LREC Workshop #8

Language Resources for Translation Work and Research

Programme

09:00 - 09:30	***Opening*** by Elia YUSTE, **Workshop Chair** (agenda - speakers introduction)
09:30 - 10:00	Silvia HANSEN & Elke TEICH, Computational Linguistics and Translation and Interpreting Departments (respectively), Saarland University, Saarbrücken, Germany *The creation and exploitation of a translation reference corpus*
10:00 - 10:30	Keynote Speaker - Maeve OLOHAN, CTIS, UMIST, Manchester, UK *Comparable Corpora in Translation Research: Overview of Recent Analyses Using the Translational English Corpus*
10:30 - 11:00	**Keynote Speaker** - Federico ZANETTIN, Università per Stranieri di Perugia, Italy *Corpora in Translation Practice*
11:00 - 11:20	Morning coffee break
11:20 - 12:00	Toni BADIA, Gemma BOLEDA, Carme COLOMINAS, Agnès GONZÁLEZ, Mireia GARMENDIA, and Martí QUIXAL, Universitat Pompèu Fabra, Barcelona, Spain *BancTrad: a web interface for integrated access to parallel annotated corpora*
12:00 - 12:30	Michael BARLOW, Department of Linguistics, Rice University, USA *ParaConc: Concordance software for multilingual parallel corpora*
12:30 - 13:00	Belinda MAIA, Faculdade de Letras, Universidade do Porto, Portugal *Corpora for terminology extraction - the differing perspectives and objectives of researchers, teachers and language service providers*
13:00 - 13:30	Lynne BOWKER, School of Translation and Interpretation, University of Ottawa, Canada *Working Together: A Collaborative Approach to DIY Corpora*
13:30 - 15:00	Lunch break
15:00 - 15:30	Elia YUSTE, Centre for Computational Linguistics, University of Zurich, Switzerland *Language Resources and the Language Professional*
15:30 - 16:00	Marita KRISTIANSEN & Magnar BREKKE, Norwegian School of Economics and Business Administration, Bergen, Norway *Textual and terminological bridgeheads for traversing the language gap*
16:00 - 16:40	Natalie KÜBLER, Intercultural Centre for Studies in Lexicology, University Paris 7, France *Creating a Term Base to Customize an MT System: Reusability of Resources and Tools from the Translator's Point of View*

16:40 - 17:00 Afternoon coffee break

17:00 - 17:30 Angelika ZERFASS, Language Technology Consultant, Germany
Evaluating Translation Memory Systems

17:30 - 18:00 Marie-Josée DE SAINT ROBERT, Chief, Terminology and Technical Documentation Section, Languages Service, United Nations Office at Geneva, Switzerland
Language resources at the Languages Service of the United Nations Office at Geneva

18:00 - 18:30 **Keynote Speaker** - Gerhard BUDIN, Department of Translation and Interpretation, University of Vienna, Austria
Global Content Management - challenges and opportunities for creating and using digital translation resources

18:30 - 19:00 ***Round-up Session***

Workshop Organisers and Programme Committee

[in alphabetical order]

Ms Elia YUSTE ***(Workshop Chair)***
Computerlinguistik, Institut für Informatik der Universität Zürich
Winterthurerstrasse 190
CH – 8057 ZÜRICH
Switzerland
yuste@ifi.unizh.ch

Dr Frank AUSTERMÜHL ***(Programme Committee Member and Main Adviser)***
Johannes Gutenberg-Universität Mainz
Fachbereich 23: Angewandte Sprach- und Kulturwissenschaft
Institut für Anglistik und Amerikanistik
An der Hochschule
D-76726 GERMERSHEIM
Germany
frank@austermuehl.de

Dr Gerhard BUDIN ***(Programme Committee Member and Keynote Speaker)***
Department of Translation and Interpreting Studies
University of Vienna
Gymnasiumstrasse 50
A-1090 VIENNA
Austria
gerhard.budin@univie.ac.at

Dr Maeve OLOHAN ***(Programme Committee Member and Keynote Speaker)***
Centre for Translation and Intercultural Studies
UMIST
PO Box 88
MANCHESTER M60 1QD
UK
maeve.olohan@umist.ac.uk

Dott. Federico ZANETTIN ***(Programme Committee Member and Keynote Speaker)***
Università per Stranieri di Perugia
Palazzo Gallenga - Piazza Fortebraccio, 4
I - 06122 PERUGIA
Italy
fz@federicozanettin.net
zanettin@unistrapg.it

Table of Contents

N.B. Papers are displayed here in the same order as they were presented on the Workshop day (please, refer to the **Programme** *above).*

Author Index

The creation and exploitation of a translation reference corpus

Silvia Hansen*, Elke Teich†

*Computational Linguistics, Saarland University
Postfach 151150, 66041 Saarbrücken, Germany
hansen@coli.uni-sb.de

† Applied Linguistics, Translation and Interpreting, Saarland University
Postfach 151150, 66041 Saarbrücken, Germany
e.teich@mx.uni-saarland.de

Abstract

While in many branches of linguistics monolingual reference corpora are widely used, in translation research as well as translation practice the concept of a translation reference corpus has not yet assumed a similarly important role. In this paper, we present the design of a German-English and French-English translation corpus and explore its use as a reference corpus for translatologists as well as translators. First, we introduce the basic computational techniques needed to build such a translation reference corpus, covering the preparation of the corpus as well as its linguistic annotation. Second, discussing some typical translation problems that occur in English-German and English-French translations, we show how the corpus can be queried making use of the linguistic annotation.

1. Introduction

In the last decade or so natural language corpora have assumed an increasingly important role in descriptive linguistics. Not only are they employed to inform lexicologists, lexicographers and grammarians in the construction of dictionaries and grammars, but also they gain importance as works of reference for linguists more generally. There are many corpora—especially for English (e.g., BNC[1], ICE[2], Bank of English[3])—that have been made accessible via the Internet with special user interfaces which allow one to query a corpus by means of KWIC concordances.

Also in translation research, corpora have started to become acknowledged as an important source of information in the investigation of theoretical issues in translatology, such as the question about the status of translations as a special kind of text with specific, possibly universal, properties. Here, the typical corpus is a parallel corpus consisting of two subcorpora, one containing source language (SL) original texts and the other containing translations of those texts into a target language (TL), where SL and TL texts are aligned (e.g., the Chemnitz corpora[4]). Some researchers advocate a three-way corpus design, where original texts in the TL are included as well (e.g., the Oslo corpora[5] as well as the work carried out at Saarbrücken (Teich & Hansen, 2001; Teich, 2001)), the latter being called a comparable corpus (cf. Baker, 1995; 1996). Also in translation practice, parallel corpora are increasingly being used in the form of translation memories. The compilation of such translation memories is supported by translation corpus workbenches. Thus, parallel corpora assume an increasingly important role both in theory and practice.

In this paper we explore the role of translation corpora as works of reference for translatologists as well as translators. It seems to us that there is a lacking interaction between the developers of corpus tools and researchers and practitioners in the field of translation. The goal of the present paper is to initiate such an exchange. We proceed in the following way. First, we discuss the basic computational techniques needed to make a corpus usable as a translation reference corpus (Section 2). We show how a corpus needs to be prepared (alignment, encoding) and how it should be enriched with linguistic information, so that it becomes possible to pose queries to it that are interesting and relevant from a translation point of view. Second, we show how a translation corpus can be queried with a parallel concordancing tool. We illustrate the use of an English-German-French translation reference corpus for solving some typical translation problems that occur in translating from English into German and from English into French (Section 3). Section 4 concludes the paper with a summary and some issues for future work.

2. Computational techniques

Corpus preparation. For the creation of a translation reference corpus, a parallel corpus needs to be aligned. For this purpose, an alignment program must be applied. One such program is Déjà Vu (Atril, 2000). Figure 1 shows a German SL and an English TL text aligned with this tool.

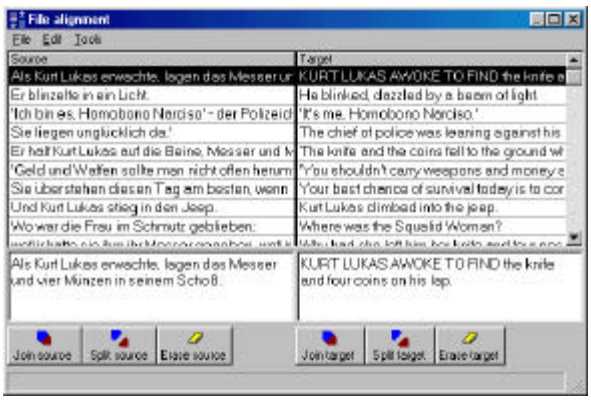


Figure 1: Multilingual corpus alignment

[1] http://sara.natcorp.ox.ac.uk/lookup.html
[2] http://www.ucl.ac.uk/english-usage/ice-gb/sampler/download.htm
[3] http:// titania.cobuild.collins.co.uk/form.html
[4] http://www.tu-chemnitz.de/phil/InternetGrammar/
[5] http://www.hf.uio.no/german/sprik/english/index.shtml

Déjà Vu aligns a text and its translation on sentence basis, storing the aligned texts in one file or in two separate files depending on the requirements of the query tool used in later stages of analysis. Files can be exported to translation workbenches and to Microsoft Excel and Access. Figure 2 shows a Déjà Vu output in a TSV (tab separated vector) format.

