005, a group of standards experts met together with users of standards from the language industry with the goal of examining the efficiency, interoperability, and quality of the wide range of normative efforts addressing issues in the industry. The goal of the conference was both to inform the participants of evolving trends and projects and to examine whether these standards are meeting the needs of software developers, publishers, government agencies, and potential users in other communities. These discussions generated a call for two major initiatives: the creation of a standards portal and the establishment of a permanent Web-based forum for discussions among different communities of practice across the spectrum of organizations and users involved in developing and implementing the standards.

Interest in a permanent discussion forum has currently taken the form a Wiki for Standards. The forum page is presently little more than a shell, but the full Wiki functionality is in place for interested parties to begin posting discussion items to the site (Wiki for Standards 2006).

The notion of a standards portal already has a sound beginning. The author has been tracking the development of these standards for the last five to ten years. A major contribution to this effort is already available online (Mitre 2006). Readers are urged to turn to this webpage in order to find specific names of standards and references to standards bodies. The length of this short paper precludes a detailed listing of the names and designations for individual standards, many of which are discussed and fully documented in further detail in Wright 2006, *forthcoming*, which is a contribution to K. Dunne's *Issues in Localization.* The longer article includes relevant history behind some of the standards, discussions of the relative merits of various conflicting standards, and projected future directions. A comprehensive Appendix lists detailed information about standards relevant to the localization industry, but some of the broader, non-localization-related items are not included.

2. Categorizing the Standards

The complexity of the Mitre Web page as a resource is reflected in the categorization of the types of standards listed in the index list that introduces the webpage:

2.1. Base Standards

Markup Languages (SGML, XML, HTML, etc.): Markup languages provide the basis for the creation and representation of content in both print and Web environments. The evolution of XML (Extensible Markup Language) also offers new modalities for recording and manipulating all manner of data, foremost on the Web, but also in a variety of other interactive modes.

Metadata Resources: Metadata (classically defined as data about data) standards provide clear methodologies for harmonizing data elements (data field names and defined content values) for use in diverse data processing environments. ISO/IEC Joint Technical Committee 1, Subcommittee 32, which is responsible for the metadata standards, is currently in the process of revamping the core standards that govern the development of MDRs. Terminological practice has informed the creation of effective, concept-oriented metadata registries (MDRs) and metadata practice has facilitated the work of terminologists and other language specialists. The metadata and terminology management communities maintain close contact, but as is so often the case where diverse groups converge, there remain issues of discussion between the communities based on differing disciplinary approaches and theoretical views.

Character Encoding: The Unicode standard offers a uniform character encoding scheme designed to replace the plethora of frequently platform-dependent, and to a certain extent, region-specific, encoding methodologies, which have over recent decades prevented easy interchangeability of information in the many scripts and languages of the world. The standard is designed to enable text processing in non-European scripts, in bidirectional language environments, using syllabary and logographic languages, and particularly in situations requiring multiple scripts. To date, full implementation of the standard is by no means complete, and the challenges remaining at all levels of data and text processing are significant, but progress is being made more rapidly than anticipated.

Access Protocols and Interoperability: These standards are basic Internet-related protocols designed for data handling in Web environments. They are listed in the Web portal for sake of completeness because they are essential for content management, especially for the implementation of Websites related to e-commerce and other interactive pages involving information management. They are not, however, generally elaborated by language professionals. Nevertheless, they must be aware of the consequences of these standards, and the standards themselves need to account for special issues involving multilingual and multicultural issues.

2.2. Content Creation, Manipulation, and Maintenance

Authoring Standards: Authoring standards take the form of style and terminology guides and tend to be proprietary and enterprise specific. In some state-of-the-art environments, they may be linked to controlled languages and automatic style and usage checkers. Nevertheless, some special areas (aeronautical design and software development, for instance) have fielded specific industry-wide standards and criteria governing content creation within specific limited disciplines.

Text and Content Markup [in Localization Environments]: Beyond the fundamentals of text markup covered by the base standards, the localization industry has undertaken to create a variety of XML-based standards designed to facilitate the processing of localizable materials throughout a sometimes complex series of steps and despite the interaction of multiple individuals and different localization service providers. The goal of these standards is a seamless workflow throughout the course of complex projects involving the manipulation of multilingual content.

Translation Standards: Aside from the relation of translation per se to other standards activities involving the localization industry, a series of translation-oriented topics present themselves. *Translation process* has been addressed in a series of national and regional standards, the most recent of which are currently in the finalization stages in Europe and the United States (Melby 2005, Arevalillo Doval 2005, DePalma 2006). *Translation quality*, particularly in the functional sense as defined in client/vendor specifications, is addressed in these standards. *Vendor certification* is a component of the new European standard, while it has been avoided in its American counterpart. *Translator competency* tends to be the subject of national and regional certification schemes, some of them dictated by law and/or administered by professional bodies (Stejskal 2004). *Translation metrics and measures* have been addressed both in standards and as a function of the certification process.

As national and regional norms proliferate, the call for uniform international standards grows. At the Berlin summit, the demand for ISO standards was both vocal and persistent. The need for activity at the international level must nonetheless be carefully balanced against the evolution and maturity of young standards that need to establish themselves at local levels. One strong facet of the standards discussion in this area called for open discussions throughout the language community, together with dissatisfaction with standards that are developed in closed environments. Given the amount of time that it takes to elaborate standards, there are many who feel that it would be productive to begin work on translation-related vocabulary and procedural standards soon.

2.3. Terminology and Lexicography Standards

*ISO TC 37 Standards***:** ISO TC 37 (Terminology and language and content resources) is responsible for a variety of standards that in themselves cut across the categories defined here. Initially dedicated to the creation of standards for terminology activities within ISO committees, TC 37 focused on the general principles of terminology management, the layout of terminological entries, and computer applications for terminology management. In recent years the scope of the committee has expanded considerably. A complete listing of current activities is impossible here, but the scope of the group includes language codes (see *locales* below), standards related to natural language processing, lexicography, and liaisons with a variety of communities of practice, such as the metadata community, localization environments, developers of ontologies and taxonomies, and the Semantic Web community. In the context of this workshop, it should also be noted that some core terminology standards lend themselves well to the pedagogy of translation and terminology studies.

Technical Interchange Standards: These technical standards enable the exchangeability and interoperability of data stored in resources modelled according to terminological, lexicographical, and machine-readable lexical principles, in addition to the exchange of lexical and terminological data included in machine translation lexicons. Besides these lexis-based interchange standards, text-based exchange standards facilitate the interchange of translation memories and, as noted above, of procedural text and content markup in the localization framework. There is also considerable interest at this juncture to create crosswalk modalities to ontology standards

Controlled Language Standards: Like authoring standards, with which they are closely related, controlled language standards tend to be proprietary or enterprise specific, with a few exceptions cited in the Web page.

2.4. Ontology and Knowledge Ordering Standards

The evolution of semantic concept-oriented standards has resulted in the evolution of the OWL (Web Ontology Language) and SKOS (Simple Knowledge Organization Standard) projects under the auspices of the World Wide Web Consortium. These standards are further augmented by the continuing development of standards and best practices for the development of concept systems within terminology management projects as reflected in the ongoing work of ISO TC 37. As implied above, the creation of linkages between ontologies, knowledge ordering systems, and terminological concept systems constitutes an intriguing field of exploration that promises to enrich both the scope of terminological resources as well as the semantic content of ontological data collections.

2.5. Corpus Management Standards

Various European projects (e.g., Eagles/Isle), the Text Encoding Initiative (TEI), as well as working groups of TC 37, Subcommittee 4 have been responsible for creating a number of corpus management standards for marking up corpora, e.g., with respect to feature structures and morpho-syntactic content. As the Semantic Web grows, these methodologies hold the promise of more intelligent access to a wide range of marked up documents, which it is hoped will contribute to the automatic manipulation of content, particularly in online environments.

2.6. Locale-related Standards

Locale identifiers are expressed with a combination of language and country identifiers based primarily on ISO two and three-letter language and country codes. In computing environments they express far more than just regional language preference, however, in that they are used to specify a wide range of regional and national conventions, such as currencies, decimal and date conventions, script handling procedures, etc. The core codes upon which these computing locales are based have evolved over time and in a variety of communities of practice.

ISO TCs 37 and 46 (Information and Documentation), the Internet Assigned Numbers Authority (IANA), the Internet Engineering Task Force (IETF), and the Unicode Consortium converge in multiple ways, along with a cast of other players (see Wright 2006) to elaborate the base standards and to specify formats and usage for locale identifiers. The uppermost level of Figure 2 attempts to represent the overlapping responsibilities of the various normalizing bodies in the generation, maintenance, and use of the various identifiers and codes.

Furthermore, ISO TC 37/SC 2 is currently moving toward the approval of a broader set of 3-letter codes developed by the Ethnologue initiative (SIL 2006) to accommodate nearly 7,000. In addition to expanding the number of languages listed in the language code, the ISO group is also working to create a listing of world dialects (an even more daunting project) and of regional groupings that have been adopted in certain communities of practice (e.g., es-LA as a code representing Latin American Spanish).

2.7. Standards Organizations

It is not easy to neatly assign the different types of standards to specific standards organizations because the different groups crisscross the field, sometimes with several groups working on the same topics, or individual groups addressing a variety of topics. Primary players include (See Wright 2006 for expansion of acronyms):

ISO: The International Organization for Standardization, primarily Technical Committee 37, but also TC 46
IEC: International Electrotechnical Commission, with programming-related standards
ISO/IEC Joint Technical Committee, JTC 1: Markup languages, metadata
National mirror bodies such as *ANSI, DIN, ÖNORM, AFNOR, BSI,* etc.
US specialised groups under the ANSI umbrella, e.g., *ASTM, SAE*: Translation quality
ANSI-based Technical Committees such as *NISO* (ANSI Z39): Thesaurus and knowledge organization
IETF, IANA, and the *World Wide Web Consortium (W3C)* Standards governing the Internet and the World Wide Web
Unicode Consortium: Character encoding and related matters, locale identifiers. Locale Markup Language
Localisation Industry Standards Association (LISA): Localization-related standards, such as Translation Memory Exchange (TMX) and Termbase Exchange (TBX), along with a set of standards related to localization quality issues
Organization for the Advancement of Structured Information Standards (OASIS): XLIFF 1.1 Specification, XML Localization Interchange File Format, and DITA, the Darwin Information Typing Architecture
Quasi-standards, such as the *American Translators Association (ATA)*: Standard Framework for Error Marking, which being used in pedagogical environments as well as for certification testing
TEI: Text encoding and corpus-related formalisms
EAGLES/ISLE: Human language technologies, primarily natural language processing

3. Presenting a Multifaceted View

The current portal (Mitre 2006) reflects a focus on standards for GILT[1]-oriented activities, particularly translation and localization, with detailed information on the standards of ISO TC 37 and related formal standards. As noted above, two areas that have experienced significant development in recent years and that are not treated in adequate detail in the existing portal are Knowledge Ordering Schemes and content management. Figure 1 covers the breadth of existing and evolving standards efforts in terms of specialized subject fields, while Figure 2 provides a view of the same efforts seen from the perspective of the various standardizing bodies. More detailed views are necessary to see the full linkage between topical and group activities.

[1] Globalization, Internationalization, Localization, Translation

This revised representation starts with the most basic standards, with *Text Representation* (Unicode), *Markup Standards*, *Metadata*, and *Identifiers* of various sorts treated first at the top of the chart. Section 2.6 has already dealt with the complex interplay of normalizing groups involved with the identifiers.

3.1. Knowledge Ordering Schemes

Figure 1 moves on to knowledge ordering schemes (KOS, discussed briefly in 2.4 above) and controlled vocabularies, which have also seen new standards emerge in the last year. Together with terminological concept systems and taxonomies, these resources provide structural frameworks for organizing different kinds of semantic content based on related, but non-identical methodological approaches in different communities of practice.

Controlled vocabularies: Essentially controlled vocabularies are used to describe documents and other "content objects" that are typically maintained in collections, such as libraries and museums, but most-importantly from the standpoint of modern information retrieval, from digital libraries and other collections as well. Controlled vocabularies include, as shown in Figure 2, a variety of resource types, all designed as systems for identifying and reusing information.

Most obviously, authority files designed for libraries and similar repositories are used to retrieve objects from documented collections as needed. More and more, however, the same or similar tools can also be used to locate and recover individual pieces of information embedded in document objects in machine-readable contexts. While printed indexes began as tools for accessing information residing in physical books, the evolution of automated indexing systems used for mining information from unmarked texts has fuelled the conflagration of search engines that has swept across the landscape of the Worldwide Web. The huge difference between traditional indexes and automatic indexing and retrieval is that traditional methodologies involve the careful examination and manipulation of known documents and the planning of strategies for re-accessing clearly identified information as needed in the future. As an example, a book documented and identified in a standard system such as the Library of Congress or the Universal Decimal System can be relocated by referencing the resource that documents and traces this particular object in any given collection. In contrast, systems for automated retrieval from *unmarked* collections require the ability to envision potential types of information or even strategies for recognizing entirely novel elements of knowledge in order to locate currently unpredictable, as-yet unknown information.

Uncontrolled vocabularies: Controlled vocabularies require the specification of standardized word forms that *shall* be used to identify and relocate objects and information. They create a relatively predictable semantic environment designed to reduce the potential for uncertainty and "noise" that exists in natural state linguistic. Terminological concept systems document semantic relations in actual, uncontrolled language, which is potentially much less predictable, even in cases involving standardized terminology. Thesauri and classification systems tend to be selective in the documentation of concepts, mapping subordinate concepts to superordinate concepts or less interesting related concepts to broader categories for retrieval purposes. In contrast, terminologists creating concept systems generally try to fill in all levels and to accommodate all notions related to a system. Here the purpose of the system is to create semantic maps of concept fields for the purpose of representation and not just for retrieval. (It should be noted that both controlled vocabularies and concept systems can be multifaceted and multidimensional, but that discussion goes beyond the limits of this short paper.) By the same token, terminological concept systems, especially ones that involve so-called "ad hoc" terminologies for commercial environments, are frequently text-driven and do not necessarily provide a full view of conceptual and knowledge networks.

Lists: Lists are sometimes included in the catalogue of objects covered by controlled vocabularies. Of course, not all lists are *controlled* in the sense described here. Nevertheless, there are important lists that, if properly managed, can provide a critical dimension to the knowledge management task; these list-like resources include terminologies, lexicographical dictionaries, and machine-translation lexicons, as well as both structured and traditional gazetteers. It remains to be seen to what extent these resources can be utilized in integrated knowledge organization environments, and to what extent the vast array of they will in future be made available for such integration, but the notion of creating linkages between ontological resources on the one hand and lex-term resources on the other offers an intriguing possibility for future information management and retrieval.

Cross-walks: Despite differences in approach and methodology, as reflected in the motivation and form of the standards created for the various communities of practice, the conceptual hierarchies and associative cross-references included in the different types of resources (taxonomies, ontologies, concept systems, controlled vocabularies, and topic maps) provide critical semantic orientation across the various planes that make up multidimensional semantic space. As such they all provide valuable information relevant to the disambiguation of conceptual reference, the application of consistent vocabulary in the creation of texts, and the retrieval of information from a variety of knowledge resources. As implied in the previous paragraph, coordination and integration of these assets with comprehensive lexicographical and terminological collections would enable further leveraging of available information. The critical lynch pins in such a linkage include the LISA TBX (Termbase eXchange) and the Lexical Markup Framework standards noted in Figure 2, as well as possibly the OLIF 2 standard. Missing elements include a future LBX (Lex-Base eXchange) format to interact with TBX, as well as a cross-walk between the terminology world of terminological concept systems and the thesaurus and controlled vocabulary-oriented environment reflected in the W3C's SKOS standard.

3.2. Communities of Practice

This short paper has attempted to provide a brief guided tour through the most critical segments making up the language standards maze. The concept maps shown in Figures 1 and 2 more clearly represent the diversity and scope of these endeavours than has been possible in this short summary. Unfortunately, limiting the view to two major overviews does not do justice to the interlinking and yet sometimes contradictory relationships that exist between the thematic elements represented in Figure 1 and the overlapping activities of the diverse groups documented in Figure 2. The above discussion has highlighted two major examples of interlacing objectives, styles, and disciplinary needs, namely involving language identifiers and knowledge organization systems. Even in relatively young areas of endeavour such as the Internet or metadata environments, differences in perspective and disciplinary orientation result in variations in approach and knowledge representation.

Furthermore, the diachronic evolution of standards in different areas sometimes introduces discordant elements into efforts at harmonization in that one group will base its work on the status of another group's work at a given point in time, only to discover later that this "standard" has evolved in the meantime to take on a different configuration. At times groups start work on a particular topic without being aware that it is already being addressed in another venue. In addition, some groups restrict membership, thus barring even qualified individuals from participation. The second-wave publication and sales mentality of some main-line standards organizations continues to make it difficult for a wide range of potential implementers to acquire the standards (BSI's new thesaurus standard, for instance, costs $330. or €275.) It is difficult to visualize how even the best standards can achieve wide acceptance under such circumstances. The advantage of ongoing close communication through a Wiki would offer opportunities for broader participation and possibly enhance adoption of critical standards.

4. Outlook

Trends in the industry would indicate that the variety of standards needed and proposed across the wide spectrum of the language industry is not likely to diminish in the future, nor is the potential for divergences and incompatible developments likely to disappear. At the same time, the great advantage of collaborative evolution in these areas is paramount in light of the drive to create interactive, interoperative knowledge management systems on the World Wide Web, in enterprise-oriented private data environments, and in content production venues, such as the localization industry.

The purpose behind the proposed Wiki-based discussion forum is to establish a venue within which the various interest groups can come together to share information and to coordinate efforts. Expansion and updating of the Mitre Web portal has the goal of maintaining current information on standards activities, although this task is not an easy one, given the speed with which new projects evolve, old ones transform, and standards even move around from group to group or merge with other efforts. Given the fact that conflicting standards are sometimes created, especially in different regions and discipline-specific situations, broad-based opinion supports establishing uniform, international standards in support of global information exchange and management. Most importantly, the two projects offer a window on the standards activities from the outside looking in by virtue of the information portal, as well as a two-way conversation by virtue of the Wiki between the potential implementers of the standards and the standards bodies.

5. References

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252-262.

Arevalillo Doval, J. J. (2004/05). Focus on Standards (English): The EN-15038 European Quality Standard for Translation Services: What's Behind It? In *The Globalization Insider*. Geneva: LISA, http://www.lisa.org/globalizationinsider/.

Baer, B. and Koby, G.S. (2003). Adapting the ATA framework for standard error marking for translator training. Paper read at the 44th Annual Conference of the American Translators Association, Phoenix, AZ.

Berlin, (2005). Language Standards for Global Business: http://www.internationalization-conference.org/languagestandards/index.html

Cover, R. *Cover Pages. Hosted by OASIS.* Misc. standards-related updates. http://xml.coverpages.org/

DePalma, D. (March 2006) European Standards Body Approves Translation Quality Specification. In *Common Sense Advisor: Global Watchtower™*. http://www.commonsenseadvisory.com/news/global_watchtower.php

Melby, Alan K. (2005). Focus on Standards: Quality from the Ground Up. In *The Globalization Insider*. Geneva: LISA, http://www.lisa.org/globalizationinsider/.

Mitre. (2006). *Standards for Content Creation and Globalization*. http://flrc.mitre.org/References/Standards/index.pl

Standards-Wiki. (2006). http://www.wikiforstandards.org/en/index.php/Main_Page

Stejskal, Jiri. (2004). *International Certification Study*. Alexandria, Virginia: American Translators Association.

Wright, Sue Ellen. (2006). "The Creation and Application of Language Industry Standards." In: Dunne, Keiren, ed., *Issues in Localization*. Amsterdam and Philadelphia, John Benjamins Publishing Company,. (This book also includes a detailed Appendix, listing the major standards organizations together with their most prominent language-related standards.)

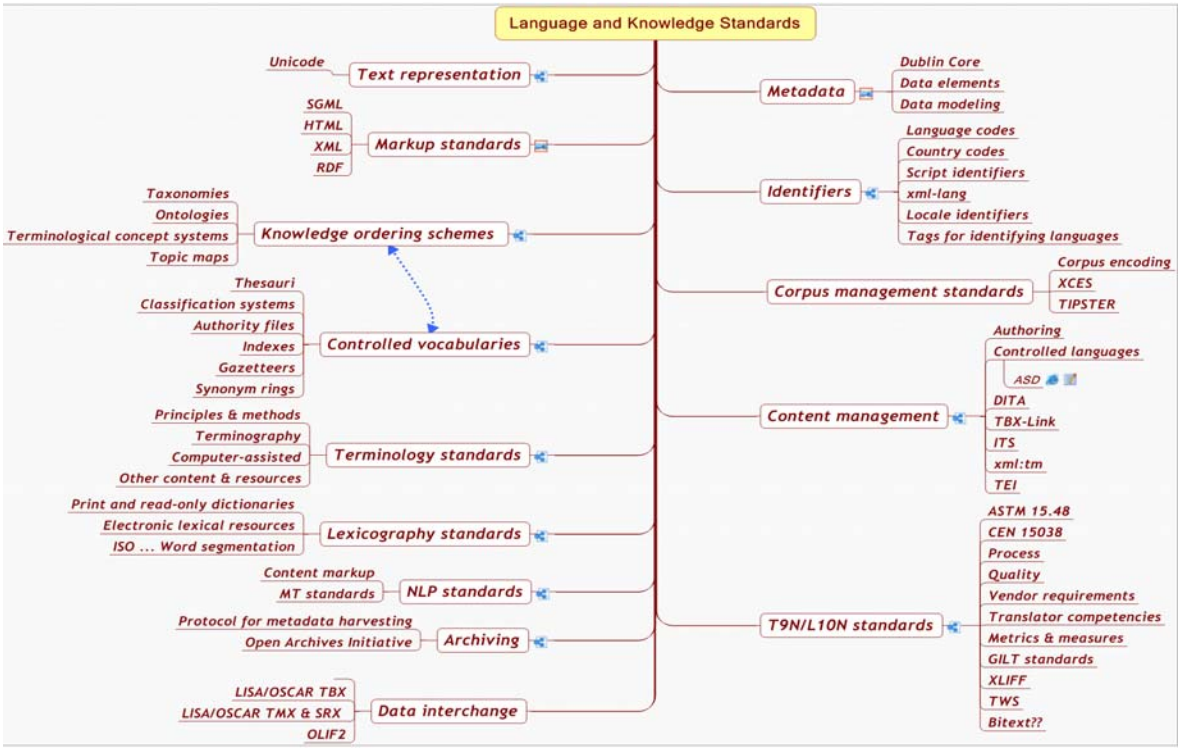


Figure 1: Language standards organized in thematic groups

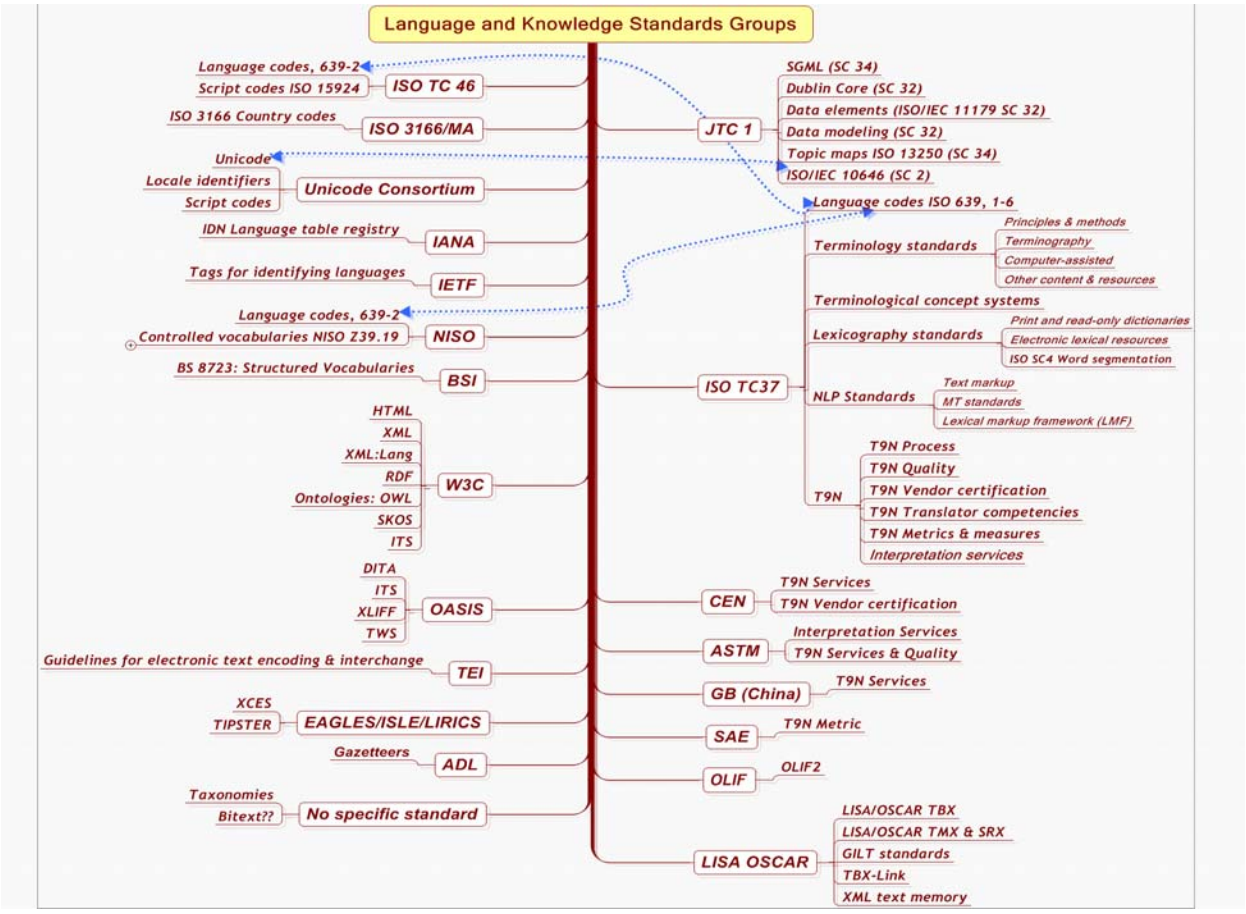


Figure 2: Language standards organized by standards groups

Corpora for translator education and translation practice
Achievements and challenges

Silvia Bernardini

School for Translators and Intepreters
University of Bologna at Forlì, Italy
Corso della Repubblica 136
47100 Forlì, Italy
silvia@sslmit.unibo.it

Abstract

This paper aims to chart some of the ground we have covered in the last decade or so in the area at the interface between corpus linguistics and translator training/translation practice, and to point to some of the challenges (and prospects) lying ahead. Two related issues of central importance for the translation professionals of tomorrow will be focussed upon: the current impact (or lack thereof) of corpus-informed pedagogy on the training of translators, and the increasing availability of tools that facilitate the construction of corpora from the web. The further spread of corpus resources in the translation profession is suggested to crucially depend on two main developments taking place: a greater focus on awareness-raising uses of corpora in translator education, and a greater ease of access to and greater integration of corpus tools with CAT technology.