"Als Kurt Lukas erwachte, lagen das Messer und vier Münzen in seinem Schoß." "Kurt Lukas awoke to find the knife and four coins on his lap."
"Er blinzelte in ein Licht." "He blinked, dazzled by a beam of light."
"'Ich bin es, Homobono Narciso' - der Polizeichef stand an seinen Jeep gelehnt -, 'fast hätte ich Sie überfahren. Sie liegen unglücklich da.'" "'It's me, Homobono Narciso.' The chief of police was leaning against his jeep."
"Er half Kurt Lukas auf die Beine, Messer und Münzen fielen herunter, Narciso hob sie auf." "The knife and the coins fell to the ground when he helped Kurt Lukas up."

Figure 2: Déjà Vu alignment format

Also, we encode each text of the corpus in terms of a header that provides meta-information such as title, author, publication, translator, etc as well as text type/register information (domain, tenor and mode of discourse). This is important to enable corpus queries according to register or other independent variables.

Text files are encoded in XML using a modified version of the Text Encoding Initiative (TEI) standard[6] (a short header including meta-information is illustrated in Figure 3) and employing a standard XML editor (here: XML Spy[7]). The text body is annotated for headings, sentences, paragraphs, etc.

```
<tei.2>
  <teiHeader>
    <fileDesc>
      <filename>infanta_tl_e.txt</filename>
      <subcorpus>fiction (trans_en)</subcorpus>
      <language>English</language>
      <titleStmt>
        <title>Infanta</title>
        <author>
          <name>J. M. Brownjohn</name>
        </author>
      </titleStmt>
      <translation>
        <direction>German-English</direction>
      </translation>
      <sourceText>
        <title>Infanta</title>
        <language>German</language>
        <author>
          <name>Bodo Kirchhoff</name>
        </author>
      </sourceText>
    </fileDesc>
    <encodingDesc>Modified TEI</encodingDesc>
  </teiHeader>
  <text>
    <body> </body>
  </text>
</tei.2>
```

Figure 3: XML corpus encoding

Corpus annotation. A translation reference corpus should at least be annotated with part-of-speech and syntactic information. Part-of-speech tagging is carried out fully automatically, either using a rule-based or a statistical approach, where recently, statistical approaches prevail. For multilingual applications, it is important that the tagger can be used for more than one language. Analyzing a corpus in terms of syntactic structure is still a challenging task and cannot be carried out automatically with satisfactory accuracy yet. Recently researchers in computational linguistics who are interested in the accurate parsing of large amounts of text promote what has been called interactive parsing, where a parser carries out a shallow parse and a human may correct or add information to the proposed parse. For example, the parser assigns syntactic labels to the elements of a clause, but does not resolve syntactic ambiguities of particular kinds, such as PP-attachment, leaving this to the human to deal with.

One system which combines part-of-speech tagging and shallow parsing is the ANNOTATE system (Plaehn & Brants, 2000) under development in the TIGER[8] and NEGRA[9] projects. ANNOTATE uses the TnT tagger (Brants, 2000) that can be applied multilingually and has been trained on a number of languages, including English and German. The tag set used for English is the Susanne tag set (Sampson, 1995); the one for German is based on the Stuttgart-Tübingen tag set (Hinrichs et al., 1995). ANNOTATE carries out an analysis of phrase categories as well as grammatical functions using a program based on Cascaded Markov Models (CMM (Brants 1999a, 1999b)). During the interactive annotation with ANNOTATE (see Figure 4), terminal nodes are labeled for parts-of-speech and morphology, non-terminal nodes are labeled for phrase categories and edges are labeled for grammatical functions.

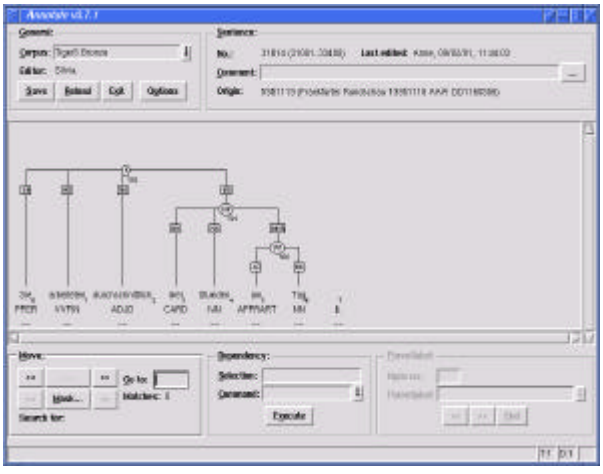


Figure 4: Interactive annotation with ANNOTATE

The tagged and parsed corpus data are stored in the form of a relational database, but can be exported to text format.

Corpus querying. For parallel concordancing, query tools such as the IMS Corpus Workbench (Christ, 1994)

[6] http://www.tei-c.org/index.html
[7] http://www.xml-spy.com
[8] http://www.ims.uni-stuttgart.de/projekte/TIGER/
[9] http://www.coli.uni-sb.de/sfb378/projects/NEGRA-en.html

can be employed. Its query processor (CQP) allows queries for words and/or annotation tags on the basis of regular expressions. For an example of a query executed on a parallel English-German corpus see Figure 5.

```
# Query: DE_EN; passives-de = [pos="VB.*"] [] {0,1} [pos="VVN.*"];
#------------------------------------------------------------------------------
   729: newspaper . A ferry had <been sunk> just off the island . ' I
  -->de_de: In den Gewässern vor der Insel war eine Fähre gesunken .
   850: country ' s future will <be decided> today . Yours too , perha
  -->de_de: Zukunft des Landes entscheidet sich heute .
   927: nced , because shots had <been fired> at a remote polling stati
  -->de_de: Der Schriftsteller und er müßten aufbrechen , in einem
```

Figure 5: Sample query with CQP

3. Solving translation problems with a translation reference corpus

With a corpus annotated in the way described in the preceding section, we now have available a translation resource that is searchable in a meaningful way. While with a raw text corpus we can only formulate string searches, we can now make use of the annotations in querying the corpus. In the following, we discuss some examples of translation problems between English, German and French. The examples are taken from two genres, narrative and factual writing. For querying the corpora selected, we use CQP (cf. Section 2).

English present and past perfect. While both English and German have present and past perfect tenses, their usage conditions differ cross-linguistically and it is sometimes hard to tell whether a one-to-one translation is the appropriate choice. The French tense system also has present and past perfect, but there are other options as well. Figure 6 shows two parallel concordances for English present and past perfect in narrative texts.

```
# Query: DE_EN; [pos="VH.*"] [pos="RR.*"] {0,1} [pos="VVN"];
#------------------------------------------------------------------------------
 509: night , he said . Adaza <had run> them off and was selling
-->de_de: Der Fotograf Adaza habe sie angefertigt und verkaufe sie für
 1120: igure and the blood that <had discoloured> a whole patch of grass
-->de_de: Die Fahrt endete vor einer Zwergschule , in der das Wahllokal
war , vor einer Blutlache , die ein ganzes Rasenstück färbte , vor einer
 2779: footsteps . Their guest <had appeared> on the terrace . Kurt Luk
-->de_de: Der Gast hatte auf die Terrasse gefunden .
 2953: ' Very few of our guests <have ever found> their way to this
-->de_de: ' Nur wenige unserer Gäste haben bisher auf diese Terrasse
gefunden

# Query: FR_EN; [pos="VH.*"] [pos="RR.*"] {0,1} [pos="VVN"];
#------------------------------------------------------------------------------
 1239: ver ventured there ; she <had even built> a low wall with her own
-->fr_fr: l ' épouse du pasteur avait même construit de ses mains un
 1395: sk , until the last rose <had dropped> into his open handkerchie
-->fr_fr: Il continua sa besogne , jusqu ' à ce que la dernière tête de rose
fût tombée dans son mouchoir ouvert .
 1478: , ' Do you realize what <has happened> to you ? When you
-->fr_fr: - Te rends-tu compte de ce qui vient de se produire en toi ?
 1499: ted Sheikh , and now you <have turned> into a thief ! I have
-->fr_fr: En arrivant ici ce matin , tu étais un cheikh respecté , et
maintenant tu es devenu un voleur !
```

Figure 6: Parallel concordances for English perfect

What can be seen here is that in translations into German, the translational choice is in fact often one-to-one, but also, past tense or present subjunctive is used. In the French parallel texts, we find direct translations, but also *passé anterieur* and "venir de".

English reduced relative clauses. Reduced relative clauses are a typical feature of English and French, but not so much of German. We can thus expect translational problems from English into German. A concordance query to a parallel corpus shows the translational options available (cf. Figure 7).

```
# Query: DE_EN; [pos="N.*"] [pos="VVN"];
#------------------------------------------------------------------------------
 197: g away under tin roofs . <Carcasses suspended> from chains
-->de_de: An Ketten hängend , bluteten zuckende Rinder aus . Schweine
 2180: ed behind on his own . A <crucifix reposed> on his lap in place of
-->de_de: An Stelle des Buchs lag ein Kreuz in seinem Schoß .
 2833: And the mountains wore <cloud-caps frayed> at the edges by
-->de_de: Und die Berge trugen Wolkenhüte , die zur Sonne hin
ausfransten .

# Query: FR_EN; [pos="N.*"] [pos="VVN"];
#------------------------------------------------------------------------------
1864: of him . This time , the <instrument provided> by Providence was
-->fr_fr: L ' instrument de la Providence fut cette fois un passe-temps
 2812: the presence of all the <people gathered> on the Blata , and in his
-->fr_fr: ' Le cheikh Francis et le patriarche se donnèrent l ' accolade
devant le peuple réuni sur la Blata , et dans son sermon , sayyedna parla
```

Figure 7: Parallel concordances for English reduced relative clauses

We see that English reduced relative clauses are indeed translated into French one-to-one (or zero-equivalent), whereas in German translations we find the present participle or full relative clauses (or zero-equivalent).