1. Corpora in the translation classroom

1.1. Achievements

Translation is in many senses an ideal field for corpus applications. The analysis of source texts against specialised and reference corpora can make the identification of stylistic traits, idiosyncrasies and *register*- and *genre*-specific conventions (Trosborg, 1997) easier. 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, p. 131). Table 1 shows a concrete example of the kinds of insights one can gain in this way. Given a turn of phrase typical of the wine tasting domain in Italian (*il vanigliato del legno*), a translator with a specialised corpus for the target language at her disposal can extract and evaluate several likely translation candidates. In this case, the results of a simple search for *vanilla* 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, among others, *vanilla notes*, *nuances*, or *hints*.

Original Italian	avere il sopravvento sul *vanigliato del legno*
Original English	...*Vanilla and oak layers*... ...*vanilla and subtle oak undertones*... ...*vanilla characteristics* especially if oak-aged... ...*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*...

Table 1: Snippets from a search for *vanilla* in a web-derived bilingual comparable corpus on wine tasting

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., in corpus parlance, 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 to 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 practical and accessible introductions to (aspects of) corpus use aimed at students and professionals have appeared. Bowker and Pearson (2002) 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).

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, there are also signs that substantial efforts still have to be put into place to convince the majority of translation students and teachers that corpus use can help to reflect on tasks and raise awareness of strategies, and that these are among the central goals translation courses should set themselves. Secondly, as we shall see, professionals still appear to be largely unaware of or unacquainted with corpora. Clearly, a second challenge for translator educators is to reach them as well. Section 1.2. discusses these issues.

1.2. Challenges

1.2.1. 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 expected to be acquirable through practice with the kinds of tools and tasks one will be faced with during one's future professional carreer, 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, pp. 32ff) 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 I am obviously sympathetic to the general issue of technology in the translation classroom, I think 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 (something that even Gouadec would probably not want to suggest) 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 professionals and non-professionals. 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 little flexible, such that there is the danger, pointed out e.g. by Wills (1994, p. 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 professionality.

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, p. 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:

1. Get the information and knowledge required
2. Find the terminology
3. Find the phraseology
4. Translate
5. Proofread
6. Rewrite
7. Manage their task(s)
8. Manage a project (and other people)

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, and they all involve wordplay. In these cases, to "fully understand the material to be translated" one needs to understand the relationship between the facts being recounted or places being described and the lexicalised expressions used. While the texts themselves no doubt provide hints for getting at the more "congruent" sense, the less congruent sense is normally not as easily inferrable from the texts, since it is assumed to be available to the reader (this is in fact a precondition for the success of the wordplay). A student who is not aware of these layers

Title	Topic	Senses
Get into the habit	Montserrat Monastery	*in the habit of doing* something: having a habit [...] of so doing. So *to [...] get into the habit* (OED) *the habit*, monastic order or profession (OED)
Getting high	*Castells* (human towers)	*high*: under the influence of drugs or alcohol (OED)
Death on the mountain	Montjuïc (site of executions)	James Still poem Japanese movie
On the tiles	The work of Architect J.M. Jujol	*on the tiles*: on a spree, on a debauch (OED) *Josep Maria Jujol*: Catalan architect, his activity ranged from furniture designs and painting, to architecture (wikipedia)

Table 2: Titles and senses: wordplay in the *Barcelona Time Out Guide*

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, I am arguing that 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.

1.2.2. 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. This impression 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.

A more thorough investigation of the perception professional translators have of corpora is being conducted in the framework of the LEONARDO-funded MeLLANGE project, as part of an attempt to define user needs for learning materials on translation technology.[1] In the first round of submissions 623 questionnaires were returned, 90.8% of which completed by professional translators from the UK (the majority), France, Germany and Italy, and 9.2 by students of translation in the same countries. Out of the total respondents, 40.5% reported collecting reference materials, and more than half of them specified that they collect texts in eletronic format (69.4% of those who reported collecting materials). 46.9% read these collections of texts (rather than *searching through* them), and, of those who do search through them, a majority use search facilities in word processors (65.9%), with only a minority using a concordancer (19%, recall that data are for professionals *and* students).

While many translators are not acquainted with corpora, there seems to be widespread interest in learning more about them: 78.6% of respondents would be interested in a service which provides domain specific corpora, 77.9% in a tool for extracting terms from corpora, and 82.4% in learning more about their potential (MeLLANGE, 2005) (results are summarised in Table 3). Thus, there is clearly 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 (but not uncritical) perspective.

2. Building corpora

2.1. Achievements

Bowker (2004) mentions different 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 their consultation require minimal effort. Similarly, the fact that a large majority of the questionnaire respondents (above) reported consulting the Web through *Google* (93.4%), despite several drawbacks (that most of them are aware of), suggests that, for corpora to be successful with translation professionals, their construction and use has to be made substantially easier and faster.

One of the achievements of the past decade has certainly been the creation of tools that facilitate the extraction of textual information from the World Wide Web in ways that are more amenable to linguistic analysis. 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 Fletcher's *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, Matthias Hüning's *TextStat* concordancer[2], 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 and searching texts.

While *KWiCFinder* is designed mainly with language learning applications in mind (searching for a given word

[1] `http://mellange.upf.es/`

[2] `http://www.niederlandistik.fu-berlin.de/textstat/software-en.html`

Do you collect domain specific texts?	59.5% No 40.5% Yes
How do you collect them?	69.4% In electronic form 30.6% On paper
How do you use them?	53.1% Search through with software 46.9% Read them
Do you use corpora in your translation practice?	60.2% No 39.8% Yes
If yes, do you use?	26.1% Corpora of the target language 23.1% Corpora of the source language 19.7% Parallel corpora 15.3% Domain specific corpora 13.6% Comparable corpora 2.3% General language corpora
What do you use to search them?	65.9% Search facility in word processor 19.0% Concordancer 14.4% Other search tools (specify: Trados, Concordance in translation memory) 0.7% UNIX utilities
If you do not use corpora, why?	41.9% Never heard about them 19.9% I don't have time to build them 17.8% I don't know how to use a concordancer 8.7% I can't see any advantage over *Google* 6.8% I can't see any advantage over translation memories 5.0% 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?	78.6% Yes 21.4% No
Would you be interested in a tool for extracting terms from a domain-specific corpus?	77.9% Yes 22.1% No
Would you be interested in learning more about the potential that corpora offer?	82.4% Yes 17.6% No

Table 3: Corpus section of MeLLANGE questionnaire (first round, closed questions)

or expression as one would search the Internet), and *TextStat* only offers basic web-search facilities (i.e. it does not interact with a search engine, but simply spiders a specified URL), the *BootCaT* toolkit[3] (Baroni and 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 *Google*, 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, e.g. 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 Internet corpora for several languages (see Sharoff's website[4] and Ueyama (forthcoming)).

The results in Table 1 were derived from a comparable corpus of English and Italian texts on wine tasting 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 this genre both in English and in Italian are unknown to virtually all the students in this course. A specialised comparable corpus is indespensable to (learn to) search for genre-restricted phraseology and terminology, two of the central know-hows identified by Gouadec (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.

[3] `http://sslmit.unibo.it/~baroni`

[4] `http://www.comp.leeds.ac.uk/ssharoff/`

2.2. Challenges

While the new tools at our disposal make the construction of corpora from the Web easier for translators, certain obstacles still have to be overcome. First, the *BootCaT* toolkit at the moment requires basic Unix skills and access to a Unix server. A Web interface is a crucial next step if these tools are to reach the average translator.

In the longer term, widespread use of corpora and corpus construction and search facilities 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 POS 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: fully automatic matching for golden-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, p. 124)

Such a system would also arguably limit some of the drawbacks associated with the use of TM. 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*,[5] a commercial 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 corpora can be found on the *Multitrans* page: yet another proof that corpora are currently not a buzzword in the translation market?

3. Summing up: prospects for the future

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

[5] `http://www.multicorpora.ca/`

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 the importance of basic "translation" skills be restored to its 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 corpora. These materials should ideally be contrastive in focus (i.e., why/when use corpora instead of the Web/TMs/Dictionaries?). They should also 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 wordlisting). Finally, corpus construction and corpus searching tools should be made more user-friendly, and ideally integrated with CAT tools.

4. References

M. Baroni and S. Bernardini 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC 2004*, pp. 1313-1316.

L. Bowker. 2002. Corpus Resources for Translators. In S. Tagnin, editor *Corpora and Translation, TradTerm 10 Special Issue*.

L. Bowker. 2004. *Computer-aided Translation Technology.* Ottawa: University of Ottawa Press.

L. Bowker and J. Pearson. 2002. *Working with Specialized Language. A Guide to Using Corpora.* London: Routledge.

W. Fletcher. 2004. Facilitating the Compilation and Dissemination of Ad-hoc Web Corpora. In G. Aston, S. Bernardini and D. Stewart, editors, *Corpora and language learners*, Amsterdam: Benjamins.

D. Gouadec. 2002. Training Translators: Certainties, Uncertainties, Dilemmas. In B. Maia, J. Haller and M. Ulrych, editors, *Training the Language Services Provider for the New Millennium*, Oporto: Universidade do Porto, pp. 31-41.

M. Hoey. 2005. *Lexical priming.* London: Routledge.

R. Jääskeläinen. 1997. *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating.* Doctoral dissertation. Joensuu: University of Joensuu.

D. Kenny. 1999. CAT Tools in an Academic Environment. *Target*, 11(1), pp. 65-82.

MeLLANGE. 2005. *Corpora and E-learning Questionnaire: Results Summary.* Internal document, 20.06.05.

B. Mossop. 1999. What Should be Taught at Translation School? In A. Pym, editor, *Innovation in Translator and Interpreter Training - An Online Symposium*. online: http://www.fut.es/ apym/symp/mossop.html [visited: 23.02.06]

K. Pearson. 2003. Using Parallel Texts in the Translator Training Environment. In F. Zanettin, S. Bernardini and D. Stewart, editors, *Corpora in translator education*, Manchester: StJerome, pp. 15-24.

J. McH. Sinclair. 2003. *Reading Concordances.* London: Longman.

M. Stubbs. 2001. *Words and Phrases.* London: Blackwell.

E. Teich. 2003. *Cross-linguistic Variation in System and Text.* Berlin: Mouton.

M. Thelwall. 2005. Creating and Using Web Corpora. *International Journal of Corpus Linguistics*, 10(4), pp. 517-541.

E. Tognini-Bonelli. 2001. *Corpus Linguistics at Work.* Amsterdam: Benjamins.

A. Trosborg. 1997. Text Typology: Register, Genre and Text Type. In A. Trosborg, editor, *Text Typology and Translation*, Amsterdam: Benjamins, pp. 3-23.

M. Ueyama. forthcoming. Evaluation of Japanese Web-based Reference Corpora. In M. Baroni and S. Bernardini, editors, *Wacky! Working Papers on the Web as Corpus*, Bologna: Gedit.

K. Varantola. 2003. Translators and Disposable Corpora. in F. Zanettin, S. Bernardini and D. Stewart, editors, *Corpora in Translator Education*, Manchester: StJerome, pp. 55-70.

H.G. Widdowson. 1984. English in Training and Education. *Explorations in Applied Linguistics II*, Oxford: Oxford University Press, pp. 201-212.

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

F. Zanettin. 2001. Swimming in Words. In G. Aston, editor, *Learning with Corpora*, Houston, TX: Athelstan, pp. 177-197.

Translation as problem solving: uses of comparable corpora

Serge Sharoff

Centre for Translation Studies
University of Leeds, Leeds, LS2 9JT, UK
s.sharoff@leeds.ac.uk

Abstract

The paper describes an approach that uses comparable corpora as tools for solving translation problems. First, we present several case studies for practical translation problems and their solutions using large comparable corpora for English and Russian. Then we generalise the results of these studies by outlining a practical methodology, which has been tested in the course of translation training.

1. The problem

It is widely accepted that translation can be viewed as problem solving: in the process of producing a translation the translator encounters problems of various sorts and uses a set of tools and resources to solve them, cf. (Levý, 1967; Reiß, 2000; Varantola, 2003). Possible problems can involve detecting properties of the source and target audiences, determining the extent of the translation brief, designing the structure of the translated document, etc.

However, problems that occur most frequently in translation of practically every sentence are those of choosing the right target word for rendering source word X in context Y. One type of word-choice problems occurs in translation of terminology: the translator may lack knowledge about the exact translation of term X in domain Z. Another type of problems concerns the choice of words from the general lexicon: the translator knows a word and the standard set of its translations, but cannot find a target word that is suitable for the current context. The obvious way to find a solution for the word-choice problem is by consulting dictionaries. However, dictionary lookup may fail in both cases: a term can be missed in available dictionaries, while translation equivalents for general words suggested in the dictionary may not be usable in the target context. In the worst possible case, a dictionary can mislead the translator by listing a term or source expression with its translation, whilst the translation is NOT used in the target language in the suggested way.

In the following sections I will present several case studies of word-choice problems of the two types and outline ways to solve them using large monolingual corpora. Parallel corpora consisting of original texts aligned with their translations offer the possibility to search for examples of translations in their context. In this respect they provide a useful supplement to decontextualised translation equivalents listed in dictionaries. However, parallel corpora are not representative: millions of pages of original texts are produced daily by millions of native speakers in major languages, while translations are produced by a small community of trained translators from a small subset of source texts. The imbalance between original texts and translations is also reflected in the size of parallel corpora, which are simply too small to account for variations in translation of moderately frequent words. For instance, *frustrate* occurs 631 times in 100 million words of the BNC, i.e. this gives on average about 6 uses in a typical parallel corpus of one million words.

The procedure is illustrated by examples of translations between English and Russian using the corpora listed in Table 1.

All corpora used in the study are quite large, i.e. their size is in the range of 100-200 million words (MW), so that they provide enough contexts for moderately frequent words such as *frustrate*. The size is especially important for the detection of collocates, as even a 10 million-word corpus with its 63 hypothetical instances of *frustrate* does not provide sufficient grounds for deciding whether a single instance of *frustrate one's efforts* represents a recurrent pattern (there are 10 instances of this expression in the BNC).

However, the requirement for large corpora does not significantly limit the applicability of this study to other language pairs, as corpora of this size are increasingly available in a variety of languages. The size of about 100 million words is now the standard for so called "National Corpora", such as Czech (Kučera, 2002), Hungarian (Váradi, 2002) or Polish (Lewandowska-Tomaszczyk, 2003). The availability of huge amount of texts on the Internet in a great number of languages can produce Internet-derived corpora of practically arbitrary size, cf. (Kilgarriff and Grefenstette, 2003). What is more, an analysis of Internet corpora used in this study (they were produced by making a random snapshot of 50,000 pages indexed by Google) shows that an Internet-derived corpus is not radically different from the BNC in terms of its coverage of text types and domains. For more information about the properties of Internet-derived corpora see (Sharoff, 2006a).

Access to all corpora is available via a uniform interface (Sharoff, 2006b), which is powered internally by IMS Corpus Workbench (Christ, 1994). In comparison to other approaches using webdata as a corpus, e.g. Linguistic Search Engine (Resnik and Smith, 2003) and WebCorp (Renouf, 2003), the interface offers standard options for concordancing, queries for part-of-speech (POS) tags, detection of collocations and other statistical operations. Thus dealing with Internet corpora is not different in any respect from dealing with standard corpora, such as the BNC or British News.

Corpus	Size	Time frame
The British National Corpus	100 MW	1970-1992
A corpus of major British newspapers	200 MW	2004
The English Internet Corpus	130 MW	2005
The Russian National Corpus, a representative Russian corpus comparable to the BNC in its design(Sharoff, 2005)	100 MW	1970-2004
A corpus of major Russian newspapers	70 MW	2002-2004
The Russian Internet corpus,	130 MW	2005

Table 1: English and Russian corpora used in the study

2. Case studies

The general principle followed in the case studies below assumes gathering a set of expressions in the source language (most typically collocates of the source word or expression), making hypotheses about their translations and testing the hypotheses in the context of target language expressions. All original examples are taken from one of the corpora used (mostly from the newspaper corpus), while translations are provided by the author.

2.1. Terminology detection

Rapid development of a field of scientific research or political process produces a host of new concepts which are somehow rendered in both the source and target languages, but are not reflected in dictionaries. However, if they can be found in corpora, there is a possibility of finding a link between them.

For instace, recent political changes in Russia produced a new expression представитель президента ('representative of president'), which is as yet too novel to be listed in dictionaries or glossaries. At the same time we can use news corpora to identify the people that perform this duty: Драчевский, Латышев, Полтавченко, Черкесов. This can be done by building the list of collocates for the original expression (представитель президента) or by simply browsing through concordance lines. The hypothesis for translation is straightforward: we can search for the English transcription of their names, because they offer more or lesss stable translations. However, even in this simple case there is some variation in the way Cyrillic characters are rendered in English, e.g. letters like -ы- or endings like -ский, which can be rendered as *-sky, skiy, ski* or *-skij*. So it is safer to make a query:

[lemma='$Drachevsk.*|Lat.shev|Poltavchenko|Kirienko$'][1]

Note that it is unwise to include the first name of the person in question, even if it is frequently supplied in the original Russian text, because it can be omited in English or again transliterated in a less-standard way. The target names in British newspapers are accompanied with the following expressions *Putin's personal envoy* (twice) and *Putin's regional representative* (once). From this we can assume that no specific term has been established for this purpose in the British media, but either translation should be acceptable.

A similar technique can be used for the detection of possible translations of a technical term *environmental enforcement*, which is not listed in major English-Russian dictionaries. The most frequent collocates of this expression (counted for the span of 3 words) are *agency, authorities, government, office*. Given that the standard translation of *environment* is окружающая среда, we can make a query in which this term combines with a variety of expressions for government offices, agencies, etc. The frequency of окружающая среда in the three Russian corpora is about 3600, which gives sufficient evidence for detecting its collocates. The range of expressions to be found in this way includes departments and agencies for: охрана ('protection', 724 instances), защита ('guarding', 234), гигиена ('hygiene', 26) of the environment, as well offices for природопользования ('nature use monitoring', 82). Again this suggests the lack of a single translation equivalent, but corpora can guide translators about the range of expressions possible for naming environmental enforcement agencies in Russian.

2.2. Translating words from the general lexicon

Terminology in any established domain should be stable and allow one-to-one correspondence between the source and target languages. However, as we noticed in the examples above, there is some variation in the use of newly coined terms in domains of rapid development. Anyway we can assume that terminology in such domains will eventually settle down, be recorded in dictionaries and translated consistently. On the other hand, translations of words from the general lexicon depend on the context of their use, so that a dictionary can never give a complete record for all possible translations.

For instance, the Oxford Russian Dictionary lists three Russian translations for *frustrate*: разочаровывать, расстраивать, обескуражен. Yet in the majority of cases the most natural translation into Russian uses a word that does not belong to this set, e.g.

(1) En: *Saddam's ambition ... is frustrated by the presence of UN inspectors.*
Ru: Стремлению Саддама ... мешает пребывание инспекторов ООН.
Gloss: 'Saddam's ambition ... is hampered by the presence of UN inspectors.'

(2) En: *The share offer opens the possibility for thousands of frustrated commuters to air their grievances*

[1]The dot character in regular expressions refers to an arbitrary character, the asterisc to a sequence of such characters, the pipe character (|) to the disjunction operator.

Ru: Благодаря этой доле тысячи недовольных пассажиров получают возможность выразить свои жалобы
Gloss: 'Thanks to this offer thousands of angry passengers get the opportunity to express their complaints'

There are natural limits on the number of translation equivalents to be listed in a bilingual dictionary, imposed by its size and usability. A printed dictionary cannot afford to give separate translations for derived forms or list dozens of translation equivalents for a relatively unambiguous word, such as *frustrate* (for instance, English monolingual dictionaries list no more than two or three senses for it). As for usability, it is impossible to use a (printed or electronic) dictionary in which the relevant translation is buried in the long list of potential translation equivalents: a translator or a student will not find a translation they want. Entries for polysemous words have already too many suggested translations. For example, the entry for *strong* in the Oxford Russian Dictionary has 57 subentries and yet it fails to mention many word combinations frequent in the BNC, such as *strong feeling, field, opposition, sense, voice*.
The obvious strategy for finding translation equivalents for such examples is to check collocates of target words that are more straightforward for translation. For instance, *voice* in the context of *Her voice was surprisingly strong and powerful* can be reliably translated as голос, so we can produce a list of adjectives collocating with it. The resulting list is long (over 100 adjectives), varied and similar to the collocates the English word *voice* has, including женский (female), громкий (loud), глухой (husky), слабый (feeble), ровный (level), etc. The last adjective is particularly interesting, as the Oxford dictionary gives no suggestion on translating ровный голос, the expression *level voice* is possible in English, but it is nowhere as frequent as the corresponding Russian expression (11 vs. 327 instances in BNC-sized corpora). However, ровный голос fits perfectly into the context for the source example giving a smooth translation

(3) Она сказала это на удивление ровным и властным голосом
'She said this in a surprisingly level and powerful voice'

What is more this expression ровный голос can be used in the majority of contexts in which *strong voice* occurs in the BNC (unless *strong voice* implies 'loud voice'), so it can be treated as a reliable translation equivalent worth including in dictionaries.
In the next case study we will encounter a shift in the link between the two languages. If we want to find a translation equivalent for *strong feeling* as in

(4) *In Eastern Europe, meanwhile, . . . nationalist feeling is exceptionally strong*

neither of the two words (*feeling* and *strong*) provides a bridge between the source and target languages. However, *nationalistic* is translated in a restricted number of ways, which helps in building this bridge in two steps. First, we can find nouns correlating with националистский and националистический as two possible translations of *nationalistic*. Nouns that can be relevant in the current context include проявления (manifestations), риторика (rhetoric), убеждения (beliefs), настроения (attitudes), страсть (passion), etc. A separate study of concordance lines discovers that intensifiers for words from the list combined with *nationalist* do not typically come in the form of adjectives (like *strong* in English); they are either nouns or verbs: разгул (raging), разжигать (to fuel), усиление (strengthening). The latter expression can be further intensified by резкий (sharp), if this is what the translator wants to emphasise:

(5) В Восточной Европе тем временем произошло резкое усиление националистических настроений
'In Eastern Europe, meanwhile, sharp strengthening of nationalistic attitudes has happened'

In the last case study, the context of a problematic expression does not provide any reliable clues for its translation. The translation of *daunting experience* in the following examples:

(6) *Hospital admission can prove a particularly daunting experience.*

(7) *Even though you knew that what you said didn't matter, it was a daunting experience.*

does not depend on hospital admission or cross-examination, while neither *daunting* nor *experience* can be reliably translated using dictionary equivalents. One way to generalise the context in this case is by using "similarity classes", i.e. groups of words with lexically similar behaviour, cf. Chapter 8.5 in (Manning and Schütze, 1999). The similarity class of a word defines the paradigmatic relationship between it and other words that can appear in similar contexts. This is analogous to the definition of the relationship of synonymy in a thesaurus, but there is a difference, in that the notion of similarity classes is based on the affinity between the contexts in which the words occur. For instance, *strong* has the following similarity class: *powerful, weak, strength, potent, heavy, good, overwhelming, intense, robust, tough, weaken, compelling, fierce*.[2]. It is not the case that *strong* is synonymous with *good, heavy* or *weak*, but this is the case that they all occur in similar contexts. The notion of similarity classes provides an automatic procedure for generalising the contexts of a word in question.
If we compute similarity classes for *daunting* and *experience*,[3] we will get:

(8) daunting ∼ insurmountable (0.347), apprehensive (0.338), alarming (0.328), onerous (0.317), unfamiliar (0.314),

[2]There is no requirement that words in the similarity class have the same POS, even though it happens quite frequently that their POS is also the same because of the similarity of contexts.