English cleft sentences. Cleft (and pseudo-cleft) constructions are a typical feature of the English grammatical system (cf. Erdmann, 1990). While they do exist in German as well, German has other options of realizing information distribution patterns, e.g., by word order variation. Because here, the search space for a translational choice is rather wide, finding a translational equivalent for an English cleft construction is therefore a notorious problem in translating from English into German. Again, a parallel concordance can provide help (cf. Figure 8).

```
# Query: DE_EN; [word="it|It"] [pos="VB.*"] [pos!="JJ.*"] {1,2}
[pos="DDQ.*|PNQ.*|CST"];
#------------------------------------------------------------------------------
 8620: simply as N , because <it is N that> makes this one-way function
-->de_de: Es ist dieses N, das die Einwegfunktion umkehrbar macht,
 8967: cells growing . <It is these properties that> make them attractive
-->de_de: Gerade diese Eigenschaften lassen sie als Wirkstoffe gegen
Krebs vielversprechend erscheinen.
 9112: is in control . <It is they alone that> persist from one generation to
-->de_de: Nur die Gene bleiben in der Generationenabfolge erhalten.
 9523: History records that <it was Galileo who> was foremost in
-->de_de: Die Geschichte belegt, daß vor allem Galilei die Zeit als eine
fundamentale Größe im gesetzesgleichen Wirken des Kosmos etablierte.
```

Figure 8: Parallel concordance for English clefts

The concordance shows that for compensation a focus particle or adverb (e.g., `gerade´) can be used to signal the syntactic focus.

4. Summary and conclusions

In this paper, we have suggested that translation corpora can assume the role of works of reference for translators and translatologists. In order for translation corpora to serve this purpose, they need to be enriched with linguistic information (Section 2). We have shown that some minimal linguistic annotation (part-of-speech, shallow phrase structure) can already make a translation corpus a valuable resource for dealing with some typical translation problems (Section 3).

While parallel concordancing tools operating on the basis of syntactic annotations already offer useful information, there are a number of further developments that can increase the value of a translation corpus. First, in corpus searches, it may be useful to be able to express constraints on the target language expression as well. Only few parallel concordance programs allow for this. Second, it could be very useful to be able to refer to a comparable TL corpus as well for a comparison of the translations with original TL texts. Third, for dealing with more complex kinds of translation problems, a translation corpus should be annotated with more abstract kinds of linguistic information, e.g., semantic and discourse information. This requires more comprehensive annotation methods and more sophisticated query facilities – both of which are current research issues in computational linguistics (cf. Teich et al., 2001).

Finally, from the perspective of the developers of corpus tools, translation corpora are an invaluable source for testing the applicability of such tools in multilingual contexts.

5. References

Atril, Development SL, 2000. *Déjà Vu. Productivity system for translators*. Software Manual. (http://www.atril.com/).

Baker, M., 1995. Corpora in translation studies: An overview and some suggestions for future research. In *Target* 7(2): 223-245.

Baker, M., 1996. Corpus-based translation studies: the challenges that lie ahead. In H. Somers (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam: Benjamins: 175-186.

Brants, T., 1999a. *Tagging and Parsing with Cascaded Markov Models - Automation of Corpus Annotation*. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 6, German Research Center for Artificial Intelligence and Saarland University.

Brants, T., 1999b. Cascaded Markov Models. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*. Bergen.

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

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

Erdmann, P., 1990. *Discourse and grammar. Focussing and defocussing in English*. Tübingen: Niemeyer.

Hinrichs, E., H. Feldweg, M. Boyle-Hinrichs, and R. Hauser, 1995. *Abschlußbericht ELWIS. Korpusunterstützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik*. Technical report, University of Tübingen.

Plaehn, O., and T. Brants, 2000. Annotate - An Efficient Interactive Annotation Tool. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000)*. Seattle.

Sampson, G., 1995. *English for the Computer*. Oxford: Oxford University Press.

Teich, E., S. Hansen, and P. Fankhauser, 2001. Representing and querying multi-layer corpora. In *Proceedings of IRCS Workshop on Linguistic Databases*. Philadelphia.

Teich, E., and S. Hansen, 2001. Methods and techniques for a multi-level analysis of multilingual corpora. In *Proceedings of Corpus Linguistics 2001*. Lancaster.

Teich, E., 2001. *Contrast and commonality in English and German system and text. A methodology for the investigation of the contrastive-linguistic properties of translations and multilingually comparable texts*. Habilitationsschrift (submitted for publication), Saarland University.

Comparable Corpora in Translation Research: Overview of Recent Analyses Using the Translational English Corpus

Maeve Olohan

Centre for Translation and Intercultural Studies
UMIST
PO Box 88
Manchester
M60 1QD
maeve.olohan@umist.ac.uk

Abstract

This paper discusses the use of a comparable corpus in translation research, where a comparable corpus comprises, on the one hand, a corpus of translations and on the other hand a corpus of non-translated texts, both corpora being similar in composition, size and other attributes. The Translational English Corpus, housed at the Centre for Translation and Intercultural Studies in Manchester, is presented as an example of a comparable corpus used in researching translation. The rationale for using a corpus of this kind to research translation is addressed. Results of a number of empirical analyses are then summarised, and the potential development and future exploitation of this corpus resource are outlined.

1. Corpora and Translation Studies

According to Michael Stubbs (2001: 151), corpus linguistics is concerned with "what frequently and typically occurs", as opposed to isolated, unique instances of language: "Corpus linguistics […] investigates relations between frequency and typicality, and instance and norm. It aims at a theory of the typical, on the grounds that this has to be the basis of interpreting what is attested but unusual". The corpus-based approach to studying translation has rapidly gained in popularity over the past eight to ten years, with a wealth of data now emerging from studies using parallel corpora, multilingual corpora and comparable corpora. In addition, corpora, whether of the ad-hoc or the reference kind, are proving a useful tool in the translator training classroom. Furthermore, most specialised translators would now be lost without their translation memory system, i.e. essentially an aligned parallel corpus of source texts and their translations.

This paper focuses on the first of these applications of corpora, namely corpora in translation research. The special issue of *Meta* on this topic published in 1998 is useful for an overview of work in this area, as is Chapter 3 of Kenny, 2001). Olohan (forthcoming b) highlights some of the strengths and limitations of corpus-based translation studies, based primarily on views put forward by Maria Tymoczko (1998) and Ian Mason (2001). This paper therefore does not present an overview of the literature nor does it address the criticisms levelled at corpus-based translation studies. Instead it assumes an understanding of corpus-based translation studies as the application of corpus analysis techniques, both quantitative and qualitative, to the study of aspects of the product and process of translation. Built into this is the recognition that there are differing opinions as to what aspects of translation we can apply these techniques to, and that the methodology requires refinement through application, discussion of findings and critical assessment. This process is now being undertaken by an ever-growing number of scholars in translation studies and it will ultimately lead to a better understanding of the scope, significance, usefulness and appropriateness (or not) of corpora to study translation processes and products.

2. Translation as Process and Product

The empirical study of the translation process emerged almost twenty years ago in translation studies, following on the heels of developments in second language research. It has since involved the identification, description and analysis of what happens during translation, i.e. of the mental steps taken by translators between, and including, reception of the source text and production of the target text. Introspection (in particular the think-aloud method) has been the principal methodological tool used in investigations of the translation process, and the introspective studies carried out to date have been largely data-based and descriptive, often focusing on specific aspects of the translation process (e.g. use of reference material, decision-making criteria). While a number of researchers have carried out descriptive empirical research in this area using the think-aloud method, there are methodological difficulties with research of this nature and, as a result, these attempts to investigate the cognitive processes at work during translation have met with scepticism from some quarters. Criticism has focused in particular on the methodology for data elicitation and collection, including its inability to provide access to thought processes which are subconscious or automated, but also on issues of scale and object of investigation.

While translation process researchers have readily acknowledged the potential shortcomings of this data elicitation method, it has been welcomed as a means of gaining some insight into something which is otherwise not accessible to the researcher. However, an alternative approach to translation process research is suggested by Bell (1991), who proposes that a model can and should be developed through a combination of induction (i.e. inferring processes from the product) and deduction (i.e. using introspective data such as diaries) (ibid.: 29). He suggests describing "translation competence in terms of generalizations based on inferences drawn from the

observation of translator performance" (ibid.: 39). He proposes to observe translator performance by analysing the translation product: "by finding features in the data of the product which suggest the existence of particular elements and systematic relations in the process" (ibid.). This approach lends support for the suggestion that the compilation and use of corpora of translations would allow us to analyse features of translation products which can provide evidence of translation processes, both conscious and subconscious, particularly if we can investigate "relations between frequency and typicality, and instance and norm", as advocated by Stubbs (2001: 151).

3. TEC – Translational English Corpus

TEC (Translational English Corpus) is a corpus of translated English held at the Centre for Translation Studies in Manchester. It consists of contemporary written translations into English of texts from a range of source languages and it was designed specifically for the purpose of studying translated texts. There are currently just under 7 million words in the corpus, made up of full running texts falling into four text types – fiction, biography, newspaper articles and in-flight magazines – with fiction representing more than 80% of the total. The translations are by native speakers of English, both male and female, and mostly date from 1983 onwards. In addition to the texts themselves, information is held on the translator and translation process, compiled via questionnaires to translators and publishers, and stored in header files.

One of the fundamental concepts in corpus-based translation studies has been the notion of comparable corpus, defined by Mona Baker (1995: 234) as "two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages…both corpora should cover a similar domain, variety of language and time span, and be of comparable length". Baker's initial groundbreaking work posited a number of features of translation which could be investigated using comparable corpora (Baker, 1996), for example, that translations tend to be more explicit on a number of levels than original texts, and that they simplify and normalise or standardise in a number of ways.

Much of the empirical analyses carried out thus far have focused on the literary component of TEC, namely fiction only, or fiction and biography. Thus, the corpus of original English put together for use as a comparable corpus is a set of texts selected from the imaginative writing section of the British National Corpus (BNC). It has been constructed specifically to match TEC in terms of both composition and date of publication (from 1981 onwards). As in the case of TEC, the BNC texts are produced by both male and female authors, all native speakers of English. Unlike TEC, however, some of the texts in the BNC subcorpus are extracts – albeit as long as 40,000 words. This was not deemed a significant difference in the current studies as they investigate intrasentential patterns. The Translational English Corpus is being added to all the time, which means that successive studies present data from TEC at different stages in its growth and the composition of the BNC subcorpus is modified accordingly.

Given that TEC and the BNC subcorpus are comparable in terms of parameters such as size and composition, features of the language of translation identified in the corpus of translation may thus be compared with features of non-translated language as found in the BNC subcorpus. Much of the work with TEC carried out to date has focused on syntactic or lexical features of translated and original texts which may provide evidence of the processes of explicitation, simplification or normalisation mentioned previously. It is possible to catch glimpses of these processes in think-aloud protocols where the translators are conscious of them and are employing them as part of controlled cognitive processes. However, corpus data may provide evidence which may constitute the result of such processes operating on a more subconscious level too.

4. Examples of Comparable Corpus Analyses

It is beyond the scope of this paper to present in detail the studies which has been carried out thus far using TEC and a BNC subcorpus. However, the results of some recent studies are summarised here, followed by an outline of some future directions for translation research using comparable corpora.

4.1. Optional Reporting *that*

The first large-scale empirical study using TEC and the BNC subcorpus indicated a substantially heavier use of the reporting *that* with verbs SAY and TELL in constructions such as examples [1] to [4] in TEC than in the BNC subcorpus, and it was suggested that this may be evidence for a tendency towards explicitation in translated English (Olohan and Baker, 2000).

[1] *He **says that** the ship is now forty-eight hours overdue and he wants explanations* (BNC)

[2] *He **says** the whole army is unsettled because it's known that Famagusta will never give up while it expects a relieving ship to arrive* (BNC)

[3] *I **told** him **that** I didn't know who it was he wanted to speak to, but he was quite insistent that he had seen you come in* (TEC)

[4] *I **told** him I thought it was a stupid thing for him to do* (BNC)

Explicitation has long been considered a feature of translation and has been investigated by a number of scholars (e.g. Vanderauwera, 1985, Blum-Kulka, 1986) who have identified different means or techniques by which translators make information explicit, e.g. using supplementary explanatory phrases, resolving source text ambiguities, making greater use of repetitions and other cohesive devices. In general, explicitation has referred to the spelling out in the target text of information which is only implicit in a source text. In these corpus-based studies, however, we are interested in the making explicit in a translation of information which is less likely to be made explicit in a non-translated text of the same language.

Scott Burnett (1999) examined the behaviour of some forms of other verbs of this type, and Olohan (2001) looked

at PROMISE, which can also take an optional *that*. The same pattern of heavier use of *that* in TEC compared with BNC was also found in these smaller-scale studies.

4.2. Other Optional Syntactic Features

Olohan (2001 and forthcoming a) presents a broad overview of some other optional syntactic features in English and their occurrence in TEC and the BNC. Since the focus of the research was subconscious processes of explicitation and their realisation in linguistic forms in translated texts, optional syntactic features were pinpointed, based on the hypothesis that, if explicitation is genuinely an inherent feature of translation, translated text might manifest a higher frequency of the use of optional syntactic elements than written works in the same language, i.e. translations may render grammatical relations more explicit more often – and perhaps in linguistic environments where there is no obvious justification for doing so – than authors in English.

Working with untagged corpora only, the analysis focused predominantly on frequency of occurrence of optional features and less so on the relationship between occurrence and omission. It can thus be regarded as a first step only. However, initial findings certainly encourage more detailed examination, suggesting for example that the use of the relative pronoun *which* is twice as frequent in TEC than in the BNC subcorpus. Similarly, a study of *who* (in the following constructions: *who is, who's, who've, who have, who'd, who did, who had* and *who would*) found that TEC has a significantly higher overall occurrence of the *who* form. Closer investigation of the co-text, which would be required to differentiate interrogative from relative usage, and to determine the optional vs. non-optional nature of the relative pronoun in each case, has not yet been carried out for all of these forms. However, in the case of *who is* and *who's*, a separation into interrogative and non-interrogative use showed that 44% of BNC occurrences were interrogative, as opposed to only 15% of TEC occurrences.

The occurrence of the complementiser *to*, which is optional following HELP, was analysed (see examples 5 and 6).

[5] *You have special skills and experience which will* ***help*** *us* ***to*** *achieve our objective.* (BNC)

[6] *She only wished Antonia were there with her to* ***help*** *her think over all the things Thomas said.* (BNC)

The data showed that although the word form *help* is more frequent in TEC, its verbal use in both corpora is quite similar. Of these verbal uses, the complementiser *to* is used in 37.5% of TEC instances, compared with only 26% of the BNC occurrences.

The use of *while* preceding a gerundial, i.e. *while *ing*,* and *after* preceding *having + participle* was measured in both corpora. *While *ing* was seen to occur more than twice as often in TEC than in BNC. A count of *after *ing *ed* (which obviously does not take irregularly formed past participles into account) also shows a tendency for TEC to use this construction more frequently than BNC, although the construction was relatively rare in both corpora.

Finally, *in order* may be omitted before *to* and may occasionally be omitted before *for* or *that*. While the investigation of every instance of the items *to, that* and *for* to see whether an *in order* has been omitted is not practical, it is possible to measure usage of *in order to, in order for* and *in order that* and compare results from the two corpora. This investigation showed a marked difference in usage of *in order to*, with 250 instances in BNC compared with 1,225 in TEC. The other forms, *in order for* and *in order that*, were infrequent in the two corpora but both occurred more often in TEC than in the BNC subcorpus.

4.3. Personal Pronouns

A small-scale study of the use of personal pronouns in both corpora is also presented in Olohan (forthcoming a). Frequencies of personal pronouns occurring with verb forms *will, have, am, is, has* and *are*, both within verb contractions and within non-contracted forms, were recorded. The data show that, when used in conjunction with these particular verb forms, personal pronouns *I, you, he, she, we* and *they* are more common in the BNC subcorpus than in TEC. The differences are extremely striking in the case of *I* (23,409 in BNC; 16,178 in TEC), and also quite marked in the case of *you, she* and *we*. The pronouns *he* and *they* occur with these verbs with almost the same frequency in the two corpora.

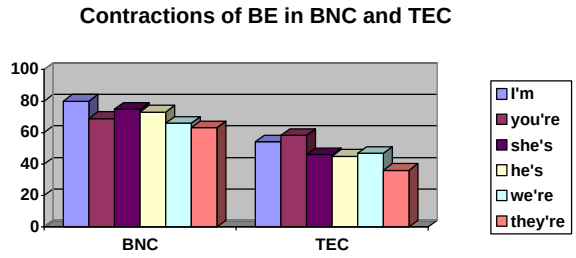
4.4. Contractions

As reported in Olohan and Baker (2000), the linguistics literature on use and omission of *that* with a range of verbs indicated that omission was more likely in informal contexts. Preliminary analysis of co-occurrence of *that* omission and contracted forms (as a crude measure of informality) revealed a definite correlation in both corpora between use of contracted forms and omission of *that.* Thus, despite lower incidence of contractions in TEC and higher incidence of that omission in BNC, the likelihood of co-occurrence of a contracted form and omission of *that* (in the same concordance line) was very similar in both corpora. In other words, the BNC texts were more likely to omit *that* and use contractions; the TEC texts were more likely to include *that* and not use contractions. This correlation suggested that contractions merited further investigation.

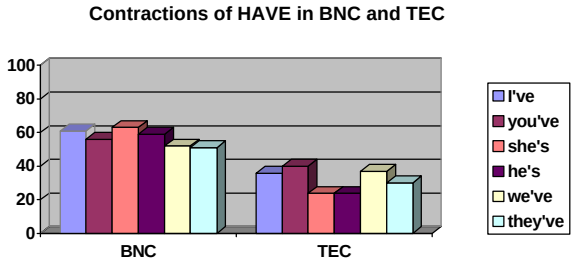
Further detailed analysis of all contracted forms in the corpora revealed that there are higher occurrences and a greater variety of contracted forms in BNC than in TEC. In many cases, the number of occurrences of a form in BNC is double that seen in TEC. (It is worth noting again at this point that the corpora under investigation are extremely similar in terms of size and composition.) In addition, there was a general preference for contracted forms over the corresponding long forms in BNC, while the TEC data showed a general tendency to use the long form in preference to the contracted one. For example, for all *'s* contractions (not including the possessive*'s*, thus for the following forms: *it's, that's, he's, there's, she's, what's, let's, who's, where's, here's, how's*), the contracted form is significantly more common than the long form in BNC. This is not true for TEC, where the long form is the more frequent in 8 out of the 11 forms. In TEC, the contracted form is more frequent only for *that's, what's,* and *let's*, but in these cases represents a smaller proportion of the

combined total occurrences of long and contracted forms than does the long form in BNC.

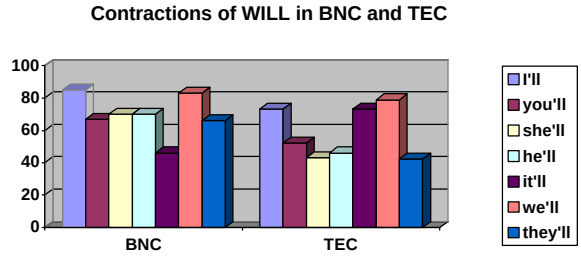
Splitting the analysis into verbs, we can see from Graphs 1, 2 and 3 that there is a greater incidence of contracted forms with personal pronouns in BNC than in TEC for present-tense forms of BE, HAVE and WILL.