[3]We use similarity classes computed using Singular Value Decomposition, as implemented by (Rapp, 2004). Figures in brackets show the relative similarity to the source word (*daunting*) according to the SVD measure.

forgivable (0.306), disconcerted (0.303), trepidation (0.300), incongruous (0.290), complicated (0.289), bleak (0.279), convincing (0.272),

(9) experience ∼ knowledge (0.357), opportunity (0.343), life (0.330), encounter (0.317), skill (0.317), feeling (0.316), reality (0.310), sensation (0.307), dream (0.296), vision (0.279), learning (0.277), perception (0.265), learn (0.263), training (0.263)

In the next step we produce an equivalence class, consisting of translations of words in the similarity class. As the list is large, it is easier to do so using an electronic bilingual dictionary (Oxford Russian Dictionary, in our case). For instance, the equivalence class of the Russian word опыт (experience) includes:

(10) ability, acquire, aptitude, capability, capacity, competence, courage, evidence, experience, experiment, expertise, feasibility, hypothesis, ingenuity, intelligence, knowledge, laboratory, learning, method, opportunity, perception, qualification, rat, research, skill, stamina, statistical, strength, study, talent, technique, test, training, vision.

The result reflects the ambiguity of опыт, which can mean 'experience', as well 'experiment' (hence the presence of *hypothesis, laboratory* and *rat* in the equivalence class), however it does preserve the semantic core of опыт, which is about skills and abilities.

In the final step we check target language corpora for uses of collocations consisting of members of the two equivalence classes. Even if an equivalence class contains some words that are not relevant to the source example, e.g. *hypothesis* or *rat*, those words create little noise, as they rarely collocate with words in the second equivalence class, e.g. *insurmountable* or *onerous*. Usually, this step brings 30-50 collocates whose relevance to the source language examples can be easily assessed, e.g. it should be obvious for the student that expressions like эффект устрашения ('deterrent effect') have nothing to do with the original query *daunting experience*. Then, the contexts of the remaining 5-7 relevant examples can be explored manually. For instance, *daunting experience* brings the following relevant collocates: безрадостный ситуация (dismal situation), волнующая возможность (worrying possibility), мрачный впечатление (gloomy impression), тягостное чувство (onerous feeling), устрашающее впечатление (intimidating impression).

Similarly, for *frustrated commuter/passenger* the procedure brings the following set of potential equivalents: пострадавший пассажир (suffered passenger), неудачный посадка (unfortunate boarding), недовольный пассажир (angry passenger), with the latter being the closest to *frustrated commuters* from the original example (2).

3. Considerations for the general methodology

This set of case studies can help in drawing generalisations about the use of corpora for problem solving. Baiscally this involves searching for 'islands' of stability in translation, around which we explore and compare contexts in the source and target languages.

In the first step we analyse the context of an expression in question (*environmental enforcement, strong voice, strong feeling*) in order to identify the functions performed by this expression in the source example and possibly in other similar contexts. The second step is to generalise the context of the original example by defining words indicative of the situation in question and extending the list with other words that can perform the same function. If contexts defy a reasonable generalisation, it is possible to use similarity classes, which statistically accumulate contexts most specific for the source expression. The third step is to build a bridge between monolingual corpora in the two languages by translating words with more obvious translation equivalents, such as names, *voice* or *nationalistic*. This step can be facilitated by the availability of a large-scale bilingual dictionary in machine-readable form, in order to produce equivalence classes without human intervention. The case studies presented above used the Russian Oxford Dictionary, some other studies conducted with my students used German and Spanish bilingual dictionaries, also provided by the Oxford University Press. However, it is possible to rely on one's intuition or to use traditional dictionaries, as it was the case with examples of *personal envoy* or *strong voice*. The final step in the methodology is to study the results of a number of queries in the target language that consist of words in the equivalence class in order to find lines which suggest suitable translation equivalents. If the number of occurrences of equivalent words is not large, as it was the case with the names of relatively obscure Russian political figures, it is possible to start with the study of concordance lines. If the number of concordance lines is too large to allow its direct exploration, as it was the case with *nationalistic* or *voice*, it is easier to study the most significant collocations for words in the equivalence class and then to study patterns consisting of these words with their collocations. Finally, if we use two very large equivalence classes, as it was the case with *daunting experience*, it is reasonable to intersect them in order to find expressions that regularly occur in the target language.

The possibility of applying this methodology is based on several assumptions. First, translators need to have skills in making queries to corpora and analysing lists of collocations and concordance lines. The latter involves skills in vertical reading of concordance lines, as the methodology crucially depends on the ability to notice and describe lexical patterns in raw data. Skills for vertical reading of concordance lines sorted around a keyword are different from those required for horizontal reading of a continuous text. Even if modern-day translators typically cannot do this type of research, a growing number of students in translation studies receive training in corpus linguistics and acquire skills for reading of concordance lines and detecting collocations.

The methodology also assumes the existence of sufficiently large source and target language corpora, such as the BNC as a general-purpose English corpus or the British news corpus for journalistic texts. As noted above, such corpora are increasingly available for a large number of languages. On the other hand, terminology in specific problem domains and register-specific word uses can be studied on the basis

of much smaller specialised corpora cf. related work (Bennison and Bowker, 2000; Zanettin, 2002b). For such tasks small disposable corpora can be even more useful, since they include more instances of terms and register-specific constructions to make generalisation specific to this domain. For instance, in a 5 MW corpus of software annotations collected from the Internet using BootCat (Baroni and Bernardini, 2004), there are 35 instances of the expressions *written in Java* and the majority of instances of *written in* are followed by the name of a prorgamming language. In contrast in the 200 MW corpus of British News there are only two instances of *written in Java*, while *written in* is typically followed by dates, locations and names of human languages.

4. Conclusions

When large corpora of the type of the BNC are used by translators, they typically provide a confirmation service: they are used to check whether a hypothetical translation equivalent is attested in authentic texts and, if yes, whether it is used in the same function as expected by the corpus user (Varantola, 2003; Zanettin, 2002a). Also students in translation classes can take part in lexicographic excercises which compare the contexts and functions of potential translation equivalents, for instance, *absolutely* and *assolutamente* (Partington, 1998).

In this study we went one step further and proposed a methodology that helps in solving the problem of choosing the right word for an expression. Even if the case studies discussed above solve problems of translation between English and Russian, we tried several exercises of this for various languages, such as Chinese, French, German and Spanish (the other language was English).

The methodology is especially useful for trainee translators. Professional translators have vast experience in finding lexical items that fit well into the context of translation. Some maintain "non-systematic" dictionaries (Palazhchenko, 2002), which highlight words that can cause troubles in translation and interpreting and explain contexts for their translations. Trainee translators on the other hand trust dictionaries, tend to use translations offered in dictionaries and feel frustrated when dictionaries do not provide them with solutions of their problems. Some of the case studies discussed above are not suitable for the practice of professional translators, either because the solution is immediately obvious for them or because finding a solution in this way takes too much of their time. However, the results are rewarding for trainees, because the final description covers not only the translation of a specific word in the context of a single example, but a wider range of contexts in which such words as *voice* and голос are used, as well as conditions for possible translations. This naturally fits into the education plan of trainee translators, which involves equipping them with a range of resources for finding contextually appropriate translations that go beyond what is offered in dictionaries.

The same methodology can be also of help for professional translators, if it is accompanied with automated means for generalising contexts and building bridges between the source and target languages. This link is explored in the ongoing ASSIST project (Sharoff et al., 2006), using semantic tags that are designed as uniform for the two languages, and USAS-EST, a software system for automatic semantic analysis of text that was designed at Lancaster University (Rayson et al., 2004). The semantic tagset used by USAS was originally loosely based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981). It has a multi-tier structure with 21 major discourse fields, subdivided into 232 sub-categories.[4] In the ASSIST project, we have been working on a tool that should assign syntactic and semantic tags to texts in comparable corpora and present source and target language examples that are similar in their semantic and syntactic contextual features. We expect that the use of similarity between contexts should reduce the number of irrelevant collocates and present only examples that can be potentially useful in the context of the current problem.

Acknowledgments

Research presented in this paper has been partly supported by EPSRC grant EP/C005902. I'm grateful to Bogdan Babych, Tony Hartley and Paul Rayson for useful comments on the earlier drafts of the paper.

5. References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of the Fourth Language Resources and Evaluation Conference, LREC2004*, Lisbon.

Peter Bennison and Lynne Bowker. 2000. Designing a tool for exploiting bilingual comparable corpora. In *Proceedings of LREC 2000*, Athens.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue of the web as corpus. *Computational Linguistics*, 29(2):333–347.

K. Kučera. 2002. The Czech National Corpus: Principles, design and results. *Literary and Linguistic Computing*, 17:245–257.

Jiří Levý. 1967. Translation as a decision process. In *To Honor Roman Jakobson: essays on the occasion of his seventieth birthday*, volume II, pages 1170–1182. Mouton, The Hague.

Barbara Lewandowska-Tomaszczyk. 2003. The PELCRA project — state of art. In B. Lewandowska-Tomaszczyk, editor, *Practical Applications in Language and Computers*, pages 105–121. Peter Lang, Frankfurt.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Tom McArthur. 1981. *Longman Lexicon of Contemporary English*. Longman.

Pavel Palazhchenko. 2002. *Moj nesistematicheskij slovar*. Valent, Moscow. (My non-systematic dictionary, in Russian).

[4] For the full tagset, see `http://www.comp.lancs.ac.uk/ucrel/usas/`

Alan Partington. 1998. *Patterns and meanings: using corpora for English language research and teaching*. John Benjamins, Amsterdam.

Reinhard Rapp. 2004. A freely available automatically generated thesaurus of related words. In *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*, pages 395–398, Lisbon.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with LREC 2004*, pages 7–12, Lisbon.

Katharina Reiß. 2000. Type, kind and individuality of text: decision making in translation. In L. Venuti, editor, *The translation studies reader*, pages 160–171. Routledge, London. Reprinted from 1981.

Antoinette Renouf. 2003. Webcorp: providing a renewable data source for corpus linguists. *Language and Computers*, 48(1):39–58.

Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Serge Sharoff, Bogdan Babych, Tony Hartley, Paul Rayson, Olga Mudraya, and Scott Piao. 2006. ASSIST: Automated semantic assistance for translators. In *Proc. of the European Association of Computational Linguistics, EACL 2006*, Trento.

Serge Sharoff. 2005. Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, and P. Rayson, editors, *Corpus Linguistics Around the World*, pages 167–180. Rodopi, Amsterdam.

Serge Sharoff. 2006a. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Serge Sharoff. 2006b. A uniform interface to large-scale linguistic resources. In *Proceedings of the Fifth Language Resources and Evaluation Conference, LREC 2006*, Genoa.

Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third Language Resources and Evaluation Conference, LREC 2002*, pages 385–389, Las Palmas de Gran Canaria.

Krista Varantola. 2003. Translators and disposable corpora. In Federico Zanettin, Silvia Bernardini, and Dominic Stewart, editors, *Corpora in Translator Education*, pages 55–70. St Jerome, Manchester.

Federico Zanettin. 2002a. Corpora in translation practice. In Elia Yuste-Rodrigo, editor, *Language Resources for Translation Work and Research, LREC 2002 Workshop Proceedings*, pages 10–14, Las Palmas de Gran Canaria.

Federico Zanettin. 2002b. DIY corpora: the WWW and the translator. In B. Maia, J Haller, and M Ulrych, editors, *Training the Language Services Provider for the New Millennium*, pages 239–248. Porto.

The Use of Corpora in Translator Training in the African Language Classroom: a Perspective from South Africa

Rachélle Gauton

University of Pretoria
Department of African Languages, University of Pretoria, Pretoria, 0002, South Africa
rachelle.gauton@up.ac.za

Abstract

This paper 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. In this paper I indicate 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 (a great many) 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 (forthcoming:5-9), Varantola (2002) and Zanettin (1998; 2002). This does not, however, seem to be the case on the African continent, and particularly in South Africa. As far as I am aware, 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 cours*e (*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 my institution, the University of Pretoria (UP), where (in 2004) I established courses on the application of Human Language Technology (HLT) in translation practice, focussing 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 paper, I therefore intend 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

With the advent of democracy in South Africa in 1994, eleven official languages were recognized. 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

[1] The so-called 'heritage languages' (the Khoi, Nama and San languages) and SA sign language, although not official languages, are promoted together with the official SA languages.

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 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.

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, I introduced translator training at the University of Pretoria at undergraduate as well as at postgraduate levels, and in 2004 added a number of postgraduate modules on the application of Human Language Technology (HLT) in translation practice. 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

In UP's Department of African Languages which hosts the translator training courses on behalf of the School of Languages, we have the following resources and infrastructure at our 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:

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

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

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 organized 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 I run a series of workshops 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

[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 myself.

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.

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.

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 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 paragraphs 3.4.1.-3.4.3 that follow.

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]:

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

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

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 following *ParaConc* screenshot in this regard, illustrating aligned English-Zulu parallel texts dealing with the *South African Qualifications Authority Act (Act 58 of 1995)*:

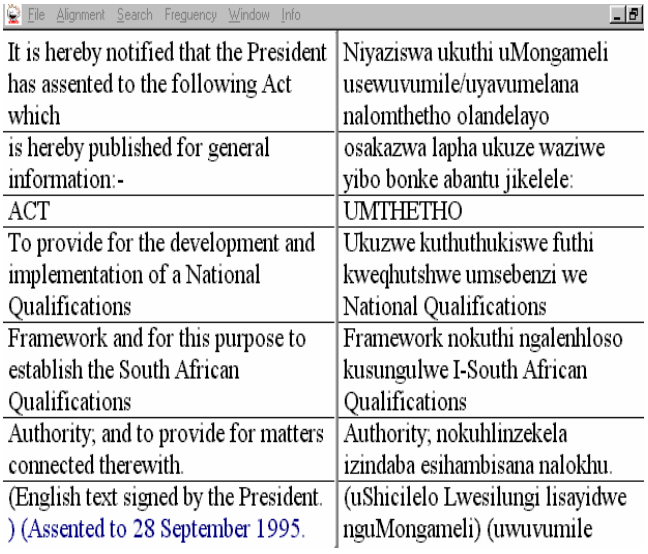


Figure 1: Screenshot of aligned English-Zulu parallel texts in *ParaConc*

[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.

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

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 Figure 1 for the format in which TL translations can be presented for the purpose of identifying translators' strategies. A glossary such as given in Footnote 5 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[5].

Due to space constraints, it is not feasible to give complete examples of students' work here, but see the following *DVX* screenshot illustrating the penultimate step in producing a translation from English into Zulu[6]:

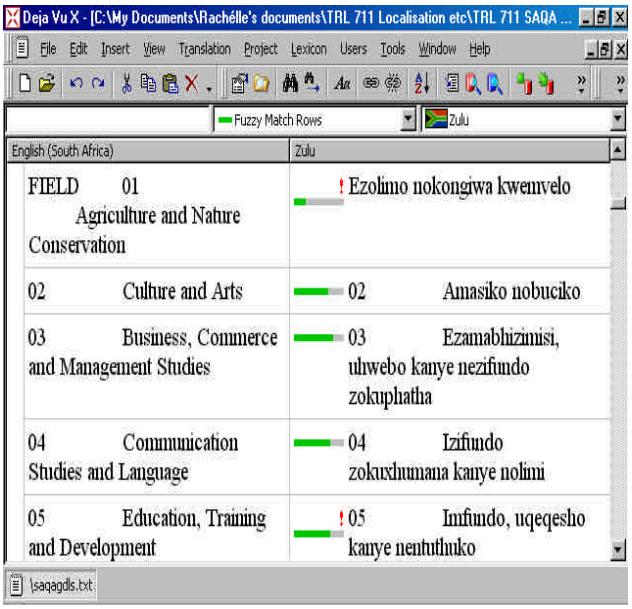


Figure 2: Screenshot of an English to Zulu translation being done in *Déjà Vu X* (*DVX*)

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% and the 2005 intake a class average of 75%. 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, I 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.

3.5. Conclusion

In this paper I have indicated how corpora are used with great success in the training of African language translators at the University of Pretoria, South Africa.

5 See for example the following extract from a Venda glossary based on the translation of a ST entitled *Cache and Caching Techniques*:

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

6 This is the so-called 'pretranslation' 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.

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. Furthermore, 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 my students, I 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.

3.6. References

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

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

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

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

Calzada Pérez, M. (2005). Applying translation theory in teaching. *New Voices in Translation Studies*, 1, pp. 1-11. Online. Available from 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. *Nordic Journal of African Studies*, 11(3), pp. 266-282.

Fictumova, J. (2004). Technology-enhanced Translator Training. In E. Yuste Rodrigo (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*, pp. 31-36. Online. Available from http://www.ifi.unizh.ch/cl/yuste/lr4trans-2/wks_papers.html

Gauton, R., Taljard, E. & De Schryver, G-M. (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, pp. 81-88.

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

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

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

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), 28 - 31 March 2003*, pp. 17 -25. Online. Available from http://www.coli.uni-sb.de/muco03/izwaini.pdf Last accessed on 12/02/2006 at http://web.archive.org/web/*/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), pp. 1-5.

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

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

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

Maia, B. (2002). Corpora for Terminology Extraction – the Differing Perspectives and Objectives of Researchers, Teachers and Language Services Providers. In E. Yuste-Rodrigo (Ed.), *Language resources for translation work and research, LREC 2002 Workshop Proceedings, Las Palmas de Gran Canaria*, pp. 25-28. Online. Available from

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), 28 - 31 March 2003*, pp. 27-34. Online. Available from http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

Maia, B. (2003b). The pedagogical and linguistic research implications of the GC to on-line 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, pp. 31-32. Online. Available from http://poloclup.linguateca.pt/docs/index.html

McEnery, A.M. and Xiao, Z. (forthcoming). Parallel and comparable corpora: What are they up to? In G. James & G. Anderman (Eds.), *Incorporating Corpora: Translation and the Linguist*. Clevedon: Multilingual Matters, pp. 2-14. Online. Available from http://www.lancs.ac.uk/postgrad/xiaoz/publications.htm

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

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

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

South African Qualifications Authority Act (Act 58 of 1995). Online. Available from http://www.saqa.org.za/publications/legsregs/index.htm#legs. Last accessed on 24/01/2006 at 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), pp. 119-132.

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

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

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

Standardizing the management and the representation of multilingual data: the MultiLingual Information Framework

Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret, Isabelle Kramer

LORIA / INRIA Lorraine
Projet “Langue et Dialogue”
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy
France
{Samuel.Cruz-Lara, Nadia.Bellalem, Julien.Ducret, Isabelle.Kramer}@loria.fr

Abstract

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 standardization. The issues related to this standardization are of industrial, economic and cultural nature. This article presents a methodology of standardization, in order to harmonize the management and the representation of multilingual data. Indeed, the control of the interoperability between the industrial standards currently used for localization (XLIFF)[1], translation memory (TMX)[2], or with some recent initiatives such as the internationalization tag set (ITS)[3], constitutes a major objective for a coherent and global management of these data. MLIF (Multi Lingual Information Framework)[4] is based on a methodology of standardization resulting from the ISO (sub-committees TC37/SC3 "Computer Applications for Terminology" and SC4 "Language Resources Management"). MLIF should be considered as a unified conceptual representation of multilingual content. MLIF does not have the role to substitute or to compete with any existing standard. MLIF is being designed with the objective of providing a common conceptual model and a platform allowing interoperability among several translation and localization standards, and by extension, their committed tools. The asset of MLIF is the interoperability which allows experts to gather, under the same conceptual unit, various tools and representations related to multilingual data. In addition, MLIF will also make it possible to evaluate and to compare these multilingual resources and tools.

1. Introduction

Standards make an enormous contribution to most aspects of our lives. People are usually unaware of the role played by standards in raising levels of quality, safety, reliability, efficiency and interoperability - as well as in providing such benefits at an economical cost. The scope of research and development in localization and translation memory process development is very large; many industrial standards have been developed: TMX, XLIFF, 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 a synergy that would make them interoperable when using one type of information in a slightly different context. Modelization 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 an uniform way of documenting such databases considering 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 organization without losing interoperability with existing standards or practices. Such an attempt should lead to more general principles and methods for analyzing existing multilingual databases and mapping them onto any chosen multilingual interchange format.

2. Contribution of standards

A multilingual software product should aim at supporting, for example, 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 (update, validation, correction) and consumption (use). To each one of these steps corresponds a specific user group, and a few specific scenarios. 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.

The development of scenarios considers the possible limits of a multilingual product, thus the adaptations required. Normalization will also allow the emergence of new needs (e.g. addition of linguistic data like some grammatical information). Scenarios help to detect useless or superseded features which it is not necessary to implement in the standardized software application. Normalization implies a specific applicative aim, in the sense that the scenarios which should satisfy the requests must be established with precision and so being based on well “on work practices” but can envisage some possible extensions. Normalization will facilitate the dissemination (export multilingual data) as well as the integration of data (import of multilingual data from an external database).

Providing normalized multilingual products and data can be considered as an advertising for a scientific community (e.g.: consulting Eurodicautom bases on the Net). Dealing with multilingual data is an expensive process, that is why a definite application would allow a return on investment, without forgetting the promotion of

the normalization experience of your entity (industry, research center...).

3. Terminology of normalization

As "Terminological Markup Framework" [5] in terminology, MLIF will introduce a structural skeleton (metamodel) in combination with chosen data categories [6], 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;
- specific information units, which can be associated with each structural node of the metamodel;
- relevant annotations, which can be used to qualify some part of the value associated with a given information unit.

3.1. What is a metamodel?

A metamodel does not describe one specific format, but acts as a kind of 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 organization of a specific standard into basic components. A metamodel should be a generic mechanism for representing content within a specific context. In fact a metamodel summarizes the organization 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 parameterize specific applications dealing with content.

3.2. 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) [6]. 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 semantic 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 [7] or 3 [8] 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 organization of an individual application, the organization of a specific domain.

3.3. Methods and representation

The means to actually implement a standard is to instantiate the metamodel in combination with the chosen data categories (DCS). This includes mappings between data categories and 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 to produce a "Generic Mapping Tool" (GMT) representation.

The architecture of the metamodel, whatever the standard we want to specify, remains unchanged. What is 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.

4. MLIF

Linguistic structures exist in a wide variety of formats ranging from highly organized 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 tagging) and sometimes to one another (e.g. References, annotations). The following model identifies a class of document structures which could be used to cover a wide range of multilingual formats, and provides a framework which 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 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.

4.1. Description of MLIF

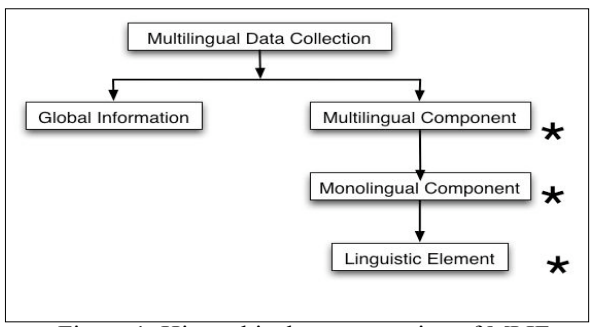


Figure 1: Hierarchical representation of MLIF

A MLIF document has a hierarchical structure as shown in Figure 1. This document will have "*Multilingual*

Data Collection" as the root level element, which content two major components: one and anly one "*Global Information*" element and one or more "*Multilingual Component*" element. The "*Global Information*" element can be considered as a header element because it contents metadata related to the document where multilingual text has been extracted and other administrative information. A "*Multilingual Component*" contains information that belongs to the linguistic unit (e.g. a single sentence or a paragraph, etc), descriptive informations (e.g. domain of application) or administrative datas (e.g. transaction, identifier, alias). Each "*Multilingual Component*" must content one or more "*Monolingual Component*" elements. A "*Monolingual Component*" is the linguistic unit in a given language. It could be a source text or a translation of this text into another language. Each of these "*Monolingual Component*" elements must contain one or more "*Linguistic Element*" elements. A "*Linguistic Element*" is the final unit of a MLIF document. It can be replaced by any metamodel such as, TMF[5], SynAF[9] or MAF[10].

For understanding what is MLIF, it is important to distinguish what depends, on the one hand, on the metamodel or, on the other hand, on the data categories. In fact, each structural node can be qualified by a group of basic or compound information units. A basic information unit describes a property that can be directly expressed by means of a data category. A compound information unit corresponds to the grouping at one level of several basic information units, which taken together, express a coherent unit of information. For instance, a compound information unit can be used to represent the fact that a transaction can be a combination of a transaction type, a responsibility, and the transaction date. Basic information units, whether they are directly attached to a structural node in the structural skeleton, or within a compound information unit, can take two non-exclusive types of values:

- an atomic value corresponding either to a simple type (in the sense of XML Schema) such as a number, string, element of a pick list, etc., or to a mixed content type in the case of annotated text;
- a reference to a structural node within the metamodel in order to express a relation between it and the current structural node.

4.2. Introduction to GMT

GMT can be considered as a XML canonical representation of the generic model. The hierarchical organization of the metamodel and the qualification of each structural level can be realized in XML by instantiating the abstract structure shown above (Figure 1) 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 [6] or that of a user-defined data category.

4.3. A practical example: MLIF and TMX

Now, we will use a very simple TMX example (see Figure 2) 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.

```
<tmx version="1.3">
    <header
        segtype="sentence"
        creationdate="19970101T163812Z"
    2   creationid="ThomasJ"
        changedate="19970314T023401Z"
        changeid="Amity">
        <note>This is an example of TMX</note>
    </header>
    <body>
        <tu tuid="0001">
            <note>Text of a note at the TU level.</note>
            <prop type="x-Domain">Translation</prop>
            <prop type="x-Project">LORIA</prop>
1           <tuv
                xml:lang="EN"
            4   creationdate="19970212T153400Z"
                creationid="BobW">
                <seg>The little cat is dead.</seg>
            3   </tuv>
            <tuv
                xml:lang="FR"
                creationdate="19970309T021145Z"
                creationid="BobW"
            4'  changedate="19970314T023401Z"
                ChangeId="ManonD">
                <seg>Le petit chat est mort.</seg>
            </tuv>
        </tu>
    </body>
</tmx>
```

Figure 2: Part of a TMX document

In Figure 2, 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 :

TMX structure	MLIF component
1 <tmx>	Multilingual Data Collection
2 <header>	Global Information
3 <tu>	Multilingual Component
4 <tuv>	Monolingual Component

Figure 3: matching TMX with MLIF components

Then, we will tag each element descriptor of TMX into 3 types: attribute, element or typed element. All these descriptors will be standardized into a MLIF descriptor element (i.e. a data category). For example the

TMX "xml:lang" attribute will be next matched with the data category named /*languageIdentifier*/ (cf figure 4).

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/*

Figure 4: typing of descriptor elements and matching with data categories.

Finally, the mapping of TMX elements into MLIF elements is represented in the following GMT file (figure 5). Note that this GMT file is nothing but a canonical representation of a MLIF document.

```
<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>
                  <feat type="primaryText">The little cat is dead.</feat>
                  <brack>
                        <feat type="transaction">creation</feat>
                        <feat type="date">19970212T153400Z</feat>
                        <feat type="author">BobW</feat>
                  <brack>
            </struct>
            <struct type="Monolingual Component">
                  <feat type="languageIdentifier">FR</feat>
                  <feat type="primaryText">Le petit chat est mort.</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>
      </struct>
</struct>
```

Figure 5: GMT representation

5. Conclusion

We have presented MLIF (MultiLingual Information Framework): a high-level model for describing multilingual data. MLIF can be used in a wide range of possible applications in the translation/localization process in several domains. This paper should be considered as a first step towards the definition of abstract structures for the description of multilingual data. The idea in a near future is to be able to implement interoperable software libraries which can be independent of the handled formats.

A first "informal" presentation of MLIF at AFNOR (Association Française pour la Normalisation - ISO's French National Body) on December 7th, 2005. We have obtained several very positive comments about our draft proposal. We are currently working on a "new work item proposal" that should be soon sent to ISO TC37 / SC4 subcommittee.

In addition, within the framework of ITEA "Passepartout" project [11], we are experimenting with some basic scenarios where MLIF is associated to XMT (eXtended MPEG-4 Textual format [12]) and to SMIL (Synchronized Multimedia Integration Language [13]). Our main objective in this project is 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.

6. Acknowledgements

We would like to thank Laurent ROMARY (TC37 / SC4 chairman) and Nasredine SEMMAR (CEA - LIC2M) for their useful comments and their kind help.

7. References

[1] XLIFF. (2003). XML Localisation Interchange File Format.http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff.

[2] TMX. Oscar / Lisa (2000) Translation Memory eXchange, http://www.lisa.org/tmx.

[3] ITS. W3C (2003) Internationalization Tag Set (i18n). http://www.w3.org/TR/its/

[4] S. Cruz-Lara, S. Gupta, & L. Romary (2004) *Handling Multilingual content in digital media: The Multilingual Information Framework*. EWIMT-2004 European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. London, UK.

[5] TMF. ISO 16642 (2003) *Computer applications in terminology -- Terminological markup framework,* Genèva, International Organization for Standardization

[6] ISO 12620 (1999) : Computer applications in terminology -- Data categories,

[7] ISO 639-1 (2002) *Code for the representation of names of languages – Part 1: Alpha-2 code,* Geneva, International Organization for Standardization

[8] ISO 639-2 (1998) *Code for the representation of names of languages – Part 2: Alpha-3 code,* Geneva, International Organization for Standardization

[9] SynAF: ISO/TC37/SC4/WG2 WD 24615 Syntactical Annotation Framework.

[10] MAF: ISO/TC37/SC4/WG2 WD 24611 Morphosyntactic Annotation Framework.

[11] ITEA "Information Technology for European Advancement". Passepartout project "Exploitation of advanced AV content protocols (MPEG 4/7)" ITEA 04017.

[12] 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.

[13] Synchronized Multimedia Integration Language (SMIL 2.0) . World Wide Web Consortium. http://www.w3.org/TR/smil20/

A Platform for the Empirical Analysis of Translation Resources, Tools and their Use

David Day, Galen Williamson, Alex Yeh, Keith Crouch, Sam Bayer, Jennifer DeCamp, Angel Asencio, Seamus Clancy and Flo Reeder

The MITRE Corporation
202 Burlington Rd., Bedford, MA 01730, USA
{day,gwilliam,asy,kcrouch,sam,jdecamp,asencio,sclancy,freeder}@mitre.org

Abstract

We have developed a software framework that will support experiments to explore the role of translator resources and tools in the performance of translation and translation-related activities. This software environment brings together a wide range of resources and tools within a single work environment that has been instrumented to measure the actions of the translator. In this paper we present an overview of the system that has been developed and describe the kinds of experiments that we intend to conduct. The platform provides detailed logs for most of the actions taken by a translator using the tool suite. We intend to use the data collected from controlled experiments to explore a number of questions, such as how resources and tools effect the productivity and quality of translators depending upon their level of experience, the texts on which they are working, the time constraints imposed on their work, and the mix of resources/tools made available.

1. Introduction and Overview

There is an increasing focus on the potential for linguistic resources and computational tools to enhance the productivity and quality of human translation. Computational linguists and tool developers are rushing forward to create a wide variety of tools and resources that they argue will provide translators, especially those without the benefit of complete mastery of the craft or working in terminologically demanding domains, with labor saving and/or insightful ways of approaching the process of transforming texts across the linguistic gulf of two languages and cultures. At the same time some experienced translators are dubious about the value of some of these proposals, arguing that the translation task is dominated by the very human process of comprehending the author's intent and finding an adequate choice of words for capturing this intent in the target language.

As part of the larger effort of coming to a greater understanding of how translators actually perform their tasks, and more specifically to identify the extent to which emerging computational tools and resources can influence the human translation task, we have developed an experimental prototype for performing empirical analyses of the translation task and the use of ancillary materials and tools. This platform has incorporated a range of resources and tools within an instrumented environment, by which we hope to record a number of relevant user behaviors in the process of performing translation and so help in understanding the true role for some of these contributions.

The emphasis of this environment is on exploring the potential advantages of tightly integrating emerging automated computational aids, This effort is not at the present time attempting to address some of the more fundamental aspects of translation and the associated cognitive processes, which have been and are continuing to be explored by others (e.g., Tirkkonnen-Condit, 1986; Danks, et al, 1997). This focus has meant that some capabilities that would support such explorations have been left out. For example, the level of detail we currently capture within the activity logs (details are described later in the paper) are suitable for addressing the specific hypotheses we wish to explore, but may not support the study of fine-grained cognitive process models, nor, at the other end of the scale, are we attempting to address larger workflow issues.

Our hope is that we will be able to identify some of the conditions under which various resources and tools provide a measurable impact on the translator's productivity or performance. While there are many different kinds of translation tasks, we are interested in broadening the notion of "translation" even more broadly, to incorporate anyone who is attempting to draw information from a foreign language source text and render it within some target language text. This broad definition will include those who wish only to generate a brief summary (or "gist") of a foreign language document, capturing the "salient points" made in the source document. It will include even those who may have extremely narrow "information needs" that are being applied against a source document, such as filling out an information template or questionnaire (e.g., "does this article mention my company's product?", "Does this article talk about cell phone technology?", etc.), or even simply assigning a topic label to a document.

Our reasons for expanding the scope of the analysis are motivated by both practical and theoretical considerations. The theoretical interest is to provide a broader continuum of behaviors which incorporate "foreign language document understanding" and so be better able to tease apart some of the issues surrounding target language composition vs. source language comprehension. The practical interest is that there is an increasingly rich set of ways in which foreign language material is being used within our highly interconnected society. Cross-language information retrieval can result in users attempting to extract meaning from documents in a foreign tongue for which the user is not fully literate. In some organizations relatively junior linguists might be responsible for routing foreign language documents to more experienced translators on the basis of intermediate language skills. Even for the task of full translation, some organizations need to sort through a large amount of

material only some of which may be relevant, requiring the translator to constantly assess the importance and level of effort required to capture the meaning of each sentence within a high-tempo work environment. It is desirable for our empirical analyses to be able to account for this wide range of interactions with foreign language material.

2. Resources as Tools, Tools as Resources

In the context of assessing the value of a resource to the human translator, it is impossible to fully divorce the "static resource" (e.g., a bi-lingual technical dictionary) from the way in which the resource is made available to the translator. Whether it is in form of a hardcopy book, a human mentor, or a computer program, the "interface" between the human and the resource is a key variable that greatly influences the utility of the underlying information. This inevitably creates the opportunity for a confounding influence on the empirical results – a lousy index or lexicographic ordering system can render the richest bilingual dictionary useless to the dictionary reader. Similarly, an annoying or ineffectual user interface in a computerized version of this same resource can thwart the ability of an empirical experiment to identify its underlying value.

These observations indicate how ambiguous is the division between "resource" and "tool" within the scope of experimental analysis of their utility in practice. They also point to a significant caveat that will need to be appended to many empirical results – sometimes a result will provide only a lower bound on the utility of a given resource, since a better interface might enable the resource to be even more useful. We have attempted to provide a relatively consistent framework for interacting with the various tools and resources provided within the experimental platform we have built, so that some of the issues might be said to be "held constant." Nonetheless, the nature of human-computer interaction is extremely complex, so the empirical results that come out of the use of this experimental translation platform must always be viewed with these issues in mind.

3. Resources and Tools to Support Translation

In as many cases as possible we have tried to incorporate strong commercial resources and tools so that the results of our empirical studies reflect as much as possible on the state-of-the-art abilities within these areas. Since we were committed to tight integration and logging, the unfortunate result is that there were many resources (e.g., dictionaries) or applications (e.g., translation memories) for which our desire for integration ruled out many excellent commercial offerings. It is important to note that we are not attempting to evaluate the performance of the individual components that are being integrated, but rather our goal is to perform an evaluation of the relative and absolute value of a class of resource/tool to the translation enterprise. Of course, we can't get around the fact that that we are limited in the

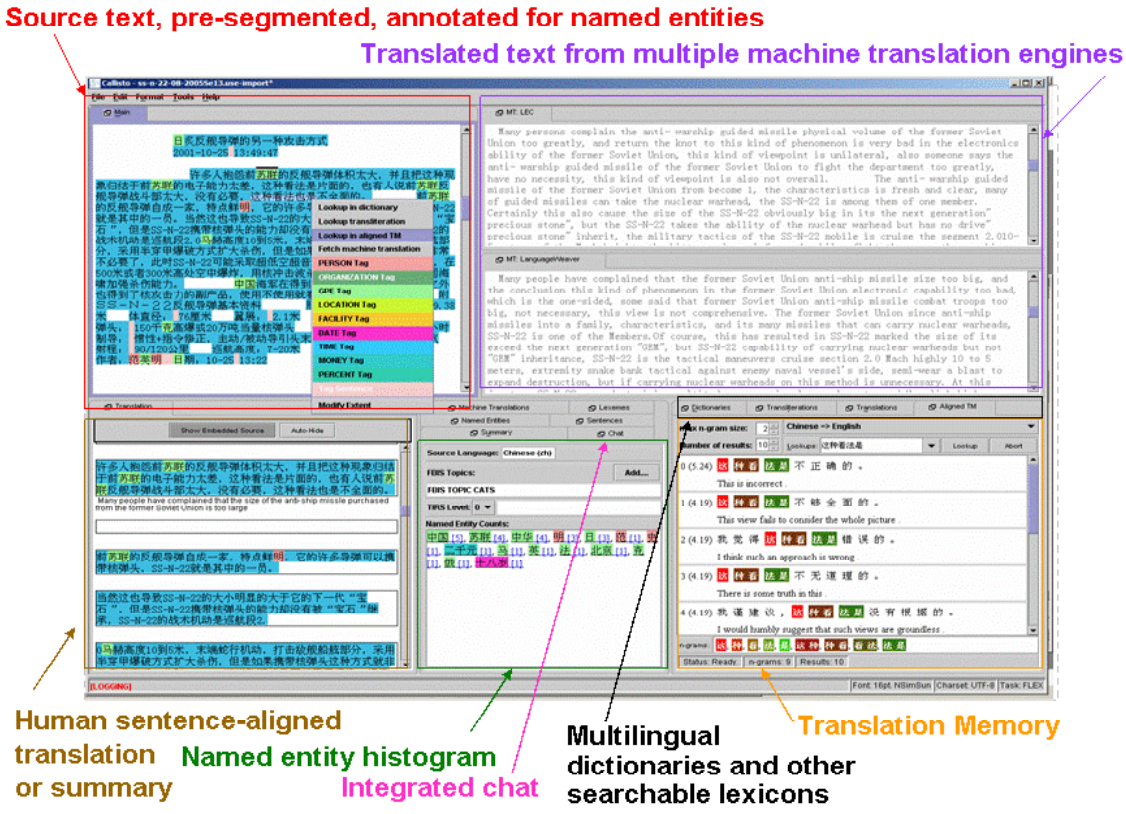


Figure 1: Screen image of the C/Flex translation experiment platform. The user has selected a sentence in the MT pane, with the result that the aligned sentences in the source text pane and the other MT output panes are also aligned and displayed. The dictionary pane displays the results from an earlier query.

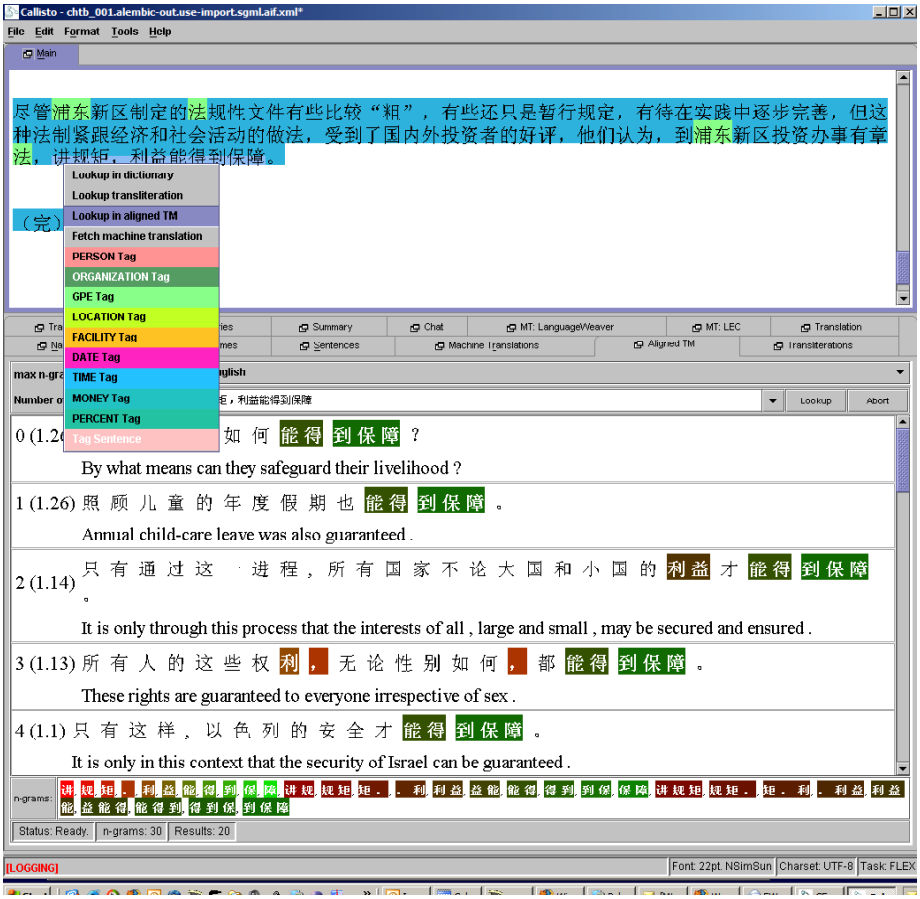


Figure 2: Screen image of a query to and search results from the translation memory (TM).

number and kind of resources and tools we have been able to put together “under one roof,” so the logging data we capture will necessarily reflect the particular resources and tools we happened to integrate.

The experiment platform that we have created is called CalliFlex (and commonly abbreviated as C/Flex). The current version of CalliFlex has been set up for handling three main source languages: Chinese, Arabic and Spanish, though it is easy to add components, resources and tools to support additional languages, if they are available. The target language in all of its configurations has been English. CalliFlex integrates the capabilities and resources listed below.

1. Multiple (usually two) machine translation (MT) systems for each language;
2. An integrated and easily accessible and searchable set of bilingual dictionaries, monolingual dictionaries, onomastica (mono- and bi-lingual person name resources), and other bilingual transliteration and translation resources and automated tools;
3. A large pre-populated translation memory;
4. Source language automatic natural language processing that can identify
   a. Sentence boundaries;
   b. Word boundaries;
   c. Parts-of-speech for each word; and
   d. Named entities (distinguishing among names for persons, organizations, locations, geo-political-entities [“GPEs”] and artifacts);
5. A text chat tool that enables translators to collaborate among CalliFlex clients.
6. A spell-checking facility in the target language translation and/or gist panes (currently restricted to English).

Each of these types of tools/resources are presented to the user within distinct “panes,” and each of these panes can be independently placed anywhere within or outside the borders of the application. The application allows for within-border pane movement in a manner similar to that supported in the Eclipse development environment, enabling an efficient means of filling up the available screen real estate with the selected components. Figure 1 shows an image of one possible layout of the application components

After the user selects a document to import into C/Flex and identifies the language of the source material, the application proceeds to invoke the natural language processing component that identifies sentence boundaries, word boundaries, part-of-speech assignments and named entity expressions. Of course, these automatically derived analyses can and will include errors. (Errors in named entity recognition can be corrected directly by the user -- the system incorporates most of the annotation editing features from the Callisto annotation tool from when it is

derived. We do not yet support editing sentence and word segmentations, or part-of-speech assignments.)

The derived sentence boundaries are then used to invoke the multiple MT engines iteratively, enabling the source sentences and the MT output sentences to all be aligned. The user is able to browse through the document, sentence by sentence, maintaining all the "views" of the sentence in synchrony – source, multiple MT renderings, user's translation.

The CalliFlex prototype also incorporates a pre-populated translation memory (TM). The query interface to this resource generates all possible distinct adjacent multi-word phrases (the user can control the maximum length of these multi-word phrases), and then searches the TM using standard information retrieval metrics for establishing sentence similarity between the collection of phrases and the target source sentences. (In the case of Chinese, these n-grams are measured in characters.) The sentence pairs (source language and target language human translation) returned by the TM search are presented in order of decreasing similarity, and the various multi-word phrases generated as the search query are separately highlighted in the returned sentence pairs. Figure 2 shows a sample of the TM interface when translating a Chinese source document.

The data used to pre-populate the Chinese and Arabic TM data sets is taken from the TIDES 2005 evaluation corpus. These are a mix of general reporting, magazine articles and parliamentary proceedings. The data we will be using in our controlled experiments will come from the general news, so we expect there to be a reasonable intersection of genres between sources articles and TM data.

Queries to any of the resources/tools (dictionaries, name resources, MT modules, transliteration modules, TM) can be invoked either directly from the source text (via word selection and selection from a pop-up menu) or via direct query type-in (assuming the user has access to the appropriate type-in methods on the client computer for that language), and can be invoked in either language direction, source→target or target→source. This ability to change directionality and enter user-generated text allows translators to explore different variations of the source words as well as explore the behavior of the tools/resources themselves under different conditions.

Translation often involves collaboration with other experts, and in some experimental contexts we wish to be able to allow this collaboration to take place while being tracked by the CalliFlex application. For this reason we have incorporated a simple multi-party text chat tool within the client.

The CalliFlex architecture has been developed to enhance the ability for rapid integration of third party tools and resources. Figure 3 illustrates how most of the resources and tools are made accessible via a Tomcat web server, enabling multiple CalliFlex clients to access them via web-based protocols. The dictionaries and transliteration resources are stored in the OLIF2 interchange format (McCormick, 2004), which are then indexed by a Lucene search engine, enable full-text and fielded search from the client.

4. Application Logging and Post-Experiment Analysis

The CalliFlex prototype captures a log of all of the following types of information, associated with each user session. Every log identifies the user ID (possibly an anonymous but unique ID), and each entry includes a time stamp.