Graph 1 Contractions of BE in BNC and TEC, represented as percentage of combined total for contracted and long forms



Graph 2 Contractions of HAVE in BNC and TEC, represented as percentage of combined total for contracted and long forms



Graph 3 Contractions of WILL in BNC and TEC, represented as percentage of combined total for contracted and long forms

As far as common *not*-contractions are concerned, the overall tendency in both corpora is to contract. However, the proportion of contracted forms is smaller in TEC than in BNC in all cases, and for 2 forms examined, *couldn't* and *wouldn't*, TEC is, in fact, more likely to use the long form. Biber et al. (1999: 1131) show that DO + *not* is contracted almost 100% of the time in conversation, around 75% in fiction, 60% in news text and 5% in academic text. From the data used in this study, on average across forms *don't, doesn't* and *didn't*, the rate of contraction of *not* with DO in BNC is 74%, thus very close to Biber et al.'s finding of 75% for fiction. In TEC, on the other hand it is 58%, thus considerably lower.

4.5. Dialectal features

Most of the contractions which featured in the analysis above were of verbs BE, HAVE and WILL or of the negation *not.* However, the BNC subcorpus had a selection of other types of contractions. Many are typical of spoken English, such as the contraction of multisyllabic modifiers e.g. *actu'lly, accident'lly, contradict'ry, prob'ly, fav'rite, gen'rous*. Some interjections also had contracted forms, e.g. *ah'm* and *fuck'em*, again characteristic of the spoken language, as were contractions of *ing* (e.g. *bleed'n*), *and* (e.g. *this'n*) and *than* (*better'n*). Some contractions were also clearly dialectal or sociolectal, with indicators of regional variations such as the dropped *h* in *be'aviour, be'ind, ware'ouse*, or the Scottish *does'na* and *hav'na* (where there is, in fact, no elision between the two words). There were 102 occurrences of *e's* in BNC (dialectal version of *he's*) and none at all in TEC. Finally, other forms found were *d'* (= *do*), *y'* (= *you*), *th'* (=*thou* or *thy*) and *t'* (= *to* or *to the*). All occur considerably more frequently in BNC than in TEC, e.g. *y'know* occurs 22 times in BNC and only once in TEC; *d'you* occurs 362 times in BNC, compared with 72 occurrences in TEC. The last two in particular indicate regional variation and do not occur at all in TEC; by contrast, *t'*, representing *to, to the* or *the* occurs in front of 99 different nouns or modifiers in BNC (see examples 7 and 8), and *th'* occurs 137 times (see example 9).

[7] *"It's a blessing it's a mild winter up ti now," he commented. "It would've been a bad time for* ***t'road*** *between* ***t'two*** *farms ti be blocked wi' snow."* (BNC)

[8] *"We're to go down* ***t'village****, to* ***t'stables****," George told his father, as he retrieved the reins.*(BNC)

[9] *"****Th'mind*** *what I say and* ***th'll*** *doubtless find there's no better place than Jarman House."* (BNC)

5. Directions of Future Research

The picture which emerges from these sets of data and the more detailed quantitative and qualitative analyses which have been done is one of a general preference for longer surface forms in TEC where there is an option between longer and shorter forms. This appears to apply as much to potential contractions of word forms as to syntactic explicitation of relations between clauses, for example in the use of the optional *that* with certain verbs or in the inclusion of relative pronouns where they are optional, i.e. in relative clauses where the co-referential NP is not the subject of the relative clause.

Furthermore, the tendency towards explicitation may extend to lexical choices, where some kind of repetition of nouns in translation may be preferred over use of pro-forms. In addition, TEC appears to contain a more standard variant of the English language, with fewer dialectal or sociolectal markers.

A tentative attempt has been made to link these findings with Biber's dimensions of English (1988 and 1995), with a view to determining to what extent TEC fiction is similar or different to the features of English fiction as analysed by

Biber. These preliminary findings seem to indicate that TEC fiction is not as typical of fiction in English as the works of fiction in the BNC subcorpus. Furthermore, some of the results suggest that TEC fiction may exhibit features more typical of academic prose in English. If this is borne out by future investigations it may contribute to an understanding of the nature of literary translation and its reception in the British literary system. However, there are many features to be investigated in the future to shed further light on this issue.

A criticism sometimes levelled at translation scholars is that we focus too much on literary text and literary translation. One area in which this research can be broadened is to add other genres to TEC. A subcorpus of non-fictional translated works of social science, politics, history etc. would provide an interesting contrast to the fiction subcorpus. Similarly, a bigger biography component would enable useful analyses of that genre to be carried out, taking into account in particular its position somewhere on the continuum between fictional and factual writing.

One aspect of research of this kind which has not been discussed in this paper is the investigation of individual translators. Due to the design of TEC and the incorporation of more than one translation by several translators, it is possible to compare translators and their practices; for example, Baker (2000) discusses the development of a methodology for investigating the style of a literary translator and Olohan (forthcoming b) examines the contraction patterns of two well-known translators across a number of translated works. There is much scope for further research of this kind.

At a conference workshop such as LREC where the emphasis is on practical application of technology in the translation process, one might question the relevance of this kind of detailed analyses of lexical or syntactic patterns in translated language. However, if studies of this nature ultimately give us a better understanding of how translators use language, i.e. how translators translate and what (cognitive) processes are involved, it will be of relevance, not just in the teaching of translation but also in the development of effective technological resources for translators in the future.

6. References

Baker, M. (1995) "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research", *Target* 7(2): 223-243.

Baker, M. (1996) "Corpus-based Translation Studies: The Challenges that Lie ahead", in H. Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*. Amsterdam and Philadelphia, John Benjamins, 175-186.

Baker, M. (2000) "Towards a Methodology for Investigating the Style of a Literary Translator", *Target*, 12(2): 241-266.

Bell, R. T. (1991) *Translation and Translating: Theory and Practice*, London and New York: Longman.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: CUP.

Biber, D. (1995) *Dimensions of Register Variation: A Cross-Linguistic Study*. Cambridge: CUP.

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

Blum-Kulka, S. (1986) "Shifts of Cohesion and Coherence in Translation", in J. House and S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr, 17-35.

Burnett, S. (1999) *A Corpus-based Study of Translational English*. Manchester: unpublished MSc dissertation, UMIST.

Kenny, D. (2001) *Lexis and Creativity in Translation: A Corpus-based Study*. Manchester: St. Jerome.

Mason, I. (2001) 'Translator Behaviour and Language Usage', *Hermes* 26: 65-80.

Meta 43(4) (1998) Special Issue: The Corpus-based Approach, http://www.erudit.org/erudit/meta/

Olohan, M. (2001) "Spelling out the Optionals in Translation: A Corpus Study", *UCREL Technical Papers*, 13: 423-432.

Olohan, M. (forthcoming a) "Leave it out! Using a Comparable Corpus to Investigate Aspects of Explicitation in Translation". In *Cadernos de Tradução*, Vol. VI.

Olohan, M. (forthcoming b) "How Frequent are the Contractions? A Study of Contracted Forms in the Translational English Corpus".

Olohan, M. and M. Baker (2000) "Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation?", *Across Languages and Cultures*, 1(2): 141-158.

Stubbs, Michael (2001) "Texts, Corpora, and Problems of Interpretation: A Response to Widdowson", *Applied Linguistics* 22(2): 149-172.

Tymoczko, Maria (1998) "Computerized Corpora and the Future of Translation Studies", *Meta* 43(4): 652-659.

Vanderauwera, R. (1985) *Dutch Novels Translated into English: The Transformation of a 'Minority' Literature*. Amsterdam: Rodopi.

Corpora in Translation Practice

Federico Zanettin

Università per Stranieri di Perugia
Palazzo Gallenga, Piazza Fortebraccio, 4 - Perugia
zanettin@unistrapg.it

Abstract

The aim of this paper is to trace links between work in the corpus linguistics community and the world of practicing translators. The relevance to translation work of corpora in general, and bilingual and parallel corpora in particular, is evaluated by comparing corpora and translation memories and by drawing an analogy between different types of corpora and more traditional reference tools, i.e. dictionaries. Corpus resources available to translators are placed along a cline going from "robust", stable corpora (e.g. large reference corpora such as the BNC) to "virtual", ephemeral corpora (e.g. DIY web corpora). Finally, a few suggestions are put forward in order to encourage a wider diffusion of corpora and concordancing software among professional translators.

1. Introduction

The translator's workplace has changed dramatically over the last ten years or so, and today the computer is undoubtedly the single most important tool of the trade for a translator regardless of whether he or she is a literary translator working for a small publisher, a technical translator working for a translation agency or a legal translator. Today, translators compose their texts on the computer screen, often receive their source texts in electronic format and sometimes their translations will only live as digital information as in the case of web site localization.

The specific hardware and software resources individual translators will resort to will vary depending on the task to be done. While in the case of most literary translators the translated text will probably take shape by means of a general purpose word processor, in the case of technical translators the target text will be produced with the help of the most sophisticated "translator workbench", equipped with all sorts of CAT tools, translation memory and terminology systems, and localization software.

The computer has also flanked, if not substituted, other technological supports in providing access to traditional tools and resources. Translation aids such as monolingual and bilingual dictionaries, terminologies and encyclopedias are now available not only on paper but also in electronic format. Colleagues and expert informants can now be consulted via e-mail and newsgroups besides via telephone, fax and face-to-face encounters. The storage capacity and processing power of personal computers have made access to linguistic and content information easier and quicker than ever before, and the Internet has opened up highways of communication and information retrieval. The problem is now not finding a piece of information, but finding the right and reliable piece of information without wasting too much time.