1. Start and end times – when a document is first imported into the tool, the source document's language, the target language of the translation, the various features within the CalliFlex tool that have been made available to the user, and when the state of the system (including translation/summary) is saved or exported at the end of a session.
2. Resource lookup – When the user queries a resource such as a mono- or bi-lingual dictionary, transliteration resource, translation memory, etc. This also includes various tools that perform automatic processing on the query string, such as machine translation engines (used as a dictionary), transliteration algorithms, etc. The logged information includes whether the string was entered directly by the user or whether it was copied (swiped) from one of the application panes, in which case the identity of the source pane will

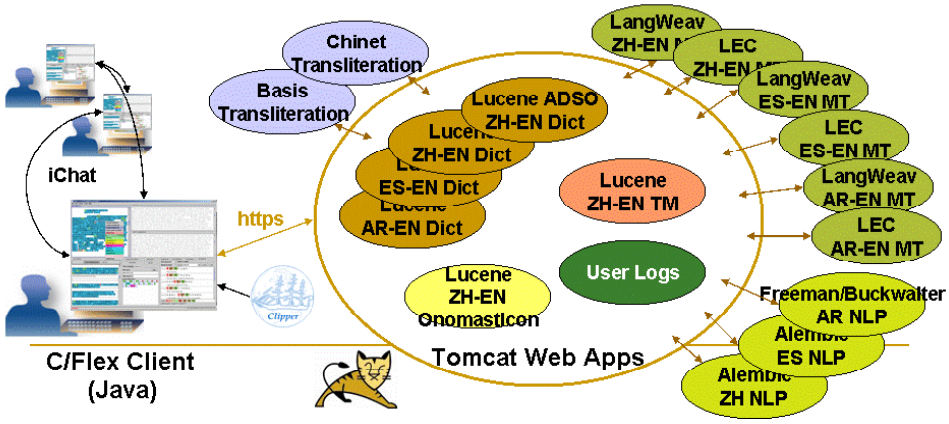


Figure 3: The Calliflex (C/Flex) Architecture. The resources and tools can be hosted locally or remotely, with remote access via TCP/IP web app. Simple application wrapper and protocols encourage rapid incorporation of new tools and resources.

also be included.

3. Translation pane updates – The source of text entered into the translation/summary pane is identified, whether it is from a copy-and-paste (e.g., from one of the MT output panes) or typed in by the user.
4. Annotation edits – As noted earlier, the source text is automatically processed to reveal predicted sentence boundaries, word-boundaries, part-of-speech assignments and named entities. Each of these types of assignments are prone to error (varying greatly depending on the genre of the source text as well as the extent to which it is complete sentences vs. abbreviated text as one often sees in web sites). The tool currently supports user editing of named entities and sentence segments, but not of the word segments and part-of-speech assignments.

As noted in the introductory section, our experimental focus is currently on the assessment of the relative contributions made by state-of-the-art computational aids such as machine translation, automatic named entity recognition, name transliteration, etc. This focus has meant that some of the more detailed levels of logging are left out that are present in other tools. For example, Translog (Jakobsen, 1999) is a powerful text editing analysis tool that has been used successfully in studies of translation (as well as other text editing tasks) that provides character-by-character update timing. We have not attempted to replicate this kind of fine-grained logging for the present experimental goals.

However, given the large number of different sources from which a given piece of target (translation) text can derive from, this is an addition and important piece of information that we are able to track. Thus, the logs will keep track of whether the text that is entered into the translation was copied from a transliteration pane, a machine translation output pane, a TM sentence, etc., and a timestamp on when this update happened.[1] We will also track deletions from the translation pane.

We have only recently brought the CalliFlex tool to a state sufficient to support experiments, so as of the date of writing this paper we do not yet have sufficient empirical experiment results to analyze. At this point we can only report on our plans for testing various hypotheses and the experimental setup and data we intend to capture to explore these hypotheses.

4.1. Translation/Summarization Productivity

One of the first questions on which we will concentrate in our studies is that of translator productivity: Can the tools brought together within the CalliFlex experiment platform support measurable productivity gains in full translation or target language summaries without any diminution in quality? Our hypothesis is that this is indeed the case for relatively junior translators and for those working in highly dynamic/technical subject disciplines. As translator skill increases, the utility for resources and tools diminish. A corollary hypothesis is that resource/tool utility in these target populations is increased in relation to the amount of time pressure imposed on productivity. While these hypotheses are modest, we hope to provide concrete empirical evidence, and also begin the process of identifying the relative contributions of different kinds of resources and tools within these different translator populations and work contexts.

External time pressures are one of the dominant aspects of the translation and summarization work contexts that we wish to study. Earlier work on time pressure in translation (e.g., Jensen, 1999) has attempted to elucidate cognitive explanations for the differences in behavior observed under different temporal constraints and with translators of different skill levels. Our data collection and analysis will concentrate initially on the differential influence that automated methods and enhanced resources play in both the translators' use of these tools, and to the extent possible the degree to which these tools actually do ameliorate the deleterious effects of tight time constraints. As in the earlier work, we anticipate, and will attempt to carefully track, the different performance characteristics that will be associated with highly experienced translators versus more junior translators.

Our experiments will be conducted in the following manner:

1. Each subject will be given a questionnaire, in which we inquire about their level of expertise in the language, translations skills and experience, the kinds of tasks they perform in their normal work, etc.
2. CalliFlex Tool Suite Training. This will be an important variable to measure. The tool is fairly complex and provides a great deal of user-customization options.
3. Summarization/Translation. The subject will be provided a fixed number of documents and will be asked to translate, or in a different experimental context be asked to provide a summary translation, of each document's contents. In the case of the summary, a specific expected length will be determined. The summary will be "domain independent" – the subject should attempt to capture as much of the "important" elements of the document as possible.
4. The subject will be provided these materials in four fixed time-period segments. Within each segment there will be one of two experiment conditions adopted: enabling all of the CalliFlex tools and resources to be available to the subject, or enabling only the ability view the source and write the translation/summary (with all other resources available only via hardcopy documents). The order in which these conditions are presented will vary among subjects.
5. A post-experiment questionnaire will be given to the subject, in which we ask a variety of subjective assessments of the tools and resources provided within the experiment, ideas for improvements, how well the experiment seemed to capture their usual work environment, etc.

[1] Note that if a user elects to type in some text from one of these other panes, as opposed to the more natural copy-and-paste action, the tool will not be able to determine the actual source of the text.

6. Post experiment analysis. The data will be analyzed from a number of perspectives. A particularly important analysis will be assessing the quality of the translation or summary, and associating this quality against the experiment conditions (with or without various tools) and the amount of time it took for the subject to generate this product. Our test data include eleven human translations for each source document. We will use various simple domain-independent Likert numerical quality scales (1 – 5) by which multiple evaluators will grade the quality of the translations/summaries. While such an evaluation metric may be crude, our focus is on measuring performance differences in relatively junior translators, where there is often a fairly high variability in translation or summary quality. In the case of full translations, we will also make use of the standard MT evaluation metrics such as BLEU (Papineni, et al, 2002), though these have limited discriminatory power. In the case of summaries, we will employ some of the techniques developed in the Document Understanding Conferences (Dang, 2005) evaluations to establish how many of the key elements of information have been included in the summary. These key elements are identified by comparing against summaries generated by multiple evaluators. These evaluators will work against the translated documents rather than the source documents.

Our "standard" experiment protocols do not presently call for any formal role for Think Aloud Protocols or TAPs (Lörscher, 1991). This is mostly because we wish to attempt to reduce the per-subject costs of the experiment sufficiently to enable a relatively large subject population, and thereby increase the opportunity for statistically significant numbers in our captured data. However, we are cognizant of the immense influence that particular user interface design elements can have on our experiments. For this reason we intend to conduct small scale, mostly in-house interactive experiments in which we record the video and audio of the experiment session, and in which we may introduce questions and ask for feedback. These sessions will not attempt to rigorously pursue the TAP methodology, however.

Due to the complexity of the tool, we have developed a fairly rigorous training and exercise regimen to familiarize subjects with the wide variety of tools and information sources. These training sessions include many opportunities to provide anecdotal feedback to us on what they think about the tool's components and their utility in performing translations or summarizations.

The discussion of the experiment platform and experiments has so far concentrated on the translation and "gisting" (summarization) tasks. If we are able to obtain sufficient experimental subjects and associated experiment materials we hope to explore a number of similar hypotheses associated with different translation "tasks" such as name finding, template filling, reading comprehension tests, etc., as well as establishing different kinds of experiment conditions in order to mimic those that might be found in different work environments – for example, subjects occupying a crowded room in which others are performing similar tasks; allowing collaboration among subjects with similar or different levels of skill, etc.

5. Current Status and Plans

We are just now, in the Spring of 2006, beginning to perform experiments with as many subjects as we are able to find. We intend to make the CalliFlex application available to other researchers without charge in order to encourage a greater investment in the empirical study of translation and how emerging linguistic resources and tools can enhance the productivity and quality of this important activity. The application will enable others to incorporate different resources and computing capabilities that may open up new experiments and test different hypotheses.

6. References

Dang. Hoa Trang (2005). Overview of DUC 2005. Presented at the Document Understanding Workshop, HLT/EMNLP Conference, October 9-10. Vancouver, B.C., Canada (http://duc.nist.gov/).

Danks, J., G. M. Shreve, S. B. Fountain, M. McBeath , eds., (1997). Cognitive processes in translation and interpreting. London: Sage Publications.

Jakobsen, A. L. (1999). Logging target text production with Translog. In G. Hansen (ed.), in Probing the process in translation: methods and results (Copenhagen Studies in Language 24). Copenhagen: Samfundslitteratur, p.9-20.

Jensen, A. (1999). Time Pressure in Translation. In G. Hansen (Ed.), *Probing the Process in Translation: Methods and Results* (pp.103-119). Denmark: Samfundslitteratur.

Lörscher, W. (1991). "Thinking-aloud as a method for collecting data on translation process". In S. Tirkkonen-Condit (ed.). Empirical research in translation and intercultural studies. Tübingen: Gunter Narr, p.67-77.

McCormick, Susan (2004). Using OLIF, The Open Lexicon Interchange Format. Presented at the 6th Conference of the Association for Machine Translation in the Americas, Georgetown University, Washington DC, September 28 - October 2.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL), Philadelphia, July, pp. 311-318.

Tirkkonen-Condit, S. (1986). Empirical studies in translation: textlinguistic and psycholinguistic perspectives. Joensuu: University of Joensuu.

Using corpus information to improve MT quality

Gregor Thurmair

linguatec
Gottfried-Keller-Str. 12, 81254 Munich
g.thurmair@linguatec.de

Abstract

The paper presents activities to improve the quality of a rule-based MT system, using corpus information. If focuses on the area of dictionary, and shows how, and to which extent, corpus-based information can improve the system quality in the different areas of dictionary development. It deals with two main sources of errors: missing entries / translations, and wrong selections of one out of several possible translations.

1. Baseline

Quality of machine translation is still a critical topic; however, research has not really focused on this issue; instead there were many attempts to start anew, hoping that a change in technology would lead to improved system quality. However, up to now, this has not proven to be the case. A comparison of the different approaches, rule-based MT and statistical MT (henceforth SMT), found two main results (cf. Thurmair, 2005, comparing German-to-English MT):

- The overall quality of SMT is outperformed by existing rule-based MT systems.
- The overall quality of both approaches is not yet sufficient. Between 20 and 30% of the evaluated sentences were ranked as being unacceptable.

A closer evaluation of the results shows that the main sources of errors in SMT (about 60%) are related to phenomena like German split verb constructions, non-standard constituent ordering, gapping etc., all of which could be rather easily described in a rule-based context; while the main sources of errors in rule-based systems (again about 60%) consist of lexical issues, and wrong selection of lexical material, which in turn a corpus-based approach can easily avoid[1].

As a consequence, efforts seem to be adequate to merge the power of corpus-based methods with the advantages of a rule-based system architecture, starting with the dictionary, identified as one of the major weaknesses of current MT systems.

Lexical mistakes, in general, result from two sources: no translations exist in the dictionary, and too many translations exist, and a wrong one is selected.

2. Missing transfers

Missing entries damage not just the translation (as they cannot provide content fidelity), they also hamper the analysis of the rest of the text. They result from three main sources: missing general purpose words, missing special terminology, and proper name issues.

2.1. Gaps in general vocabulary

The most straightforward case is dictionary gaps. But current MT dictionaries contain several 100 K entries, and gaps are not so easy to identify. On the other hand, in most existing MT dictionaries, surprising entries can be found. Experiments have shown (Dillinger, 2001) that MT dictionaries contain a significant amount of entries on which coding effort has been spent, but are nearly never used.

Obviously, corpus-based technologies of monolingual and bilingual term extraction can be used to close dictionary gaps, based on frequency information[2]. In the context of linguatec's 'Personal Translator', missing entries with a frequency more than 5000 were identified and added to the system dictionaries.

2.2. Corpus-based terminology

Beyond general vocabulary words, there is a huge amount of terms not represented in MT dictionaries, mainly terminology for special domains. Corpus-based techniques here are to be preferred to conventional dictionary entering:

1. Studies in the automotive sector showed that even special domain dictionaries with high reputation, in a significant amount of cases, propose translations which sound plausible but are not at all used in the target language[3]. Using such translations can make the text not understandable.

2. Often it is required to meet special user terminology requirements. E.g. if users allow for cross-lingual searches on their web sites, terms must be translated in a user specific way, otherwise they lead to poor search results.

Again, corpus-based work is required to provide adequate terminology, given the fact that special terminology has multiple translations.

2.3. Proper Names

Proper names are another large source of unknown words. Although they form a considerable amount of the vocabulary, only recent research (Babych and Hartley, 2003; Jiménez, 2001) into proper names shows their potential for quality improvement.

Proper names cannot be stored in dictionaries, as there is a too large and ever growing amount of them. But users often are puzzled if proper names are treated incorrectly.

1. **Not treating** proper names at all often results in parsing errors, like with other missing lexical elements in

[1] The rest consists mainly in grammatical mistakes; wrong structures are selected in a given situation.

[2] The linguatec corpus for German and English, collected for the work presented here, consists of 700-800 million word forms each.

[3] This can easily be verified by searching for them in the Internet.

the input. In addition, the fidelity criterion is always violated if names which are used in the source text do not show up in the translation.

2. Such problems can be avoided if proper names are marked to be '**don't-translate**' words, as is possible in some systems[4]. Then the proper names undergo some default system treatment (usually: noun with some default values for gender and number). However, this can be incorrect as proper names have different syntactic properties: They inflect (like in Russian or German), they differ in number (plurale tantum like *the Hebrides*, *les Pyrénnées*), they take special prepositions, etc.; so more information is needed than just the default.

3. Therefore, a full **named entity recognition** component is required to improve the analysis, by providing information about constituency and attachment (*He robbed [the Bank of Scotland]* vs. *He robbed [the Bank] [of Scotland]*) and semantic type of proper names.

Named Entity recognition often uses statistical or shallow parsing technology, and there are two options of integration into an MT system: running as some pre-processor, or being integrated into the full syntactic analysis. Full integration tends to be less robust (in case of parsing errors), but can better deal with homographs (de *Peter Maurer war Maurer* -> en *Peter Maurer was a bricklayer*) or gender issues (en *Anna Frank was a teacher* -> *Anna Frank war Lehrerin*)[5]. In addition, there is another feature of Named Entity recognisers, namely coreference analysis, which affects conventional MT system structure: Coreference is a feature which is text based, and MT systems which are sentence-based cannot really cope with it. In the following example, while the first occurrence of *Schneider* is recognised by contextual analysis, sentence-based MT systems fail to identify it in the third sentence, and therefore incorrectly translate the name there:

Das FDP-Mitglied Dr. Schneider lebt in München. Dort ist es heiß. Schneider ist der erste ausländische Politiker.

The FDP member Dr. Schneider lives in Munich. It is hot there. Tailor is the first foreign politician. (instead of: *Schneider is the first foreign politician*).

4. A special challenge consists in the **translation** of proper names. This is where MT systems need to extend standard NE recognisers, which only identify their entities. While it is a common mistake of MT systems to translate proper nouns (en *Mrs. Rice* -> de **Fr. Reis*, de *Hr. Fischer* -> en **Mr. Fisherman*), it is only true for person names that they must not be translated[6]. Dates usually must be translated to accommodate to the target language's conventions. Places behave differently: some are translated (en *Ivory Coast* -> fr *Côte d'Ivoire* -> de *Elfenbeinküste*), others are not (e.g. *Santiago de Compostela*). Often such place names are put into the dictionary.

Proper names can also have different linguistic properties in source and target language, which is relevant for generation: The *Désert du Thar* is masculine in French but *Thar Wüste* is feminine in German, and so is *Rhône* where even the lemma is identical in both languages. *Balkan* is singular in English but plural in Russian (*Балканы*). For product names, the gender seems to be dependent on the ‚base type': cars like *Renault* default to be masculine in German (derived from *der Wagen*) but feminine in French (derived from *la voiture*); determiner placement is language specific as well:

fr *L'Italie* -> de *Italien* but
fr *La Suisse* -> de *die Schweiz*.

While some of these cases can be handled by default assumptions, others are idiosyncratic, need to be detected by corpus work (cf. Jiménez 2001) and require a special resource to describe them.

5. The **evaluation** of integrating a named entity component into an MT system (the linguatec 'Personal Translator') was done as follows: A total of 1500 sentences from the news domain was selected in three language directions, 15% of which contained proper names. They were analysed with and without the proper name recogniser, and the results were compared.

The evaluation showed an increase in translation quality for sentences containing proper names by about 30% on average. The main improvements were:

- no erroneous translations of person names, esp. in coreference positions
- better contextual adaptations (correct preposition and determiner selection; and correct pronominalisation)
- better parses in some cases (e.g. segmentation of dates containing periods).

Of course the overall quality gain for a given corpus depends on the number of sentences containing proper names, and will be higher in news text translation than e.g. in computer manuals.

3. Wrong translation selection

While the problem of missing dictionary entries seems to be reducible to a tolerable size, the opposite problem is much more difficult to solve. It consists in an improper selection of a target term from a number of candidate translations. This problem aggravates with growing numbers of dictionary entries and increased system intelligence. And this is what articles like 'Have fun with MT' refer to:

Wortebene is *word level* and not *word plane*, and *Stromunternehmen* is not a *river expedition* but an *electric power producer*.

The challenge consists in the selection of the proper translation in a given context. It should be noted that dictionaries for humans contain much more translation variants than even large MT dictionaries, which increases the relevance of the problem.

3.1. Current disambiguation means

State-of-the-art systems offer two possibilities to select translation alternatives:

1. **Global settings** by users. Systems provide options for subject area settings, for customer settings (to cover customer-specific terminology), for locales (to select for $truck_{US}$ or $lorry_{UK}$), for conservative vs. progressive

[4] Babyh & Hartley (2003) tested a recogniser for named entities, and marked all of them as don't-translate words.

[5] Frank et al., 2004

[6] Albeit transliterated, which opens a problem when translating cyrillic or arabic scripts, cf. (Virga and Khudanpur, 2003).

spelling (to select for German *Gemse* vs. *Gämse*), and several other options.

These settings require user interaction, and a level of user skills which often is not available. Also, MT systems linked to search engines do not even ask users for subject area settings.

2. **Linguistic context** description. Such descriptions are coded in the dictionaries as transfer tests; they describe linguistic contexts which trigger the transfer selections:

See (gender = <feminine>) -> sea
See (gender = <masculine>) -> lake
ausführen (dir. object = <person> ->take out
ausführen (dir. object=<program>) -> execute

Such tests can be described as configurations of feature settings of underspecified tree structures[7]. Translation candidates are compared, in a specific order, to the input trees, and if their test configuration matches the input tree configuration then this translation is picked.

Such a technique has two problems to solve:

- In case of parse failures, the structures with which the transfer candidates are compared are erroneous, so the comparison may fail, and a poorer translation is selected
- There are many cases of underspecification, i.e. the information which would trigger a transfer selection is not present: In cases where
  de *Bank* (plural *Bänke*) -> en *bench / benches*
  de *Bank* (plural *Banken*) -> en *bank / bank*
  but the sentence contains only a singular (*er steht vor der Bank*), then the system cannot apply the test, and randomly has to pick a translation, which can be wrong.

Both options, parameter setting and linguistic tests, obviously need improvements in translation selection. For the parameter settings, an obvious solution is to set such parameters automatically.

3.2. Automatic subject area selection

To overcome the problem that not even the options which can be provided by the system (especially subject area selection) are used, a topic identification component has been added to the MT system, to compute to what subject area a text would have to be assigned.

1. There are two main lines of **technology** to build topic identification, or text classification, systems (Jackson and Moulinier, 2002): Selecting classification features (usually words) from an example corpus by machine learning techniques, or using manually selected key words describing the respective topic. While the former crucially depends on the similarity of test corpus and runtime text material, and therefore is less robust, the later depends on a careful selection of key words and tends to have a too small keyword basis. An e.g. in context where an MT system must translate internet material, the selection of a corpus which would be sufficiently similar to the texts to be translated at runtime is a very challenging task, so the second option seems to be preferable.

2. In an MT environment, the most plausible option seems to use the system **dictionary** as a resource for text classification. But although dictionaries are sensitive for subject area selection, they follow a different purpose:

- They use subject area tags only in cases where disambiguation is needed; and for 1:1 translations such a tag assignment often does not need to be assigned, as the respective translation is selected anyway. For a classification tool, however, this is a drawback.
- Also, there are subject areas containing only very few terms (again only the ones which need to be disambiguated), which is not suitable for good classification either.

So, although MT dictionaries can be a good starting point, more intelligence is required.

3. Therefore, a different approach was taken: A large text corpus was searched, starting with some **seed terms** (like '*sports football hockey racing*'); the system returned the highest correlated terms (both single and multiwords) to the seed words, using standard retrieval technique. From the resulting terms, the experts selected the ones which they believed to describe the topic best, and repeated this procedure. For each of the about 40 topics, between 400 and 1500 terms per language were collected to describe it.

These terms were processed with statistical classification tools to compute their relative importance related to the topic in question.

The classification is implemented in such a way that it gives the best (or the several best) subject areas if they match a given threshold, and gives no indication if it is not sure, and leave it to the users to decide; we felt that false assignments would do more harm than no assignment.

4. The **evaluation** of the component shows ambivalent results.

a. For a test corpus of several hundred documents in two languages, the correct subject area was identified in over 80% of the cases, and no false positives were returned. This is quite acceptable.

b. However, correct subject area recognition is just a prerequisite for proper selection of translation alternatives by the MT system. It depends on the organisation of the dictionaries what use of this information the system can make, and how sensitive it is to subject area coding. The result here was that the improvement was not really overwhelming, even if the classifier works fine[8].

During the evaluation, it also turned out that a subject area code rather means that a given translation alternative is unlikely outside of a certain subject area, but it does not mean that within a subject area this translation is always correct. Many general vocabulary terms occur in specific domains both with their special and their general meaning, like (in the automotive domain):

en *project* -> de *Restaurierungsobjekt* vs. *Projekt*
de *Übersetzung* -> en *gear ratio* vs. *translation*

As a result, a subject area test, even if the subject area is recognised correctly, is not the most helpful information for transfer selection; additional means need to be used.

3.3. Neural transfer

[7] An attempt to define a kind-of-standard representation for this has been made in OLIF, cf. (McCormick, 2001)

[8] This, of course, depends on the organisation of the MT dictionary, and may be different in the different systems.

Beyond improving global settings, the linguistic criteria for transfer selection should also be extended.

1. When observing **human behaviour** in transfer selection, it can be seen that people often refer to the conceptual context, to explain that 'even in the automotive domain, '*Übersetzung'* in the context of '*documentation'* and '*language'* and other such terms can only be '*translation'*, not '*gear ratio*''. The question is if such human behaviour can be modelled in an MT system to improve transfer selection using conceptual context.

The task is similar to word sense disambiguation, but applied not to abstract word senses (as in WordNet) but to concrete word senses as represented in different translations. It requires the identification of conceptual contexts which indicate a certain word sense, and consequently a certain translation of a term.

2. As a consequence, all **dictionary entries** with more than one translation were evaluated, and 'clear' cases like

en $teacher_{masculine}$ -> de *Lehrer*
en $teacher_{feminine}$ -> de *Lehrerin*

were eliminated. From the remaining set, several hundred candidates were selected for further analysis. Each of them was looked up in a standard dictionary to make sure that the most important readings of the term were represented.

3. For each term, a **corpus lookup** was done, using the linguatec corpus, resulting in a couple of thousand contexts per term. Each of these contexts was assigned a reading of the word in question, to enable the formation of clusters of concepts for each reading. These clusters were then statistically analysed, using a standard Bayesian classifier, to identify the most distinctive terms for a given reading, and represented as a neural network[9].

4. Examples of the effect are shown in the following texts, for different translations of *fan* and of *coach* into German *Fan* vs. *Ventilator* and *Trainer* vs. *Bus*, respectively:

(1) en *The fans make noise. The whole club was already drunk when they came to the stadium to support their soccer heroes, although their coaches had to leave.*
=> de *Die Fans machen Lärm. Der ganze Klub war schon betrunken, als sie zum Stadium kamen, um ihre Fußball-helden zu unterstützen, obwohl ihre Trainer abfahren mussten.*

(2) en *The fans make noise. Their rotor does not distribute the air evenly, and the electric motor is not in full operation. All the coaches full of tourists were disappointed.*
=> de *Die Ventilatoren machen Lärm. Ihr Rotor verteilt die Luft nicht gleichmäßig, und der elektrische Motor ist nicht in vollem Betrieb. All die Busse voll von Touristen waren enttäuscht.*

The first sentence is translated differently in the two contexts, although both times identical in the source language. Sentence-based translation is not able to grasp the difference.