Corpora and concordancing software can be a way of gaining access to information about language, content, and translation practices which was hardly available to translators before the present stage of ICT development. Corpora and corpus analysis software have been around for quite a long time, but their use is only now beginning to extend beyond a restricted segment of language professionals, such as lexicographers, language engineers, as well as linguists in educational and training institutions.

I would like to suggest that corpora and concordancing software could find a larger place in the translator computerised workstation, and that more corpus resources could and should be made more accessible to professional translators. In order to do so, however, corpus builders and software producers should take into account the specific needs of this group of users. Learning to use corpora as translation resources should also be part of the curriculum of future translators and become part of their professional competence.

2. Corpora and translation

According to the EAGLES text typology elaborated by John Sinclair (1996) we can make a general distinction between Monolingual and Multilingual (including Bilingual) corpora. As regards bilingual (and multilingual) corpora a further distinction can be made between Comparable corpora (corpora compiled using similar design criteria but which are not translations) and Parallel, or Translation Corpora, which are texts in one language aligned with their translation in another. This picture can be further complicated by involving variables such as direction and directness[1] of translation, number of languages, number of translations per text, etc., producing bi-directional, reciprocal, control, star and diamond corpus models (cf. Johansson, forthcoming; Teubert, 1996; Zanettin, 2000; Malmkiaer, forthcoming). Still another type of translation related corpus is the Monolingual Comparable Corpus (Baker, 1993), or a corpus composed of two sub-sections, one of texts originally composed in one language and the other of texts translated into that same language (from a number of other languages). This type of corpus, however, while undoubtedly an extremely useful tool for translation theorists, researchers and students, is arguably of less immediate relevance for professional translators dealing with actual translation jobs.

Professional translators working in the technical sector are perhaps more familiar with the parallel concordancing feature of translator memory systems. A translation memory is data bank from which translators automatically retrieve fragments of past translations that match, totally or to a degree, a current segment to be translated, which must match, totally or to a degree - an already translated

[1] (i.e. whether a translation is produced directly from the original text or via an intermediate translation in another language).

segment. But it can also be seen as a parallel corpus which translators manually query for parallel concordances of (already translated) specific terms or patterns. Aligned translation units are conveniently displayed on the screen, offering the translator a range of similar contexts from a corpus of past translations. A translation memory is, however, a very specific type of parallel corpus in that:

a) it is "proprietory": TMs are created individually or collectively around specific translation projects. They are highly specialized and very useful when used for the translation or localization of program updates – indeed that is their origin – but are not much help when starting a new translation project on a different topic or text type.
b) TMs tend to closure, to progressively standardize and restrict the range of linguistic options. This may be an advantage from the point of view of terminological consistency and of processing costs for clients or translation agency managers, but is often detrimental for readability (texts translated using a "Workbench" can become very repetitive) and the translators eyesight (translators using a well-known Workbench often testify to a "yellow-and-blue-eye-syndrome).

Translation workbenches and translation memories have indeed become the most successful technological product to be created for professional translators, but – as it often happens with MT products – their use is best limited to specific text types, such as online help files, manuals and all types of reference work which do not require sequential reading and for which the scope of translation can be limited to the sentence of phrase level (and thus left to a machine). When dealing with other types of texts translators are perhaps better off with a different kind of language resource, i.e. the type of corpora which are more familiar to lexicographers and linguists and which are only now beginning to enter the selection of tools available to professional and trainee translators.

3. Corpora as translation aids

The respective potential uses on the part of professional translators of monolingual target corpora, bilingual comparable corpora, and of parallel corpora can be illustrated drawing an analogy with other respected tools of the trade, i.e. dictionaries: Monolingual target corpora can be compared to monolingual target language dictionaries, and comparable source corpora to monolingual source language dictionaries. While dictionaries favor a synthetic approach to lexical meaning (via a definition), corpora offer an analytic approach (via multiple contexts).[2] Translators can use target monolingual corpora alongside target monolingual dictionaries to check the meaning and usage of translation candidates in the target contexts. Like source language dictionaries, source language corpora can be consulted for source text analysis and understanding. Large reference corpora (BNC, CORIS/CODIS, etc.) can function as general dictionaries, while smaller, specialized and bilingual comparable corpora can be seen as analogous to specialized monolingual dictionaries (either or both in the source and in the target language).

Parallel corpora can instead be compared to bilingual dictionaries, with a few important differences: bilingual dictionaries are repertories of lexical equivalents (general dictionaries) or terms (specialized dictionaries and terminologies) established by dictionaries makers which are offered as translation candidates. Parallel corpora are repertoires of strategies deployed by past translators, as well as repertoires of translation equivalents. In selecting a translation equivalent from a general bilingual dictionary a translator has to assess the appropriateness of the candidate to the new context by starting from a definition and a few usage examples. A parallel corpus will offer a repertoire of translation strategies past translators have resorted to when confronted with similar problems to the ones that have prompted a search in a parallel corpus.

Parallel corpora can provide information that bilingual dictionaries do not usually contain. They can not only offer equivalence at the word level, but also non-equivalence, i.e. cases where there is no easy equivalent for words, terms or phrases across languages. A parallel corpus can provide evidence of how actual translators have dealt with this lack of direct equivalence at word level. For example, in the translations by two different Italian translators of a number of novels by Salman Rushdie (Zanettin, 2001b), the word "edges", which usually collocates with a preposition, as in the phrases "around the edges," or "at the edges," was never translated literally, but rather omitted:

1. …biting the skin around the edges of a nail…
   …*mordicchiandosi la pelle attorno all'unghia…*
2. …around the edges of Gibreel Farishta's head…
   …*intorno alla testa di Gibreel Farishta…*
3. …around the edges of the circus-ring…
   …*intorno alla pista da circo…*
4. …and there was a fluidity, an indistinctiness, at the edges of them…
   …*vicinissime a loro c'erano una fluidità e un'indeterminatezza…*
5. …the horses grew fuzzy at the edges…
   …*i cavalli diventavano sempre più sfocati…*
6. …blurred at the edges, my father…
   …*con la mente annebbiata, mio padre…*
7. …looking somewhat ragged at the edges…
   …*con l'aria di un uomo distrutto…*
8. …Mrs Qureishi, too, was beginning to fray at the edges…
   …*anche Mrs Qureishi si stava consumando…*

In all these cases, the two professional translators have consistently chosen to resort to "zero-equivalence", which being a translation strategy rather than a case of comparative linguistic knowledge would be hardly reported in any bilingual dictionary.

4. Corpus resources for translators

Not all dictionaries are the same, nor are all corpora. Apart from translation memories, corpus resources which are of potential use for professional translators could be classified along a scale which goes from "robust" to "virtual." A "corpus" is a collection of electronic texts assembled according to explicit design criteria which usually aim at representing a larger textual population. "Robust" corpora are ready-made corpora created and

[2] So-called "production dictionaries", which focus on usage information, can be thought of as standing somehow in between the two.

distributed by the research community and the language industry on CD-ROM or accessible through the Internet. Prototypical examples are large reference national corpora, such as the *British National Corpus* (BNC) for British English, and the *Dynamic Corpus or Written Italian* (CORIS/CODIS) for Italian. This type of resource, which requires a large building effort, is only now becoming available to the wider public outside the (corpus) linguistics community, and will probably require some "customisation" effort in order to become more widespread among language services providers.

Parallel corpora are usually smaller and even less available to the general public than monolingual corpora. Their construction requires more work than that of monolingual corpora. Among other factors, text pairs (rather than single texts) have to be located and before they can be used they need to be aligned, at least at the sentence level (cf. Véronis, 2000).

There are of course varying degrees of robustness, according to the effort and care which has been put in achieving a balanced and representative selection of texts, in providing explicit linguistic and extralinguistic information (corpus annotation) and the means (the software) to query the corpus for that information (McEnery & Wilson, 1996). Corpus design criteria also vary according to the purpose for which a corpus is built, e.g. a comparable monolingual corpus for descriptive translation research. In this sense, the less "robust" (i.e. the more "virtual") corpora are the most truly professional type, with reference to translators, since they are "rough-and-ready" products created for a specific translation project. A distinction is usually made by corpus linguists between "corpora" and "archives" of electronic texts. An "archive" is simply a repository of electronic texts: In this sense the WWW is an immense (multimedia) text archive. Virtual or "disposable" corpora are created by a translator using the WWW as a source "archive". The WWW and HTML documents need not to be the only source for small, specialized DIY corpora, and textual archives of various types and targeted to various users (newspapers, collections of laws, encyclopedias, etc.) are available on cd-rom. The WWW is however certainly the most familiar and user friendly environment for translators: it is always available; it is the most comprehensive source of electronic texts, and corpus creation, management and analysis can be a relatively straightforward operation (Austermühl, 2001; Zanettin, forthcoming). Building a corpus of web pages basically involves an information retrieval operation, conducted by browsing the Internet to locate relevant and reliable documents which can then be saved locally and made into a corpus to then be analysed with the help of concordancing software. The additional time required by creating and consulting a corpus is compensated for by saving in other translation-related tasks, such as dictionary consultation (both on paper and electronic), paper documentation (often in the form of "parallel texts", e.g. Williams, 1996), help from experts, and by the fact that the corpus contains information not available elsewhere. Moreover, the effort is rewarded by improving quality in terms of terminological and phraseological accuracy (Friedbichler & Friedbichler, 2000).