5. The next task was the **integration** of the neural networks into the MT system. There are two challenges:

- Like in proper name recognition, neural transfer needs more context than just a sentence; systems with a only sentence-based architecture create artificial limitations. More context is required.
- The neural transfer must be integrated into the transfer selection architecture of the MT systems, and be related to the other transfer selection criteria.

5. The component was **evaluated** as follows: In the German-to-English system, 30 concepts were randomly selected for the tests, and texts containing these concepts were downloaded from the internet, without reading disambiguation. The texts contain 165 occurrences of the test concepts. These sentences were translated, and the result was compared.

Of those, 162 (98%) were correctly translated, using neural transfer. Without neural transfer, just 92 (56%, which is close to random) were correct, so there is an improvement in quality of more than 40%.

Of course the real quality gain depends on the frequency of such concepts in the complete corpus.

4. Conclusion

These examples show that the quality of MT systems is not yet at its limits; it also shows that it will develop in an evolutionary process rather than in a completely new technology.

The most promising approach seems to consist in hybrid system architectures, enriching rule-based approaches (which model the language competence) by corpus-based and statistical techniques (modelling the language performance aspects) as presented above.

5. Acknowledgements

The work presented here was mainly carried out by Vera Aleksić (who also composed the story), Alois Baumer and Thilo Will.

6. References

Babych, B., Hartley. A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. Proc. EACL-EAMT, Budapest.

Dillinger. M. (2001). Dictionary Development Workflow for MT: Design and Management. Proc. MT Summit, Santiago de Compostela. Spain.

Frank, A., Hoffmann, Chr., Strobel, M., (2004). Gender Issues in Machine Translation. Univ. Bremen

Jackson, P., Moulinier, I. (2002). Natural Language Processing for Online Applications. Amsterdam (J. Benjamins).

Jiménez, M. (2001). Generation of Named Entities. Proc. MT Summit, Santiago de Compostela. Spain.

McCormick, S. (2001). The structure and content of the body of an OLIF v.2 File. www.olif.net

Thurmair, Gr. (2005). Hybrid architectures for Machine Translation Systems. Language Resources and Evaluation, 2005, 91-108.

Virga, P., Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Language Applications. Proc. SIGIR Toronto.

[9] This is why we call this kind of transfer 'neural transfer'.

The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators

Youcef Bey[1,2], Christian Boitet[2], and Kyo Kageura[3]

[1]Graduate School of Advanced Studies, NII.
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo,
101-8430, Japan
youcefb@grad.nii.ac.jp

[2]Laboratoire CLIPS-GETA-IMAG
Université Joseph Fourier
385, rue de la Bibliothèque.
Grenoble, France
Christian.boitet@imag.fr

[3]Graduate School of Education
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku.
Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp

Abstract

The aim of our research is to design and develop a new online collaborative translation environment suitable for the way in which the online volunteer translators work. In this paper, we discuss thus how to exploit collaborative Wiki-based technology for the design of the online collaborative computer-aided translation (CAT) environment TRANSBey, which is currently under development. The system maximizes the facilitation of managing and using existing translation resources and fills the gap between the requirements 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. Two major types of online volunteer translator communities can be identified (Bey, 2005):

(i) *Mission-oriented translator communities:* mission-oriented, strongly-coordinated groups of volunteers are involved in translating clearly defined sets of documents. Many such communities translate technical documentation of project like Linux documentation (Traduct, 2005), W3C (W3C, 2005), and Mozilla (Mozilla, 2005).

(ii) *Subject-oriented translator network communities:* individual translators who translate online documents such as news, analyses, and reports and make translations available on personal or group web pages. These groups of translators do not have any orientation in advance, but they share similar opinions about events (anti-war humanitarian communities, report translation, news translation, humanitarian help, etc.) (TeaNotWar, 2005).

The aim of our research is to design and develop a new online collaborative translation environment suitable for the way in which these online translators work, with a special focus on mission-oriented translator communities, but also taking into account the needs of subject-oriented individual translators. To achieve this aim, we decided to use a Wiki-system as a base technology for developing an online collaborative translation environment that facilitates the management and use of documents and linguistic reference resources. This paper discusses the basic requirements of online translators working in a collaborative environment and reports the system functions developed for satisfying these needs within the TRANSBey system that we are currently developing.

We introduce general volunteer translators' needs by describing and analyzing various existing communities. In the second section, we attempt to outline the basic functionalities of online Wiki technology and its advantages for constructing of an online computer-aided translation environment (CAT). In the last section, the main modules of TRANSBey are described.

2. Current stat of online volunteer translators

Many translator communities are currently involved in translating various types of documents in different formats. In W3C, 301 volunteer translators translate specification documents (XML, HTML, Web service, etc.) into 44 languages. Paxhumana (PaxHumana, 2006) is another community of volunteer translators who translate report documents into four languages (English, Spanish, French, German). The twos groups show basically the same behavior during the translation process. In general, translation is done using a stand-alone personal environments. In the process, translators do not use linguistic tools on the server from which they disseminate translated documents (Bey, 2005). They communicate with each other to avoid duplicate translations.

The function of this translation processes currently not only falls short of what can be achieved using current technology but also does not satisfy translators' potential needs. The major insufficiencies, among others, of existing collaborative translation environments are that (i) different file formats (e.g., DOC, HTML, PDF, XML) cannot be automatically dealt with in the translation environment, (ii) existing translated document pairs cannot be efficiently and systematically looked up within the overall community environment in the process of translating new related documents, and (iii) linguistic tools are not sufficiently provided. Existing CAT systems, on the other hand, do not address fully the functions required by collaborative environments. As stated, our aim is to develop a system that can fill the gap between volunteer translators requirements and the existing community-based translation environment as well as CAT systems.

The insufficiencies above can be filled by functions/modules that (i) unify and consistently manage document formats and versions so that they can not only

be consistently administered but also can be processed into recyclable units for future reference as translation memory (TM), (ii) integrate the rich online wysiwyg editing environment for direct document creation on the server with various linguistic reference lookup functions, and (iii) support multilingual content. If these functions/modules are integrated into (iv) basic online community management mechanisms, we would be able to further promote the activities of mission-oriented translator communities. Let us elaborate these points in the following paragraphs:

(i) Combining TM with document management

Most existing TM systems provide little or no support for document management and versioning (Bowker, 2002). However, keeping information or traces from original documents translated sentences is useful when translators look for the context of translation and could at least allow them to reconstruct documents from translated segments in TM. Translators would be able to use TM to search general information related to the context of documents (e.g., to find the latest text translated for a particular organization). We have adopted translation memory exchange (TMX) to support document structure and TM exchange. For unit detection, we have exploited the efficiency of LingPipe tools (LingPipe, 2006), witch deal with sentence-boundary detection and linguistic unit detection (e.g, named-entity detection). This tool can be extended to support additional languages and trained resources for more precision.

(ii) In-browser wysiwyg editor

Translators have shown interest in developing online translation editors that would allow multiple translators to share TMs and documents for translation. This is particularly appealing to freelance translators and useful for sharing translation.

(iii) Multilingual content support

Another improvements that is underway is extending CAT tools to support a wider variety of languages by using encoding methods such as Unicode (UTF-8) and designing new standards and filters to support a wider variety of file formats, including formats using tags (e.g., HTML and XML).

(iv) Collaboration for enhancing translation

In terms of more general developments, 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 TM, translation and various type of linguistic resources.

We explained the principal need of volunteer translators in the above sections, witch leads us to underline the basic and relevant functionalities of Wiki technology and its advantages and facilities for the overall design of TRANSBey in the next section.

3. Basic Wiki functionalities for online CAT environment

A Wiki environments allow users to freely create and edit web page content using any web browser. On the one hand, they have simple syntax for creating new pages and links between internal pages, and on the other hand, they allow the organization of contributions to be edited in addition to the content itself. Augar stated (Augar, 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".

Browser-based access means that neither special software nor a third-party webmaster is needed to post content. Content is posted immediately, eliminating the need for distribution. Participants can be notified about new content, and they review only new content. Access is flexible. In fact, all that is needed is a computer with a browser and Internet connection (Schwartz, 2004).

The most important feature of a Wiki technology is the open editing, which means that content is open for direct editing and direct dissemination of information. Among existing open Wiki environments, we have chosen XWiki, a Java-based environment with the following features (XWiki, 2006):

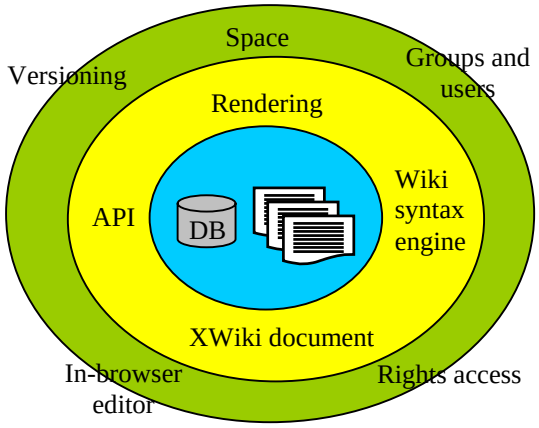


Figure 1: Main layers in XWiki

(i) Document management

Documents are managed in the core of an internal database where each entry is a document, which in the overall function of the environment is considered as the principal component or entries in the central databases. Documents are managed under the control of space and have a set of access rights for users. Documents can contain not only information to display (in HTML) but also Wiki syntax and programming codes for extending the system functionalities (Schwartz, 2004).

(ii) Versioning

Documents in the XWiki environment are supposed to be modified by direct editing. For any modification of the content, the system stores a new version in a new document and saves the old version for possible comparison or reuse.

(iii) Multilingual support

Translators are supposed to deal with content in several languages. Allowing the dissemination of translations on the web in several languages will motivate translator communities to use easily XWiki environment.

(iv) Space concept for volunteer communities

Organizing documents in space is very useful for communities and communication. Indeed, volunteers often work in an identified space, which they use it for sharing documents.

(v) Group (users) and access Rights

Access to XWiki content can be controlled. Users are identified via their IP addresses, and access can be limited for specific documents and functionalities.

Note that these features provide us with basic environment for collaborative translator communities. Within this overall environment, we have developed functions and modules specifically for translators, to which we now turn.

4. The TRANSBey prototype: integration of documents management and TM

4.1. Importing and processing source documents

Under the control of the uploader module that we have added to XWiki, documents (from source documents) can be uploaded directly into a unified format from various format types (PDF, DOC, RTF, HTML, etc.) or copied to source text areas in the in-browser editor (Figure 2). They are then stored in a unified format for document management with proper segmentation for recycling useful reference units.

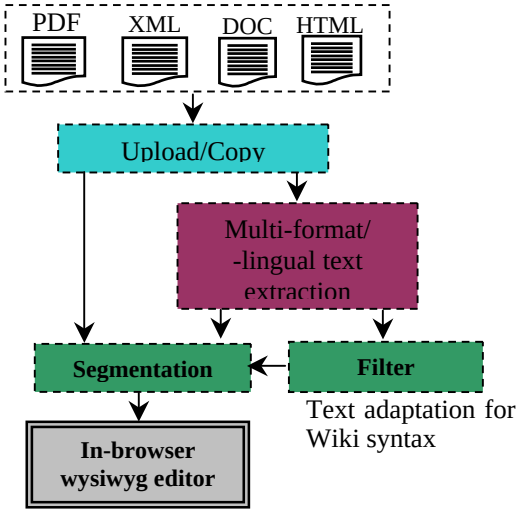


Figure 2: Monolingual document importation.

The extracted texts are segmented to logical translatable and linguistic units. The segmentation process is done semi-automatically[1], and translatable units are defined in the cores of textual sources.

4.2. TMX-C format for TM management

For the integrated management of documents and TM recycled from the documents, the document data structure should satisfy two requirements: (a) maximal facilitation of providing recyclable units and (b) unified management of translated documents. The first requirement come from individual translators, who strongly look for relevant linguistic units (especially collocations and quotations) in existing translations. The second requirement comes from the manager of the community in which translators take part or from the community itself. For this aim, we found that the translation memory exchange (TMX) standard is suitable (LISA, 2006). This standard was developed to simplify the storage/exchange of TM and to facilitate source/target sentences to be stored in a multilingual format in XML format(Bey, 2005)(Boitet, 2005).

Annotation is done in our environment in accordance with the TMX standard[2] and the 3-tiers level model proposed by Saha (Saha, 2005). The 3-tiers model for dealing with various information (Metadata annotation, sentences and linguistic unit annotations) is illustrated in Figure 3. The result of segmentation is an annotated document, which used to auto-construct TM and linguistic resources in the Wiki store.

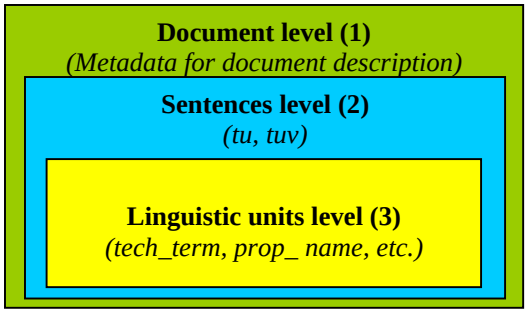


Figure 3: 3-tiers level for document segmentation.

Taking into consideration the advantages of the 3-tiers level model and the TMX standard capabilities, we have proposed TMX for Collaboration (TMX-C), witch is adapted for dealing with three levels during segmentation, for constructing the TM format, and for supporting collaborative Wiki information (Figure 4).