A number of studies have reported on experiments in translation and language teaching classes with DIY corpora, either made of "disposable" web pages (e.g. Varantola, 2000, forthcoming; Maia, 1997, 2000, forthcoming; Zanettin, forthcoming; Pearson, 2000) or of texts taken from other electronic sources such as newspapers (Zanettin, 2001a) or magazines (Bowker, 1998) on CD-ROM. Corpora created from sources other than web pages can require more time and effort to be built, and can be more or less "disposable" depending on the size of the translation project and on the resources available to create and manage them.

Reports on the use of corpora by professional translators are fewer: Friedichler & Friedbichler, drawing on their experience as translators of medical texts and trainers of technical translators, suggest that domain-specific target language corpora may usefully complement dictionaries and the Web as resources in the translation process, filling the gap between the two. Jääskläinen and Mauranen (2000) report on an experimental study involving a team of researchers from the University of Savonlinna and a team of professional translators translating for the timberwood industry. The researchers created a corpus from a variety of sources (web sites, PDF documents, etc.) following suggestions from the translators, and then trained them in using concordancing software (*WS Tools*, Scott, 1996) to analyse the corpus. In exchange, the translation team agreed to answer a questionnaire. One of the results of the study was learning that translators often complained that the user-friendliness of the concordancing software was very low. This complaint was seconded by translator trainees in other studies with "disposable" corpora where students, usually working in groups, collected a corpus of HTML documents and used them to help them translate a specific text.

These studies have underlined, nonetheless, the value of corpus building as a way of getting acquainted with the content and terminology of the translation. They have stressed the importance of type and topic of the text to be translated as well as of the target language (some text types, topics, and target languages are better helped with corpora than others) and also of adopting sound criteria in choosing suitable texts for inclusion in the corpus. Most of the corpora in these experiments were target monolingual corpora, though some use of bilingual comparable and even parallel corpora was reported.

The main benefits and shortcoming of DIY corpora may be summed up as follows:

Benefits:

- They are easy to make.
- They are a great resource for content information.
- They are a great resource for terminology and phraseology in restricted domains and topics.

Shortcomings:

- Not all topics, not all text types, not all languages are equally suitable or available.
- The relevance and reliability of documents to be included in the corpus needs to be carefully assessed.
- Existing concordancing software is not well equipped to handle HTML or XML files, i.e. web pages. There are no or few parallel corpora, since while some parallel texts (i.e. source texts + translations) can be found on the Internet, hardly all of them could be included in a parallel corpus designed to provide instances of professional standards (Maia, forthcoming).

DIY web corpora stand midway the WWW itself, which can be used as if it were a corpus and robust, "proper"

corpora. As for the Web, a "quasi-concordance" view of documents indexed and retrieved is provided by such as search engines Google (http://www.google.com) or Copernic (http://www.copernic.com). Corpus linguistics-oriented software currently being constructed for browsing the WWW as a corpus, such as *KwicFinder* (Fletcher, 2001) and *WebConc* (Kilgarriff, 2001), will certainly prove a useful tool for translators among other language professionals. However, while this "web as corpus" approach has certainly advantages in terms of time over DIY web corpora (the "corpus" is always already there), it necessarily looses in precision and reliability.

The advantages of "robust" corpora over "virtual" corpora can instead be summed up as follows:

- They are usually more reliable.
- They are usually larger.
- They may be enriched with linguistic and contextual information.
- If parallel, they are already aligned.
- They come with user-friendly, customised software (though, again, not necessarily targeted to the needs of professional translators).

5. Conclusions

Translators can tolerate the learning curve necessary to adopt corpora and concordancing software among their everyday working tools only if they derive benefits. These benefits are the fact that corpora provide information not available elsewhere at an affordable cost.

As a way of concluding, I would like to point out possible improvements for existing corpora and concordancing software:

a) "Robust " reference corpora need to become more accessible: for instance, a BNC license is still relatively expensive and the interrogation software might do with some customization; the CORIS/CODIS corpora and others have limited access.

b) In order for "virtual" corpora to become more widespread among translators, concordancing software for work with small monolingual corpora has to become capable of dealing with HTML and, increasingly, XML texts. For example, it may be useful to interface the concordancing software with the Internet browser to provide facilities for file downloading and management, and for allowing the user to switch between concordance lines and full text view, in order to take advantage of multimedia features of electronic texts.

c) Bilingual and parallel corpora are scarcely available and usually of limited size. Bilingual concordancers require bilingual corpora, and given what it takes to locate and align text pairs, it is not very likely that individual translators will resort to consulting parallel concordances unless parallel (aligned) corpora are already available. The creation of more corpora of this kind is a matter of computational resources (especially parallel concordancers and efficient aligning utilities) as well as of more awareness of the usefulness of this resource among translators and language resources providers.

6. References

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

Baker, M. (1993). "Corpus linguistics and translation studies. Implications and applications". In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and technology*. Philadelphia/Amsterdam: John Benjamins, 233-252.

BNC web site, http://info.ox.ac.uk/bnc

Bowker, L. (1998). "Using specialized monolingual native-language corpora as a translation resource: a pilot study", in *META* 43:4, 631-651.

CORIS/CODIS web site, http://www.cilta.unibo.it

Fletcher, W. (2001). "Concordancing the web with KWiCFinder", presentation given at the *Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA, 23-25 March 2001. Available at http://miniappolis.com/KWiCFinder/Corpus2001.htm.

Friedbichler, I. & Friedbichler, M. (2000). in S. Bernardini & F. Zanettin (eds.) *I corpora nella didattica della traduzione*. *Corpus Use and Learning to Translate*, Bologna: CLUEB, 107-116.

Jääskeläinen, R. & Maurannen, A. (2000) *Work Package 5: Development of a Corpus on the Timber Industry - Final Report, Project SPIRIT MLIS-programme: MLIS-3008 SPIRIT 24637*, University of Joensuu, Savonlinna School of Translation Studies.

Johansson, S. (forthcoming). "Reflections on corpora and their uses in cross-linguistic research", in F. Zanettin, S. Bernardini, & D. Stewart (eds.) *Corpora in translator education*.

Kilgarriff, A. (2001). "Web as corpus". In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 conference, UCREL Technical Papers: 13.* Lancaster University, 342-344.

Maia, B. (1997). "Do-it-yourself corpora ... with a little bit of help from your friends!" in B. Lewandowska-Tomaszczyk & P. J. Melia (eds.) *PALC '97 Practical Applications in Language Corpora*. Lodz: Lodz University Press, 403-410.

Maia, B. (2000) "Making corpora: A learning process", in S. Bernardini & F. Zanettin (eds.) *I corpora nella didattica della traduzione*. *Corpus Use and Learning to Translate*, Bologna: CLUEB, 47-60.

Maia, B. (forthcoming) "Training translators in terminology and information retrieval using comparable and parallel corpora", in F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education*.

Malmkiaer, K. (forthcoming). "On a pseudo-subversive use of corpora in translator training", in F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education*.

McEnery, T. & Wilson, A. (1996) *Corpus linguistics*. Edimburgh: Edimburgh University Press.

Pearson, J. (2000). "Surfing the Internet: teaching students to choose their texts wisely". In Lou Burnard and Tony McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Maim et al: Peter Lang, 235-239.

Scott, M. (1996). *Wordsmith Tools*. Oxford: Oxford University Press.

Sinclair, J. McH. (1996) *EAGLES Preliminary recommendations on Corpus Typology, EAG--TCWG--CTYP/P*. Online: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Teuberg, W. (1996) “Comparable or parallel corpora?” *International journal of lexicography*, 9:3, 238-264.

Varantola, K. (2000). “Translators, dictionaries and text corpora” in in S. Bernardini & F. Zanettin (eds.) *I corpora nella didattica della traduzione*. *Corpus Use and Learning to Translate*, Bologna: CLUEB, 117-136.

Varantola, K. (forthcoming). “Translators and disposable corpora” in F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education*.

Véronis, J. (2000) *Parallel text processing. Alignment and use of parallel corpora*. Dordrecht: Kluwer.

Williams, I. A. (1996) “A translator’s reference needs: Dictionaries or parallel texts”. *Target* 8, 277:299.

Zanettin, F. (2000). “Parallel corpora in translation studies: issues in corpus design and analysis”, in Olohan, M. (ed.) *Intercultural Faultlines. Research Models in Translation Studies I. Textual and cognitive aspects*, Manchester: St Jerome. 93-118.

Zanettin, F. (2001a). “Swimming in words: Corpora, translation, and language learning”, in G. Aston (ed.) *Learning with corpora*, Bologna/Houston,TX: CLUEB/Athelstan, 177-197.

Zanettin, F. (2001b). *IperGrimus*. In *inTRAlinea* (online) http://www.intralinea.it

Zanettin, F. (forthcoming). “DIY corpora. The WWW and the translator”, *Proceedings of the “Training the language services provider for the new millennium” International Conference, Porto, Portugal, 25-26 May 2001*.

BancTrad: a web interface for integrated access to parallel annotated corpora

Toni Badia, Gemma Boleda, Carme Colominas, Agnès González, Mireia Garmendia, Martí Quixal

Universitat Pompeu Fabra
Rambla 30-32 ,
E-08002 Barcelona
{toni.badia,carme.colominas,marti.quixal}@trad.upf.es, gemma.boleda@iula.upf.es

Abstract

The goal of BancTrad is to offer the possibility to access and search through (parallel) annotated corpora via the Internet. This paper presents the design of the whole process: from text compilation and processing to actually performing queries via the web, while it describes as well its technical architecture.

The languages we work with are Catalan, Spanish, English, German and French. Queries are possible from any of these languages to Spanish and Catalan and vice versa (but not between the language pairs formed by French, German and English). The texts go first through a pre-processing and mark-up stage, then through linguistic analysis and are finally formatted, indexed and made ready to be consulted. The web interface has been created through the integration some *ad hoc* applications and some ready-to-use ones. It provides three different levels of query expertise: basic, intermediate and expert.