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <tmx version="1.4">
 - <header datatype="PlainText" segtype="sentence" adminlang="en-us" srclang="EN" creationdate="01/11/2005" creationid="AuthorA"
    changedate="03/01/2006" changeid="TranslatorB" o-encoding="iso-8859-1">
    <!-- Note: comments -->
    <note>Paxhumana is an humanitarian community of volunteer translators</note>
    <prop type="domain">Politic, Humanitarian</prop>
    <prop type="community_origine">Paxhumana Community</prop>
    <prop type="space">XWiki.Paxhumana</prop>
   </header>
 - <body>
  - <tu tuid="0001" datatype="Text">
     <prop type="Domain">Politic, Humanitarian</prop>
     <prop type="Project">Paxhumana Open Translation Community</prop>
     <prop type="Target_tu_order">1</prop>
    - <tuv xml:lang="en" creationid="AuthorA">
       <prop type="Source_XWiki_DocName">TortureSecret</prop>
       <prop type="Source_XWiki_DocSpace">Paxhumana</prop>
       <prop type="Source_XWiki_version">1.2</prop>
       <prop type="Source_location">http://paxhumana.info/article.php3?id_article=538</prop>
       <seg>I recently caught a glimpse of the effects of torture in action at an event honoring Maher Arar. The Syrian-born Canadian is
        the world's most famous victim of "rendition," the process by which US officials outsource torture to foreign countries...</seg>
     </tuv>
    - <tuv xml:lang="fr" creationid="XWiki.TranslatorB" changeid="Xwiki.TranslatorC">
       <!--          **** Document information: in the store of XWiki ********  -->
       <prop type="Target_XWiki_DocName">InfameTorture</prop>
       <prop type="Target_XWiki_DocSpace">Paxhumana</prop>
       <!--          **** versioning management ********  -->
       <prop type="Target_XWiki_version">1.4</prop>
       <!--      **** order of "tu" in the translated document ****      -->
       <prop type="Target_XWiki_tu_order">1</prop>
       <seg>J'ai récemment eu un aperçu en action des effets de la torture lors d'un événement en l'honneur de Maher Arar. Ce Canadien
        d'origine syrienne est la plus célèbre victime d'un genre d'extradition spécial appelé « restitution » [rendition], qui est un
        procédé par lequel les fonctionnaires des États-Unis sous-traitent la torture dans d'autres pays...</seg>
     </tuv>
    - <tuv xml:lang="it" creationid="XWiki.AuthorB" changeid="Xwiki.AuthorD">
       <prop type="Target_XWiki_DocName">infamesegreto</prop>
       <prop type="Target_XWiki_DocSpace">Paxhumana</prop>
       <prop type="Target_XWiki_version">1.4</prop>
       <prop type="Target_XWiki_tu_order">1</prop>
       <seg>Ho recentemente avuto un compendio in azione degli effetti della tortura durante un'avvenimento in onore di Maher Arar.
        Questo Canadese di origine siriana è la vittima più famosa di un genere di estradizione speciale chiamato
        "restituzione" [rendition] un procedimento con il quale i funzionari degli Stati Uniti subappaltano la tortura in altri paesi...</seg>
     </tuv>
   </tu>
 </body>
```

Figure 4: TMX-C format: Collaborative TMX-based for managing documents and TMs (PaxHumana, 2006).

At the top level, document information is provided, which is essential for document management but also useful for translators for checking the context and/or

[1] Translators have the ability in TRANSBey to annotate text in both source and target documents. The process is done (i) automatically by direct detection of translatable units before starting translation and (ii) by translators who delimitate translatable and linguistic units during translation.

[2] A standard proposed by the Localization Industry Standards Association (LISA) communities for TM support, exchange between humans specialists (or software) for more consistency, and decreased data loss.

domain to which documents belong (Table 1). The second and third levels are concerned with language units, i.e., sentences in the second level and various linguistic units (quotations, collocations, technical terms, proper names, idioms, etc.) in the third level (Table2). These units can be automatically detected using sentence-boundary tools (LingPipe, 2006) and other basic language processing tools, but translators can manually control these units in the process of editing and translation. The segments and metadata XML tags are defined as follows:

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.
Etc.	Etc.

Table 1: Metadata annotation tags.

TU/LU	Description	Format
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[3]	TMX
Etc.	Etc.	Etc.

Table 2: Translatable/linguistic unit annotation tags.

5. Online in-browser wysiwyg editor in TRANSBey environment

Editing source and target documents in an enhanced editor is the most important module that translators look for. Offering online editing in TRANSBey means also leading translators to edit in a rich environment that, among other functions, efficiently manage document formats, includes linguistic tools for accelerating translation and increasing quality, and avoids making translators become web developers, which is in general a hard task (which includes editing html code).

Among existing in-browser editors, we have chosen HTMLArea to integrate our environment for its many advantages (HTMLArea, 2006): (i) compatibility with almost all web browser (IE, Mozilla, Firefox) (ii) production of a well-formed HTML code (iii) ease of integration with XWiki for managing Wiki syntax (iv) several wysiwyg editing features (table, images, headings, etc.).

Figure 5 and Figure 6 illustrate how without any effort web documents in English can be imported from their original web sites to the TRANSBey environment for collaborative translation without losing their format and style presentations.

Figure 5: Source document in its original web site[4].

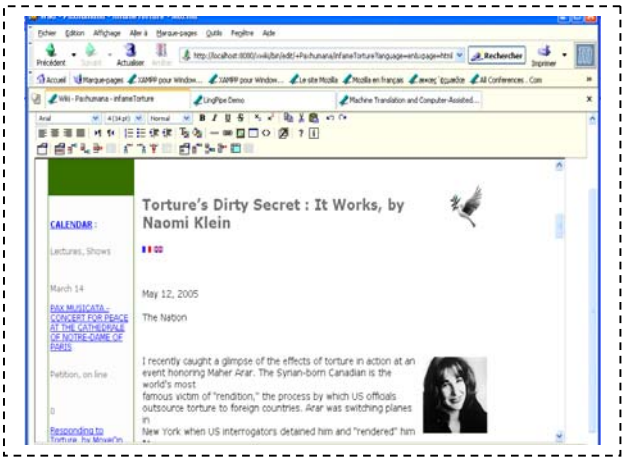


Figure 6: Imported document in TRANSBey environment[5].

Furthermore, using the integrated in-browser wysiwyg editor allows the same document to be produced in French without any modification of the source format during the translation (figure 7).

This example shows the feasibility of joining volunteer translators in comparable individual

[3]For further information about translation unit (TU) , translation unit version (TUV) refer to TMX standard (LISA, 2006).

[4]http://paxhumana.info/article.php3?id_article=538

[5]http://localhost:8080/xwiki/bin/view/+Paxhumana/TortureDirty (Wiki path for the imported document on the local server)

environments. The environment allows users easy HTML link navigation and gives them enhanced multilingual research functions for easily finding source/target documents and switching directly to the editing wysiwyg environment.

The integrated in-browser editor in XWiki is an open source, witch was developed separately by a group of volunteers called HTMLArea (HTMLArea, 2006). It contain the principal functionalities for editing and visual HTML component design (forms, tables, images, buttons, etc.). Furthermore, it manages well HTML/Wiki tag conversion and is compatible with IE and all Gecko web browsers (MOZILLA, FIREFOX, etc.).

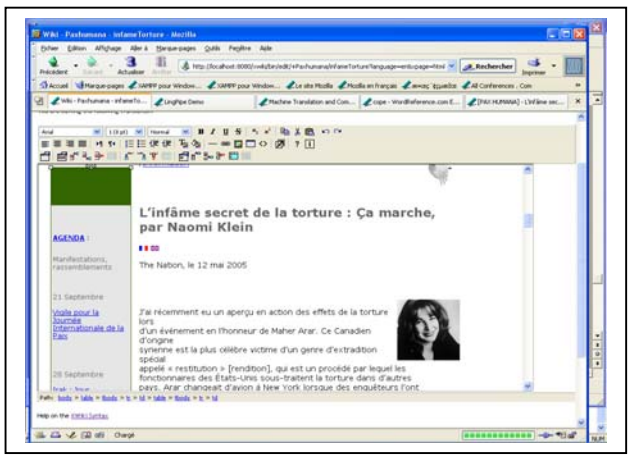


Figure 7: Target document after translation[6].

The editor, which is integrated into the collaborative environment and whose functions are currently under development, will be able to deal with different source texts in different formats in a unified framework while keeping the original format and can provide translators reference lookup and semi-automatic annotation based on TMX-C.

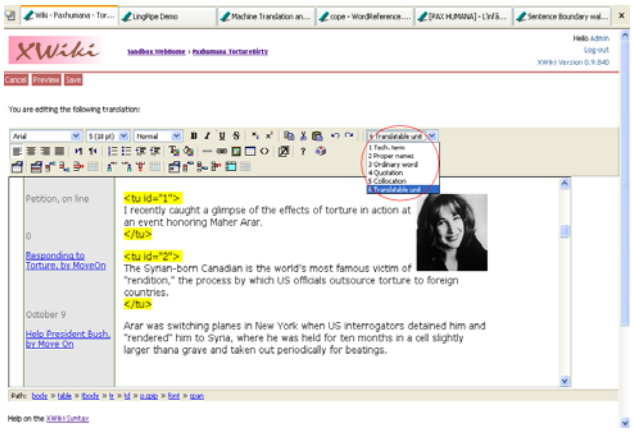


Figure 8: Wysiwyg edition and direct source annotation.

After translation is finished, source and target documents are recycled and translated segments and linguistic units are stored in linguistic resources for possible reuse for translating other documents (Figure 8).

6. Conclusion

We have proposed the TRANSBey prototype, an environment for helping volunteer translators produce high quality translations of various types of documents. This environment, which we are developing, will open a way for gathering skills and enhancing quality for all communities involved in translation. On the one hand, we used Wiki technology to exploit collaborative and open editing functionalities on the web; on the other hand, we have integrated the management of translatable units and linguistic resources using annotation system. Our aims for our environment are to offer to online volunteer translators important components for producing a quick translation with high quality in several languages.

In the near future, we are interesting for the enhancement of the integrated online editor for supporting synchronization and semi-automatic alignment between source and target documents for automatic TM construction, and integration during the translation process.

7. Acknowledgements

Special thanks go to Prof. Akiko Aizawa (National Institute of Informatics, NII, Japan) for many most valuable suggestions, advices and her invitation for the international internship at the NII.

8. References

Augar, N. , Raitman, R. and Zhou, W. (2004). Teaching and Learning Online with Wikis School of Information Technology. *In Proceedings of the 21st Australasian Society of Computers In Learning In Tertiary Education Conference.* Western Australia, Deakin University, Australia.

Bey, Y., Kageura, K. and Boitet, C. (2005). A Framework for Data Management for the Online Volunteer Translators' Aid System QRLex. *In Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation.* Taiwan, pp. 51-60.

Boitet, C., Bey, Y. and Kageura, K. (2005). Main Research Issues in Building Web Services for Mutualized, Non-Commercial Translation. *In Proceeding of the 6th Symposium on Natural Language Processing, Human and Computer Processing of Language and Speech*, Thailand.

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

HTMLArea. In-browser wysiwyg editor for XWiki. http://www.htmlarea.com/ (last accessed 02/02/2006).

Hutchins, J. (2003). Machine Translation and Computer-Based Translation Tools: What's Available and How It's Used. *Edited Transcript of a Presentation.* University of Valladolid, Spain. http://ourworld.compuserve.com/homepages/WJHutchins/ (last accessed 26/01/2006).

[6] http://localhost:8080/xwiki/bin/view/+Paxhumana/InfameTorture (Link to the French translation)

LingPipe. Linguistic Tools (Sentence-Boundary Detection, Named-Entity Extraction, Language Modelling, Multi-Class Classification, etc.). http://alias-i.com/lingpipe/demo.html (last accessed 10/01/2006).
LISA: Localization Industry Standards Association. Translation Memory eXchange. http://www.lisa.com/ (last accessed 25/01/2006).
Mozilla: French Mozilla Project. Open Software Localization. http://frenchmozilla.online.fr/ (last accessed 11/11/2005).
PaxHumana: Translation of Various Humanitarian Reports in French, English, German, Spanish. http://paxhumana.info (last accessed 30/01/2006).
Queens, F. and Recker-Hamm, U. (2005). A Net-Based Toolkit for Collaborative Editing and Publishing of Dictionaries. *Literary and Linguistic Computing Advance Access*. Oxford Journal. Lit Linguist Computing, pp. 165-175.
Radeox Engine. http://radeox.org/space/start (last accessed 15/12/2005).
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 (last accessed 26/01/2006).
Schwartz, L. (2004). Educational Wikis: Features and Selection Criteria. *International Review of Research in Open and Distance Learning*. Athabasca University, Canada's Open University. http://www.irrodl.org/content/v5.1/technote_xxvii.html (last accessed 27/01/2006).
TeaNotWar: Human Rights Documents Translation. English to Japanese News Translation. http://teanotwar.blogtribe.org/ (last accessed 20/12/2005).
Traduc Project. Linux Documentation Translation. http://wiki.traduc.org/ (last accessed 20/11/2005).
W3C Consortium. Specification Translation. http://www.w3.org/Consortium/Translation (last accessed 01/11/2005).
Walker, D.J., Clements D.E., Darwin, M. and Amtrup, W. (2001). Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality. *In Proceedings of the 8th Machine Translation Summit.* Spain.
XWiki. http://www.xwiki.com/ (last accessed 30/01/2006).

Corpógrafo – Applications

Belinda Maia & Luís Sarmento
PoloCLUP - Linguateca
Universidade do Porto
Faculdade de Letras
Via Panorâmica s/n
Portugal
E-mail: bmaia@mail.telepac.pt & las@letras.up.pt

Abstract

This paper will discuss how the Corpógrafo, a suite of on-line tools created by PoloCLUP of the Linguateca project (http://www.linguateca.pt) for the construction and analysis of corpora and the building of terminological databases, has been used for training professional linguists in corpora compilation, terminology extraction, terminology management and information retrieval. Reference will be made to the research which contributed to the development of the different tools that combine to make the suite usable, and examples will be given of the work possible using both the general language analysis tools and terminology and related data extraction tools.

1. Corpógrafo

The Corpógrafo is an on-line suite of tools for the creation and analysis of personal corpora and the creation of terminological databases that can be found at http://www.linguateca.pt/Corpografo. Although it was designed primarily for the study of terminology, translation and information retrieval, it also provides tools for the more general study of language. An individual or team may do their research independently in their own space on-line using these tools. The Corpógrafo tools are freely available on-line and anyone can sign in and start a personal or group project. Users receive access to the tools and a tutorial, but have to create the content of texts, corpora and databases.

The ideas for the Corpógrafo originated from a pedagogical idea in which special domain mini-corpora were created for the effect of teaching appreciation of text genre and register and the extraction of terminology (Maia, 1997). The creation of a branch of Linguateca, a project devoted to the Natural Language Processing of Portuguese, at the University of Porto led to the creation and implementation of technological tools to speed up this process, create integrated terminology databases, and permit the semi- automatic extraction of terms, definitions and semantic relations. The prototype was first described as the GC (Maia & Sarmento, 2003) and is now known as the Corpógrafo, now in its third version, (see Sarmento et al, 2006). We have always worked on the conviction that computer engineers and linguists have to work in harmony and that semi-automatic procedures, in which the computer programme accelerates the work of the human linguist is more satisfactory than either fully automatic or conventional human methods.

At present the Corpógrafo offers the following functions:

- Gestor (File Manager): the area where each individual or group can upload texts to the server, convert text formats like .doc, .html, .pdf, .ps, and .rtf texts to .txt, edit the texts, check for tokenization, chunk the text into sentences, register metadata on the text, and group texts into corpora.
- Pesquisa (Search): an area that allows for general corpus analysis, with tools for producing wordlists, n-grams and statistics, and studying words or phrases with sentence and KWIC concordancing which allows for sorting according to word position, as well as collocations and other phenomena.
- Centro de Conhecimento (Knowledge Centre): the area where terminology databases can be created and then linked to the corpora from which terms, definitions and semantic relations can be semi-automatically extracted. Term candidates are extracted automatically using an n-gram tool with filters to extract noun phrases from raw text. The terminologist then observes the list of term candidates, checks the term against the context of the underlying concordanced sentences, and clicks the term into the database. Each term automatically takes with it all the meta-data on the texts and corpora in which it appears if it has been previously registered.
- Centro de Comunicação (Documentation): the area where you can find a tutorial and news about the Corpógrafo as well as presentations and publications our group has produced.

2. Pedagogical applications

In the more specific environment of training at academic institutions, Corpógrafo has important pedagogical implications. Perhaps one of the most useful lessons students learn from all the technology we use in the using, making and analyzing of corpora is what they learn about the value of a corpus as a resource of information. They start by learning how to use large monolingual corpora like the British National Corpus and the Portuguese Linguateca corpus, CETEMPúblico, or a parallel corpus like the Linguateca Portuguese/English COMPARA, and they soon become enthusiastic about the advantages of corpora for providing solutions on usage and collocation that dictionaries do not offer.

Once the initial corpus linguistics methodology has been learnt, it is not difficult to build on this and encourage the compilation of corpora for a variety of uses, including terminology work. The various pedagogical exercises that are possible using Corpógrafo are very useful training as

they give a more rounded view of the theory underlying the commercial translation software, with translation memories, associated term databases and other tools, that they will use in the future as professional translators.

2.1 General corpus compilation and analysis

The exercise of constructing a corpus of any kind is important as a means of teaching students how to apply theories of genre and register in practice. Large, general purpose corpora are very useful for a wide variety of applications and research, particularly general lexicographical and language analysis. However, one often needs to work with small specialized corpora in order to study more specific aspects of different language varieties and lexicons. For this one needs to collect the texts in digital form and have access to concordancers and other language analysis tools. The Corpógrafo started out as a way of simplifying this process for the individual researcher.

Although the Corpógrafo has been developed primarily for work with special domain corpora, as we shall describe below, it is also possible to use it for other tasks, such as studying a specific author or genre. In these circumstances, the tools and methodology are those of normal corpus linguistic research.

The more general language analysis tools offered by the Corpógrafo effectively allow anyone to build their own corpus for their own personal project work, and we encourage people to do this and inform us of new ideas for improvement of this area. The work carried out under our supervision includes small individual projects in contrastive and corpus linguistics that have been varied and interesting. The Corpógrafo is often used for analysing specific lexical items or syntactic structures.

A typical piece of project work will take a lexical item that is difficult to translate, either due to its polysemous nature, with words such as *get*, *look*, and *issue*, or because they are closed system items like the adverbs *indeed*, *too*, and *just*, which rarely translate easily, or because they belong to lexical sets that do not easily find direct synonyms in the target language, such as the group *beautiful*, *handsome*, *pretty* and *good-looking*. The behaviour of these words are observed in monolingual and parallel corpora and small 'corpora' can be constructed out of the concordanced examples from these larger corpora for more minute and flexible analysis using the general language analysis tools in Corpógrafo. Similar work has been done with lexical bundles such as *I know that*, *I wonder if*, or any of the many examples in Biber et al (1999) as well as syntactic structures such as complex noun phrases or examples of the use of tense and aspect. The pedagogical objective of this type of work is to raise students' awareness of translation problems at a micro-linguistic level.

2.2 Corpora for terminology work

Most students in applied language or translation related courses come from a traditional language learning environment, and do not always find it easy to understand special domain texts. They tend to call the terminology 'jargon' and to consider the texts themselves boring. Restrictions on time usually mean that the translation teaching programme provides variety rather than subject depth, and 'terminology' is often little more than a short list of difficult words. As future translators, they sometimes ask why, when we can retrieve almost any information we need off the Internet, one should undertake the labour of building corpora for the extraction of terminology.

Clearly, in the everyday world of a professional translator, building corpora and terminology databases is apparently a luxury. However, in order to produce reliable terminology one needs good sources from which to extract information and, although the Internet contains a lot of good information, it also provides us with a good deal of rubbish. One of the objectives of the corpus and terminology building exercise is to teach the value of searching for and recognizing quality resources. As professional providers of language services now understand, proper investment of time and effort in reliable terminology means better quality control and results in the longer term.

Building special domain corpora with a view to extracting terminology encourages students to explore the domain in a certain depth and, in our experience, as the information becomes knowledge, curiosity to know more about the subject takes over. This type of exercise brings them closer to professional translation because it forces the student to become more familiar with the subject matter than is normal in most translation teaching. The exercise of choosing texts and analysing them in terms of genre and register is also useful for teaching them to find and imitate appropriate models in their own text writing or translation. They also learn to assess texts for their lexical quality and density, and consequent appropriateness for terminology extraction.

We recommend that beginners in the special domain start with encyclopaedia articles and then move on to pedagogical introductions to the subject, before including more complex texts like master's and doctoral dissertations, which usually include plenty of definitions and other relevant information. As the terminology database grows, keywords can be used to search for further appropriate texts, gradually leading to peer-to-peer publications for the extraction of more sophisticated or 'state-of-the-art' new terminology in the domain.

2.3 Extraction of Terminology, Definitions and Semantic Relations

Although there will always be a need for standardized terminology, for legal and simple administrative reasons, the emphasis in now on describing which terms are actually used in different contexts, as well on detecting the appearance of neologisms and/or mutation of terms. This information is essential for domain experts, translators and others who work with monolingual and multi-lingual documentation. The fast evolution of most technical and scientific knowledge makes it necessary to create more dynamic resources to cope with this phenomenon, and paper-based dictionaries and glossaries have given way to the terminology database.

It is very important to choose the texts for analysis by the Corpógrafo tools carefully. They will not find what is not in the texts that compose the corpus. However, once one

has a good corpus, the tools in the Centro de Conhecimento are of particular interest. The term extraction tool allows for n-grams to be filtered according to restrictions on the lexical items that can appear in proximity to possible term candidates. This tool functions for Portuguese, English, French, Italian, Spanish, and German, and we are working with the University Pompeu Fabra in Barcelona on Catalan. Although it produces a certain amount of noise, the recall is good and the human terminologist can select good term candidates and reject unacceptable ones very quickly, before submitting the results to the appreciation of the domain specialist for confirmation. The human labour of term extraction that could take months can thus be reduced to a few days.

The tools for extracting definitions and semantic relations depend on a bank of lexical patterns that is under constant development. The underlying theoretical approach is that of Pearson's (1998) 'terms in context', Partington's (1998) and Hunston & Francis's (1999) 'patterns', Biber et al's (1999) 'lexical bundles', and Hoey's (2005) 'lexical priming'. In practice, the task of finding the lexical patterns depends on combining computational expertise with human observation and analysis.

The terminology databases are conceived as essentially multilingual. This allows for terms to be extracted from the corpora in different languages and then linked within the database. The main database fields are typical of those used in terminology, but the pick-lists within them can be modified as and when the occasion arises. For example, the domain and sub-domain fields offered reflect the areas we are working on, but they can be added to on request. Also, although the more classical semantic relations are already part of the programme, researchers are encouraged to create their own as well. Experience has shown us that each domain reveals different types of semantic relation, as Sager (1990:30) demonstrates.

Once the more basic terminology has been extracted, it can be used to discover more specialized texts on the Internet. One can use a function that indicates the co-occurrence of terms in the different texts in the corpus, and this allows for further relevant texts to be found using normal Google-type searches. The tools are being improved on an on-going basis in order to provide further possibilities of extracting and structuring domain knowledge, creating further corpora and providing tools for more general information retrieval.

3. Research Applications

The Corpógrafo is being used for a variety of projects, many being prepared by people we do not even know. Here we shall concentrate on showing how the users have cooperated with us in its development, and refer to some of the projects with which it is being used.

The development of the Corpógrafo has resulted from working from an overall concept to the small details that make it workable. The process of trial-and-error that produced it is possibly as relevant to research methodology as the results themselves. Computer scientists and computational linguists clearly had a leading role, but the need to cooperate with general linguists, terminologists and translators forced them to contemplate the human + machine cooperation aspect. This attempt to create genuine understanding between two research groups which do not always work easily together was fundamental to the way the Corpógrafo developed and resulted in the coordination of the various tools and the user-friendly interfaces.

Much of the work done so far with the Corpógrafo has been experimental and has led to further improvement of the tools. The more general language work done in courses in contrastive and corpus orientated linguistics led to the way the concordancing tools developed, while the terminology work within master's degree projects was essential to the development of the terminology database. The compilation of the banks of lexical patterns and semantic relations has been carried out by research assistants and within the scope of masters' dissertations.

3.1 Terminology projects

The Corpógrafo has been very largely developed to deal with terminology projects, and version 3 now permits these to be carried out successfully, with the resulting databases being exportable in .xml for formatting in other programmes. We hope soon to develop tools for exporting the terminology data to a format that can be consulted on-line. It must be remembered that the existing system only allows consultation of the corpora, terms and other data by individual researchers, or by those individuals they authorise to consult their work.

For demonstration purposes there is a small project which is described in Portuguese on the site under the title 'Neurodemo'. This project started out as two small comparable corpora in English and Portuguese of about 25,000 words each on the subject of neurons created by an undergraduate student for a term paper. It now has comparable corpora on the same subject in five other languages, all of which have been used for the extraction of terms, definitions and semantic relations. The texts come largely from on-line popular science texts explaining neurons and have proved exceptionally useful for searching for information in a small well-defined area. The instructional nature of the texts provides the terms, as well as useful definitions and contexts from which the semantic relations between terms can easily be deduced by the human observer. The small size of the corpora in relation to their comparative success is proof that a well-selected corpus of texts is often more useful than a loosely constructed large corpus of only partially relevant texts.

There are several other on-going terminology projects that are not yet ready for publication, but we hope that they will be available in the near future.

3.2 Research for the improvement of the Corpógrafo tools

Most of the research done so far at dissertation level has involved the production and analysis of corpora, and the extraction of terms and other data as a method of testing and developing the Corpógrafo rather than for the production of full-scale databases. For example, the corpus analysis area has already proved useful for studying the instability of terminology in the fast developing area of GPS – Geographical Positioning

System (Brito, 2005). This is a study of concepts and how they are represented by several different terms, depending on who is using them and where. The Corpógrafo was used for the creation and observation of the corpus used, although our version of the terminology database at the time was not yet ready for the developments registered, and the terminology was created in another system. There is also a study of how concepts and their related terms have developed over decades in the field of Genetics and, although a much larger project is planned, (Fróis et al, forthcoming) shows how the concept behind one term has evolved seventy years, and how the expansion and subdivision of meaning within the concept has led to changes in usage of the original term and its expansion into various terms by the addition of adjectives to the original noun. These studies show how knowledge evolves and how terminology sometimes struggles to keep up with the pace of development and with the shifting concepts involved. They also show how a diachronic corpus can often be useful in explaining apparent inconsistencies in the evolution of the terminology of a certain area, and how different participants in the process contribute to the proliferation and confusion of terms.

Other work at dissertation level has also tested and provided incentive for further developments. One dissertation, by Almeida (to be defended) involves the testing and development of definition patterns in the domain of Natural Hazards. Having extracted definitions from corpora using the general language analysis concordancing function in the Corpógrafo, based on the ideas of Pearson (1998) and others, she tested her results against those obtained later using the bank of lexical patterns being built to support the Corpógrafo's definition extraction tool. Another dissertation by Jesus (to be defended) involves the building of networks of semantic relations in the area of Seismology. The resulting database should prove very useful in the development of the tool we are at present designing for the visualization of semantic networks. However, we cannot divulge further details until these dissertations have been defended.

4. Conclusions

The Corpógrafo is freely available online, which may partly account for its popularity. Sarmento et al (2006) supplies more details on who is using it and for what purposes. The original objective of producing pedagogical tools has been successful, with the users participating in the brainstorming over the development of the original tools, testing them and providing further ideas for improvement. Although we would not claim that the work done during this development has proved easy or perfect, we hope that the resulting Version 3 will soon prove its worth as a tool for more professional situations of terminology retrieval and management. However, the state-of-the-art of tools and resources in this area is moving fast and we recognize the need to refine the existing Corpógrafo and add to its potentialities in the future.

Acknowledgements

We should like to thank Linguateca, a distributed language resource center for Portuguese, for the opportunities offered to develop all the tools that are described here, and, more specifically, Diana Santos, Luís Cabral, and Ana Sofia Pinto, for all the work that has gone into their production. We should also like to thank the researchers who work with us for their ideas and for the research referred to here.

References

Almeida, A.S. (to be defended). *Pesquisa de Informação Terminológica: dos Marcadores Lexicais Aos Padrões Suporte: um Estudo no Dominio dos Riscos Naturais – Cheias*, Master's dissertation, Universidade do Porto.

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

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.* Universidade do Porto: Master's dissertation.

Fróis, C., B. Maia & A. Videira (forthcoming). 'A Case of Meaning Extension', in the *Proceedings of PALC 2005 – Practical Applications in Language and Computers*, University of Lódz, April, 2005.

Hoey, M. (2005). *Lexical Priming,* London/New York: Routledge.

Hunston, S., G. Francis (1999). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English (Studies in Corpus Linguistics)*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Jesus, C. de (to be defended) *Terminologia e Representação do Conhecimento do Domínio Específico da Geodinâmica Interna: Uma Abordagem ao Subdomínio da Actividade Tectónica*. Master's dissertation, Universidade do Porto.

Maia, B. & L. Sarmento (2003). 'GC - An integrated Environment for Corpus Linguistics'. Poster at CL2003: CORPUS LINGUISTICS 2003 - Lancaster University (UK).

Maia, B. (1997). 'Do-it-yourself corpora ... with a little bit of help from your friends'. In Lewandowska-Tomaszczyk, B. & P.J. Melia, (eds.) *PALC'97: practical applications in language corpora* (pp. 403-410), Lodz. Lodz University Press.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching (Studies in Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pearson, J. (1998). *Terms in Context*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Sarmento, L., B. Maia, D. Santos, L. Cabral, A. Pinto. (2006). "Corpógrafo V3 - From Terminological Aid to Semi-automatic Knowledge Engineering", in Proceedings of LREC 2006.

xml:tm - a radical new approach to translating XML based documents.

Andrzej Zydroń

CTO XML-INTL
PO Box 2167
Gerrards Cross
Bucks SL9 8XF
UK
azydron@xml-intl.com

Abstract

This paper describes the proposed xml:tm standard. xml:tm a revolutionary new approach to the problems of translating electronic document content. It leverages existing OASIS, W3C and LISA standards to produce a radically new view of XML documents: text memory. xml:tm has been offered to LISA OSCAR for consideration as a LISA OSCAR standard.

1. Translating XML documents

XML has become one of the defining technologies that is helping to reshape the face of both computing and publishing. It is helping to drive down costs and dramatically increase interoperability between diverse computer systems. From the localization point of view XML offers many advantages:

1. A well defined and rigorous syntax that is backed up by a rich tool set that allows documents to be validated and proven.
2. A well defined character encoding system that includes support for Unicode.
3. The separation of form and content which allows both multi target publishing (PDF, Postscript, WAP, HTML, XHTML, online help) from one source.

Companies that have adopted XML based publishing have seen significant cost savings compared with SGML or older proprietary systems. The localization industry has also enthusiastically used XML as the basis of exchange standards such as the LISA OSCAR TMX[1] (Translation Memory eXchange), TBX[2] (TermBase Exchange), SRX[3] (Segmentation Rules eXchange) standards, as well as GMX[4] (Global Information Management Metrics eXchange) set of proposed standards (Volume, Complexity and Quality). OASIS has also contributed in this field with XLIFF[5] (XML Localization Interchange File Format) and TransWS[6] (Translation Web Services). In addition the W3C ITS[7] Committee under the chair of Yves Savourel is working towards a common tag set of Elements and Attributes for Localization (Translatability of content, localization process in general etc.).

Another significant development affecting XML and localization has been the OASIS DITA (Darwin Information Technology Architecture) standard. DITA[8] provides a comprehensive architecture for the authoring, production and delivery of technical documentation. DITA was originally developed within IBM and then donated to OASIS. The essence of DITA is the concept of topic based publication construction and development that allows for the modular reuse of specific sections. Each section is authored independently and then each publication is constructed from the section modules. This means that individual sections only need to be authored and translated once, and may be reused many times over in different publications.

A core component of DITA is the concept of reuse through a well defined system for establishing a usable level of granularity within document components. DITA represents a very intelligent and well thought out approach to the process of publishing technical documentation. At the core of DITA is the concept the 'topic'. A topic is a unit of information that describes a single task, concept, or reference item. DITA uses an object orientated approach to the concept of topics encompassing the standard object oriented characteristics of polymorphism, encapsulation and message passing.

The main features of DITA are:

1. Topic centric level of granularity
2. Substantial reuse of existing assets
3. Specialization at the topic and domain level
4. Meta data property based processing
5. Leveraging existing popular element names and attributes from XHTML
6. The basic message behind DITA is reuse: 'write once, translate once, reuse many times'.

2. xml:tm

xml:tm[9] is a radical new approach to the problem of translation for XML documents. In essence it takes the DITA message of reuse and implements it at the sentence level. It does this by leveraging the power of XML to embed additional information within the XML document itself. xml:tm has additional benefits which emanate from its use. The main way it does this is through the use of the XML namespace syntax.

xml:tm was developed by XML-INTL and donated to the LISA OSCAR steering committee for consideration as a LISA OSCAR standard. In essence xml:tm is a perfect companion to DITA - the two fit together hand in glove in terms of interoperability and localization.

At the core of xml:tm is the concept of “text memory”. Text memory comprises two components:

1. Author Memory
2. Translation Memory

3. Author Memory

XML namespace is used to map a text memory view onto a document. This process is called segmentation. The text memory view works at the sentence level of granularity – the text unit. Each individual xml:tm text unit is allocated a unique identifier. This unique identifier is immutable for the life of the document. As a document goes through its life cycle the unique identifiers are maintained and new ones are allocated as required. This aspect of text memory is called author memory. It can be used to build author memory systems which can be used to simplify and improve the consistency of authoring.

The following diagram shows the how the tm namespace maps onto an existing xml document:

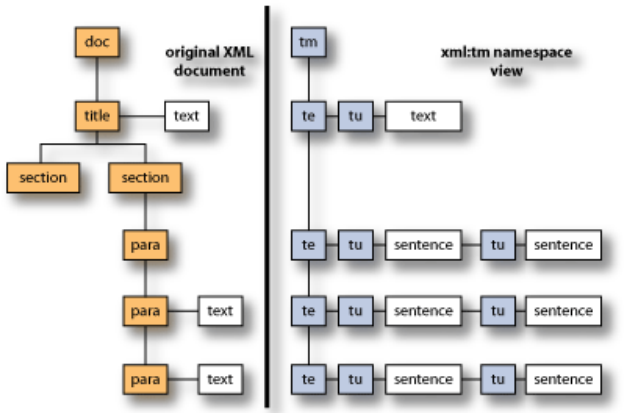


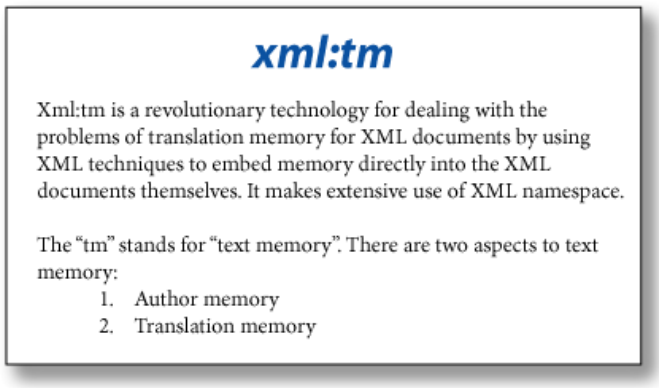
Figure 1. How xml:tm namespace maps onto an existing xml document.

In the above diagram "te" stands for "text element" (an XML element that contains text) and "tu" stands for "text unit" (a single sentence or stand alone piece of text).

The following simplified example shows how xml:tm is implemented in an XML document. The xml:tm elements are highlighted in red to show how xml:tm maps onto an existing XML document.:

```
<?xml version="1.0" encoding="UTF-8" ?>
<office:document-content
   xmlns:text="http://openoffice.org/2000/text"
   xmlns:tm="urn:xmlintl-tm-tags"
xmlns:xlink="http://www.w3.org/1999/xlink">
   <tm:tm>
     <text:p text:style-name="Text body">
       <tm:te id="e1" tuval="2">
           <tm:tu id="u1.1"> Xml:tm is a
revolutionary technology for dealing
          with the problems of translation
memory for XML documents by using
          XML techniques to embed memory
directly into the XML documents themselves.
</tm:tu>
           <tm:tu id="u1.2"> It makes extensive
use of XML namespace. </tm:tu>
       </tm:te>
     </text:p>
     <text:p text:style-name="Text body">
       <tm:te id="e2">
           <tm:tu id="u2.1"> The “tm” stands for
“text memory”. </tm:tu>
           <tm:tu id="u2.2"> There are two
aspects to text memory: </tm:tu>
       </tm:te>
     </text:p>
     <text:ordered-list text:continue-
numbering="false" text:style-name="L1">
       <text:list-item>
         <text:p text:style-name="P3">
           <tm:te id="e3">
               <tm:tu id="u3.1"> Author
memory</tm:tu>
           </tm:te>
         </text:p>
       </text:list-item>
       <text:list-item>
         <text:p text:style-name="P3">
           <tm:te id="e4">
               <tm:tu id="u4.1"> Translation
memory</tm:tu>
           </tm:te>
         </text:p>
       </text:list-item>
     </text:ordered-list>
   </tm:tm>
</office:document-content>
```

And the composed document:



xml:tm

Xml:tm is a revolutionary technology for dealing with the problems of translation memory for XML documents by using XML techniques to embed memory directly into the XML documents themselves. It makes extensive use of XML namespace.

The “tm” stands for “text memory”. There are two aspects to text memory:

1. Author memory
2. Translation memory

Figure 2. The composed document.

4. Translation Memory

When an xml:tm namespace document is ready for translation the namespace itself specifies the text that is to be translated. The tm namespace can be used to create an XLIFF document for translation.

4.1. XLIFF

XLIFF[5] is another XML format that is optimized for translation. Using XLIFF you can protect the original document syntax from accidental corruption during the translation process. In addition you can supply other relevant information to the translator such as translation memory and preferred terminology.

The following is an example of an XLIFF document based on the previous example:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE xliff PUBLIC "-//XML-INTL XLIFF-XML
1.0//EN" "file:xliff.dtd">
<xliff version="1.0">
  <file datatype="xml" source-language="en-USA"
target-language="es-ESP">
   <header>
    <count-group name="Totals">
     <count count-type="TextUnits"
unit="transUnits">40</count>
     <count count-type="TotalWordCount"
unit="words">416</count>
    </count-group>
   </header>
   <body>
     <trans-unit id="t1">
        <source> xml:tm</source>
        <target> xml:tm </target>
     </trans-unit>
     <trans-unit id="t2">
        <source> Xml:tm is a revolutionary
technique for dealing with the problems of
translation memory for XML documents by using
XML techniques and embedding memory directly
into the XML documents themselves.
        </source>
        <target> Xml:tm is a revolutionary
technique for dealing with the problems of
translation memory for XML documents by using
XML techniques and embedding memory directly
into the XML documents themselves.
        </target>
     </trans-unit>
     <trans-unit id="t3">
        <source> It makes extensive use of XML
namespace.
        </source>
        <target> It makes extensive use of XML
namespace.
        </target>
     </trans-unit>
     <trans-unit id="t4">
        <source> The “tm” stands for “text
memory”. </source>
        <target> The “tm” stands for “text
memory”. </target>
     </trans-unit>
     <trans-unit id="t5">
        <source> There are two aspects to text
memory: </source>
        <target> There are two aspects to text
memory: </target>
     </trans-unit>
     <trans-unit id="t6">
        <source> Author memory </source>
        <target> Author memory </target>
     </trans-unit>
     <trans-unit id="t7">
        <source> Translation memory </source>
        <target> Translation memory </target>
     </trans-unit>
    </body>
  </file>
</xliff>
```

The magenta colored text signifies where the translated text will replace the source language text as shown below:

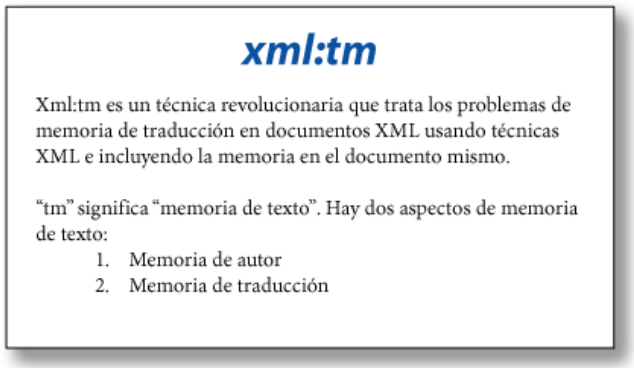
```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE xliff PUBLIC "-//XML-INTL XLIFF-XML
1.0//EN" "file:xliff.dtd">
<xliff version="1.0">
  <file datatype="xml" source-language="en-USA"
target-language="es-ESP">
   <header>
    <count-group name="Totals">
     <count count-type="TextUnits"
unit="transUnits">40</count>
     <count count-type="TotalWordCount"
unit="words">416</count>
    </count-group>
   </header>
   <body>
     <trans-unit id="t1">
        <source> xml:tm</source>
        <target> xml:tm </target>
     </trans-unit>
     <trans-unit id="t2">
        <source> Xml:tm is a revolutionary
technique for dealing with the problems of
translation memory for XML documents by using
XML techniques and embedding memory directly
into the XML documents themselves.
        </source>
        <target> Xml:tm es un técnica
revolucionaria que trata los problemas de
memoria de traducción en documentos XML usando
técnicas XML e incluyendo la memoria en el
documento mismo.
        </target>
     </trans-unit>
     <trans-unit id="t3">
        <source> It makes extensive use of XML
namespace.
        </source>
        <target> E sta técnica hace extensor uso
de XML namespace.
        </target>
     </trans-unit>
     <trans-unit id="t4">
        <source> The “tm” stands for “text
memory”. </source>
        <target> “tm” significa “memoria de
texto”. </target>
     </trans-unit>
     <trans-unit id="t5">
        <source> There are two aspects to text
memory: </source>
        <target> Hay dos aspectos de memoria de
texto: </target>
     </trans-unit>
     <trans-unit id="t6">
        <source> Author memory </source>
        <target> Memoria de autor </target>
     </trans-unit>
     <trans-unit id="t7">
        <source> Translation memory </source>
        <target> Memoria de traducción </target>
     </trans-unit>
    </body>
  </file>
</xliff>
```

When the translation has been completed the target language text can be merged with the original document to create a new target language version of that document. The net result is a perfectly aligned source and target language document.

The following is the translated xml:tm document in Spanish:

```
<?xml version="1.0" encoding="UTF-8" ?>
<office:document-content
   xmlns:text="http://openoffice.org/2000/text"
   xmlns:tm="urn:xmlintl-tm-tags"
xmlns:xlink="http://www.w3.org/1999/xlink">
   <tm:tm>
     <text:p text:style-name="Text body">
       <tm:te id="e1" tuval="2">
           <tm:tu id="u1.1"> Xml:tm es un
          técnica revolucionaria que trata los
problemas de memoria de
          traducción en documentos XML usando
técnicas XML e
          incluyendo la memoria en el documento
mismo. </tm:tu>
           <tm:tu id="u1.2"> E sta técnica hace
extensor uso de XML namespace. </tm:tu>
       </tm:te>
     </text:p>
     <text:p text:style-name="Text body">
       <tm:te id="e2">
           <tm:tu id="u2.1"> “tm” significa
“memoria de texto”. </tm:tu>
           <tm:tu id="u2.2"> Hay dos aspectos de
memoria de texto: </tm:tu>
       </tm:te>
     </text:p>
     <text:ordered-list text:continue-
numbering="false" text:style-name="L1">
       <text:list-item>
         <text:p text:style-name="P3">
           <tm:te id="e3">
                <tm:tu id="u3.1"> Memoria de
autor</tm:tu>
           </tm:te>
         </text:p>
       </text:list-item>
       <text:list-item>
         <text:p text:style-name="P3">
           <tm:te id="e4">
                <tm:tu id="u4.1"> Memoria de
traducción</tm:tu>
           </tm:te>
         </text:p>
       </text:list-item>
     </text:ordered-list>
   </tm:tm>
</office:document-content>
```

This is an example of the composed translated text:



xml:tm

Xml:tm es un técnica revolucionaria que trata los problemas de memoria de traducción en documentos XML usando técnicas XML e incluyendo la memoria en el documento mismo.

“tm” significa “memoria de texto”. Hay dos aspectos de memoria de texto:

1. Memoria de autor
2. Memoria de traducción

Figure 3. The composed translated document.

The source and target text is linked at the sentence level by the unique xml:tm identifiers. When the document is revised new identifiers are allocated to modified or new text units. When extracting text for translation of the updated source document the text units that have not changed can be automatically replaced with the target language text. The resultant XLIFF file will look like this:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE xliff PUBLIC "-//XML-INTL XLIFF-XML
1.0//EN" "file:xliff.dtd">
<xliff version="1.0">
  <file datatype="xml" source-language="en-USA"
target-language="es-ESP">
   <header>
    <count-group name="Totals">
     <count count-type="TextUnits"
unit="transUnits">40</count>
     <count count-type="TotalWordCount"
unit="words">416</count>
    </count-group>
   </header>
   <body>
     <trans-unit translate="no" id="t1">
        <source> xml:tm</source>
        <target state-qualifier="exact-matched">
xml:tm </target>
     </trans-unit>
     <trans-unit translate="no" id="t2">
        <source> Xml:tm is a revolutionary
technique for dealing with the problems of
translation memory for XML documents by using
XML techniques and embedding memory directly
into the XML documents themselves.
        </source>
        <target state-qualifier="exact-matched">
Xml:tm es un técnica revolucionaria que trata
los problemas de memoria de traducción en
documentos XML usando técnicas XML e incluyendo
la memoria en el documento mismo.
        </target>
     </trans-unit>
     <trans-unit translate="no" id="t3">
        <source> It makes extensive use of XML
namespace.
        </source>
        <target state-qualifier="exact-matched">
E sta técnica hace extensor uso de XML
namespace.
        </target>
     </trans-unit>
```

```
     <trans-unit translate="no" id="t4">
        <source> The “tm” stands for “text
memory”. </source>
        <target state-qualifier="exact-matched">
“tm” significa “memoria de texto”. </target>
     </trans-unit>
     <trans-unit translate="no" id="t5">
        <source> There are two aspects to text
memory: </source>
        <target state-qualifier="exact-matched">
Hay dos aspectos de memoria de texto: </target>
     </trans-unit>
     <trans-unit translate="no" id="t6">
        <source> Author memory </source>
        <target state-qualifier="exact-matched">
Memoria de autor </target>
     </trans-unit>
     <trans-unit translate="no" id="t7">
        <source> Translation memory </source>
        <target state-qualifier="exact-matched">
Memoria de traducción </target>
     </trans-unit>
    </body>
  </file>
</xliff>
```

4.2. Exact Matching

The matching described in the previous section is called “exact” matching. Because xml:tm memories are embedded within an XML document they have all the contextual information that is required to precisely identify text units that have not changed from the previous revision of the document. Unlike leveraged matches, perfect matches do not require translator intervention, thus reducing translation costs.

The following diagram shows how Exact Matching is achieved:

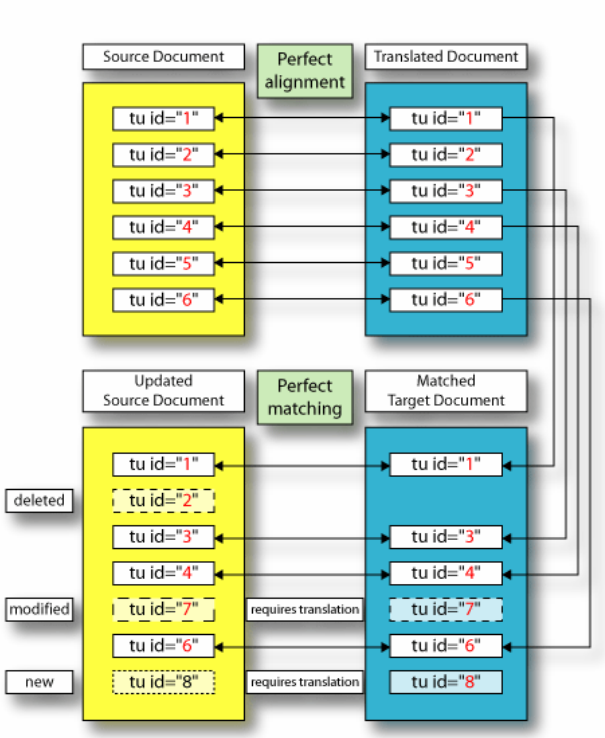


Figure 4. Exact Matching.

4.3. Matching with xml:tm

xml:tm provides much more focused types of matching than traditional translation memory systems. The following types of matching are available:

1. Exact matching

   Author memory provides exact details of any changes to a document. Where text units have not been changed for a previously translated document we can say that we have a “Exact match”. The concept of Exact Matching is an important one. With traditional translation memory systems a translator still has to proof each match, as there is no way to ascertain the appropriateness of the match. Proofing has to be paid for – typically at 60% of the standard translation cost. With Exact Matching there is no need to proof read, thereby saving on the cost of translation.

2. In document leveraged matching

   xml:tm can also be used to find in-document leveraged matches which will be more appropriate to a given document than normal translation memory leveraged matches.

3. Leveraged matching

   When an xml:tm document is translated the translation process provides perfectly aligned source and target language text units. These can be used to create traditional translation memories, but in a consistent and automatic fashion.

4. In document fuzzy matching

   During the maintenance of author memory a note can be made of text units that have only changed slightly. If a corresponding translation exists for the previous version of the source text unit, then the previous source and target versions can be offered to the translator as a type of close fuzzy match.

5. Fuzzy matching

   The text units contained in the leveraged memory database can also be used to provide fuzzy matches of similar previously translated text. In practice fuzzy matching is of little use to translators except for instances where the text units are fairly long and the differences between the original and current sentence are very small.

6. Non translatable text

In technical documents you can often find a large number of text units that are made up solely of numeric, alphanumeric, punctuation or measurement items. With xml:tm these can be identified during authoring and flagged as non translatable, thus reducing the word counts. For numeric and measurement only text units it is also possible to automatically convert the decimal and thousands designators as required by the target language.

5. xml:tm and other Localization Industry Standards

xml:tm was designed from the outset to integrate closely with and leverage the potential of other relevant XML based Localization Industry Standards.

In particular:

1. SRX[3] (Segmentation Rules eXchange)

   xml:tm mandates the use of SRX for text segmentation of paragraphs into text units.

2. Unicode Standard Annex #29[11] Text Boundaries

   xml:tm mandates the use of Unicode Standard Annex #29 for tokenization of text into words.

3. XLIFF[5] (XML Localization Interchange File Format)

   xml:tm mandates the use of XLIFF for the actual translation process. xml:tm is designed to facilitate the automated creation of XLIFF files from xml:tm enabled documents, and after translation to easily create the target versions of the documents.

4. GMX-V[4] (Global Information Management Metrics eXchange - Volume)

   xml:tm mandates the use of GMX-V for all metrics concerning authoring and translation.

5. DITA[8] (Darwin Information Technology Architecture)

   xml:tm is a perfect match for DITA, taking the DITA reuse principle down to sentence level.

6. TMX[1] (Translation Memory eXchange)

   xml:tm facilitates the easy creation of TMX documents, aligned at the sentence level.

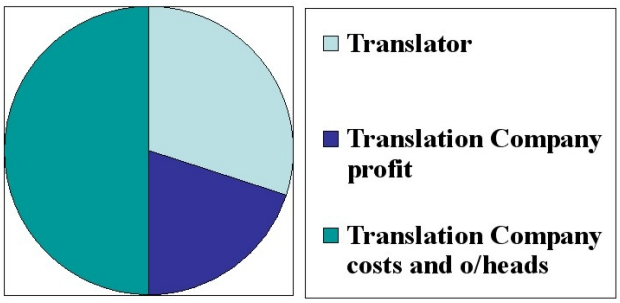
6. Controlling Matching and Word counts

You can use xml:tm to create an integrated and totally automated translation environment. The presence of xml:tm allows for the automation of what would otherwise be labour intensive processes. The previously translated target version of the document serves as the basis for the exact matching of unchanged text. In addition xml:tm allows for the identification of text that does not require translation (text units comprising solely punctuation or numeric or alphanumeric only text) as well as providing for in-document leveraged and fuzzy matching.

In essence xml:tm has already pre-prepared a document for translation and provided all of the facilities to produce much more focused matching. After exhausting all of the in-document matching possibilities any unmatched xml:tm text units can be searched for in the traditional leveraged and fuzzy search manner.

The presence of xml:tm can be used to totally automate the extraction and matching process. This means that the customer is in control of all of the translation memory matching and word count processes, all based on open standards. This not only substantially reduces the cost of preparing the document for translation, which is usually charged for by localization service providers, but is also much more efficient and cost effective as it is totally automated. The customer now controls the translation memory matching process and the word counts.

In a study conducted in 2002 by the Localization Research Centre the typical cost of the actual translation accounted for only 33% of the cost of localization for a typical project. Over 50% of the cost was consumed by administrative and project management charges. With xml:tm in an automated translation environment you can substantially reduce the costs of translation.



Source Professor Reinhard Schäler LRC - ASLIB 2002

Figure 5. The true costs of a traditional translation process.

The output from the text extraction process can be used to generate automatic word and match counts by the customer. This puts the customer in control of the word counts, rather than the supplier. This is an important distinction and allows for a tighter control of costs.

Traditional translation scenario:

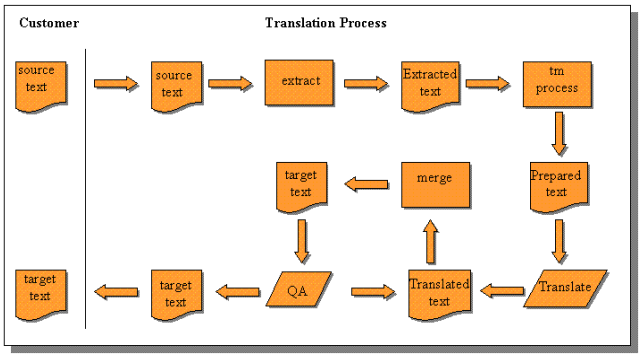


Figure 6. Traditional translation scenario.

In the xml:tm translation scenario all processing takes place within the customer's environment:

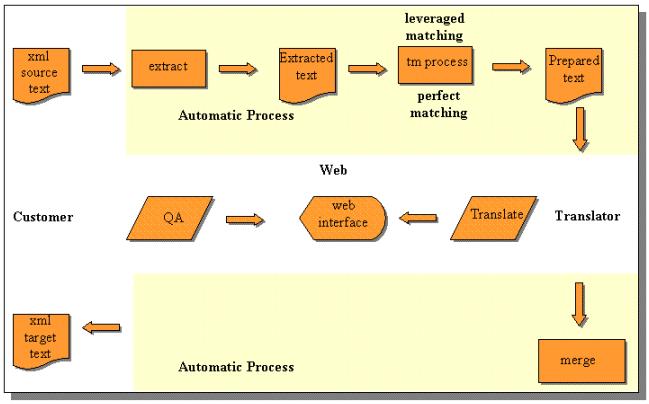


Figure 7. xml:tm translation scenario.

7. On line translation.

xml:tm mandates the use of XLIFF as the exchange format for translation. XLIFF format can be used to create dynamic web pages for translation. A translator can access these pages via a browser and undertake the whole of the translation process over the Internet. This has many potential benefits. The problems of running filters and the delays inherent in sending data out for translation such as inadvertent corruption of character encoding or document syntax, or simple human work flow problems can be totally avoided. Using XML technology it is now possible to both reduce and control the cost of translation as well as reduce the time it takes for translation and improve the reliability.

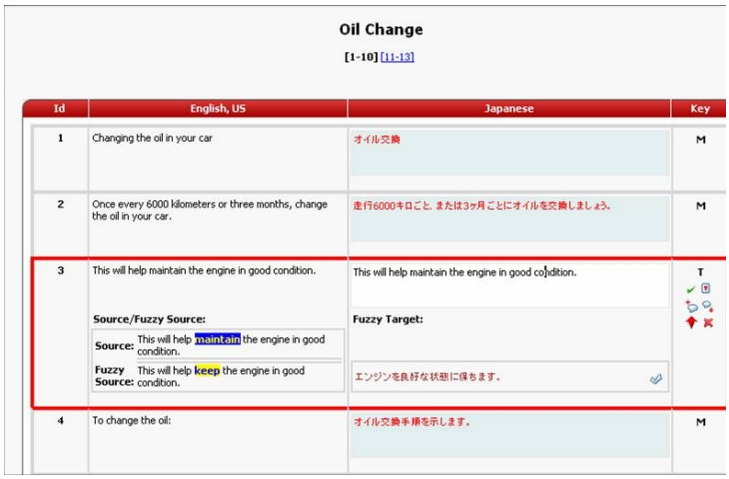


Figure 8. An example of a web based translator environment:

8. Benefits of using xml:tm

The following is a list of the main benefits of using the xml:tm approach to authoring and translation:

1. The ability to build consistent authoring systems.
2. Automatic production of authoring statistics.
3. Automatic alignment of source and target text.
4. Aligned texts can be used to populate leveraged matching tm database tables.
5. Exact translation matching for unchanged text units.
6. In-document leveraged and modified text unit matching.
7. Automatic production of word count statistics.
8. Automatic generation of exact, leveraged, previous modified or fuzzy matching.
9. Automatic generation of XLIFF files.
10. Protection of the original document structure.
11. The ability to provide on line access for translators.
12. Can be used transparently for relay translation.
13. An open standard that is based and interoperates with other relevant open standards (SRX[3], Unicode TR29[11], XLIFF[5], TMX[1], GMX-V[4]).

9. Summary

xml:tm is a namespace based technology created and maintained by XML-INTL based on XML and Localization Industry Standards for the benefit of the translation and authoring communities. Full details of the xml:tm definitions (XML Data Type Definition and XML Schema) are available from the XML-INTL web site (http://www.xml-intl.com).

The xml:tm approach reduces translation costs in the following ways:

1. Translation memory is held by the customer within the documents.

2. Exact Matching reduces translation costs by eliminating the need for translators to proof these matches.
3. Translation memory matching is much more focused than is the case with traditional translation memory systems providing better results.
4. It allows for relay translation memory processing via an intermediate language.
5. All translation memory, extraction and merge processing is automatic, there is no need for manual intervention.
6. Translation can take place directly via the customer's web site.
7. All word counts are controlled by the customer.
8. The original XML documents are protected from accidental damage.
9. The system is totally integrated into the XML framework, making maximum use of the capabilities of XML to address authoring and translation.

10. References

[1] TMX - Translation Memory eXchange format : *http://www.lisa.org/tmx/*

[2] TBX - TermBase eXchange format : *http://www.lisa.org/tbx/*

[3] SRX - Segmentation Rules eXchange format : *http://www.lisa.org/oscar/seg/*

[4] GMX - Global Information management Metrics *: http://www.lisa.org/standards/gmx/*

[5] XLIFF - XML Localisation Interchange File Format *: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff*

[6] Translation Web Services *: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=trans-ws*

[7] W3C ITS *: html/www.w3.org/International/its*

[8] DITA - Darwin Information Technology Architecture *: html/www.oasis-open.org/committees/tc_home.php?wg_abbrev=dita*

[9] xml:tm - detailed specification *: http://www.xml-intl.com/docs/specification/xml-tm.html*

[10] The Localisation Research Centre (LRC) *: http://www.localisation.ie/*

[11] Unicode Standard Annex #29 *: http://www.unicode.org/reports/tr29/*