The paper is structured as follows: section 1 gives an overview of the project; section 2 describes the text compilation process; section 3 explains the corpora building and parsing stages; section 4 details the search machine architecture; finally, section 5 describes foreseen applications of BancTrad.

1. Overview

The original idea of BancTrad[1] was to obtain a tool with pedagogic applications (see work done e.g. by Gaspari, Hansen, S.) especially thinking of translation and interpreting courses held at the Translation and Interpretation Faculty (FTI) of the University Pompeu Fabra (UPF). It was meant to be a translation databank that could serve both teachers and students to search for prototypical translations or texts containing special features that would make them interesting from the translator's point of view. Afterwards, the target user of BancTrad was broadened to e.g. professional translators and linguists (see section 5), through the creation of different search modes and the expansion of the expressiveness of the queries, in order to adapt to the user needs or knowledge.

As an annotated translation databank, BancTrad offers the possibility to work with Catalan, Spanish, English, German and French. Queries are possible from any of these languages to Spanish and Catalan and vice versa (but no queries are possible between the language pairs formed by French, German and English), as well as between Catalan and Spanish in both directions. The web page of the project can be accessed from http://glotis.upf.es/bt/index.html

2. Text collecting, extra-linguistic tagging and alignment

The corpora in BancTrad aim at being representative for translated texts. In other words, they don't have a normative character but a descriptive one. Therefore we have chosen to collect documents from very different sources, representing a variety of text types, subjects and registers.

The main sources we have focussed on are faculty professors, work done in translation courses, publishing houses and the Internet. Many faculty professors work also as freelance translators, which constitutes a good source of high quality translations. Besides, the fact that we include (supervised) work done in translation courses can have many advantages regarding academic self-evaluation. Specially, because they give evidence of the text types, subjects, etc., which have been worked on with pedagogical purposes. As for translations from the Internet, some supervision is done on them before they are selected to be introduced in BancTrad (for the sake of quality).

Selected texts are semi-automatically processed to be marked up with SGML tags and aligned with their respective original texts. Both the originals and the translations are marked up with some extra-linguistic information by means of a special MS Word form coded in Visual Basic (see Fig. 1).

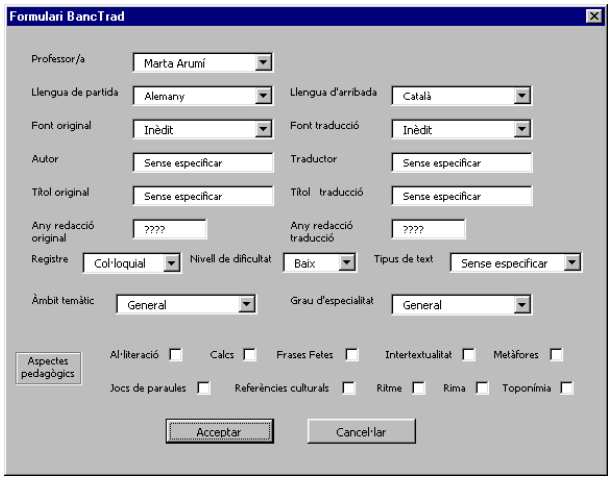


Figure 1: MS Word form used for the mark-up of extralinguistic features of the texts

[1] This project is running under the auspices of the "Programa d'Innovació Docent" (Educational Innovation Program) sponsored by our university (Universitat Pompeu Fabra) and has also been partially financed by the Spanish Government and by the 2001FI 00582 grant from the autonomous Government of Catalonia.

This mark-up takes the following parameters into account:

- name of the person who introduced the aligned texts (i. a., in order to track translation quality)
- source and target languages
- original and translation references
- publication date (for both the original and the translation)
- register (colloquial, standard, learned, etc.)
- type of text (normative, descriptive, literary, etc.)
- subject matter (economy, science, politics, etc.)
- degree of specialisation (low, middle, high).

Besides these parameters, and bearing in mind that BancTrad was originally conceived as a tool with pedagogic applications, we include information on certain aspects such as idioms, metaphors, puns, degree of difficulty, etc. All of these parameters, as well as the information coded within them, were consensuated with the teachers and researchers of the FTI. It is relevant to note that this mark-up allows us not to make a rigid classification of the texts in the corpus (see section 3).

By clicking on the *Acceptar* ("Accept") button, the options selected in the form are marked in the text in SGML format and a script tags the paragraph structure of the document. Otherwise, this very valuable piece of information on the text structure would be lost in the alignment step.

Texts are aligned at a sentence level with the align tool of the DéjàVu Database Maintenance, software by Atril (http://www.atril.com). DéjàVu aligns texts and allows editing in quite a user-friendly way.

The tasks described so far, although only semi-automatic, require neither special skills in computing nor much time (the time to go through them for a 400 word-long text -both source and target texts- is 5 to 10 minutes). We could have chosen to tackle the alignment task fully automatically instead, but the error rate of automatic aligners (notably errors in sentence identification) would have increased too much the error rate in the subsequent linguistic analysis. However, it should be kept in mind that, according to our architecture, the use of a particular tool for the mark-up and alignment independent of the rest of the process, so that other tools could be used in the future.

Finally, the texts are transferred to our Linux server to proceed with the text processing, which from this moment on will be completely automatic.

3. Linguistic Processing and Corpus Building

Once the texts are in the server, they undergo two further steps: linguistic tagging and corpus formatting. Both steps are completely automatic.

3.1. Linguistic Processing

Each language follows a different tagging process. On the one hand, Catalan texts are parsed with CATCG (Badia *et al.* 2000), a Catalan shallow morphosyntactic parser based on a constraint grammar developed by the Computational Linguistics group at UPF. Spanish texts will be handled with a Spanish version of it in a year's time. On the other hand, the linguistic analysis for English, German and French texts is made with TreeTager, a part-of-speech tagger developed at the IMS (see Schmid 1995, 1997). Both CATCG and TreeTager are shallow parsers.

La	**noia**	**de**	**el**	**port**	**de**	**Barcelona**	**dorm**
the	***girl***	***of***	***the***	***harbour***	***of***	***Barcelona***	***sleeps***

```
<s id="1">
La          el          Det         AFS         DN>
noia        noi         Nom         N5-FS       Subj
<contrac forma="del">
de          de          Prep        P           <NA
el          el          Det         AMS         DN>
</contrac>
port        port        Nom         N5-MS       <P
de          de          Prep        P           <NA
<enty>
Barcelona               Barcelona               Nom         N4G6S  <P
</enty>
dorm                    dormir      Verb        VRR2S-      VPrin
.           .           .           .           PT
</s>
```

Figure 2: Input and output of CATCG

It is important to note that, despite the use of different tagging tools for exploiting the linguistic information of our texts, all languages receive a minimum of uniform kind of information: lemma and POS tag (syntactic function is only there for Catalan). Thus, all the languages can be processed and made queries upon in the same fashion, independently of the tagging tool used. This favours modularity, for the linguistic processing of a certain language can be modified without changing any of neither the other linguistic processes nor the interface. We now proceed to roughly characterize CATCG and TreeTagger.

3.1.1. CATCG

CATCG is a linguistic-based parser that assigns each word a lemma, a POS tag and a syntactic function. It uses three major devices:

a) a Perl module for the preprocessing
b) a morphological tag mapping tool that uses a word-form dictionary created with a morphological generator developed at UPF (Badia *et al.* 1997)
c) three grammars using the Constraint Grammar formalism developed at the University of Helsinki (Karlsson et al. 1995, Tapanainen 1996), which perform the morphosyntactic disambiguation task and the partial syntactic analysis.

Fig. 2 gives an example of the input and output of our system. The SGML tags are the result of the preprocessing, and in the example they mark a contracted form, an entity and the sentence boundaries. The columns list the linguistic information: word form, lemma, part of speech tag, complete morphological information in an compressed tag and syntactic function (in order of appearance). The last piece of information is shallow and partial in the sense that it doesn't fully indicate dependency: note that the

preposition *de* (“from”) in the PP *de Barcelona* gets a tag indicating that it modifies a noun to its left (<*NA*, left adjoining Nominal Adjunct); however, no clue is given about whether it modifies *Barcelona* or *port*.

3.1.2. TreeTager

TreeTager is a probabilistic tagger that uses decision trees. It provides each word with a lemma and a POS tag (at the moment, no syntactic information is given).

3.2. Corpus formatting

After being annotated, the text files are eventually formatted and processed with the Corpus WorkBench (CWB) tools, a set of linguistic information exploitation tools developed at the IMS in Stuttgart (Christ 1994; Christ *et al*. 1999[2]). Thus we build the actual corpora making them ready to be consulted with CQP, the Corpus Query Processor, a tool from the CWB. This tool allows very flexible and expressive queries for any of the pieces of information encoded (be it the word form, lemma, POS tag or syntactic function). In fact, as a far as one gives corpora the adequate structure, one can have as a many attributes as one pleases.

One of the most significant (to us) features of the CWB is the fact that it can process aligned corpora. Not only is it possible to view the aligned sentences, but it is also possible to place restrictions both on the source and on the target language in a query (see section 5). It has also been crucial to us the special module that lets CQP interacting with the web (see next section).

4. The search machine and the web Interface

Technically speaking, the novelty of BancTrad is the integration of several tools that make available parallel annotated corpora via the Internet. This entails that the system has to be able to (1) interpret the query made by the user, (2) search for the query, (3) present the results. For this purpose, two devices were needed: a graphical user interface (GUI) with a fill-in form and an external program interface (to allow browser/server communication)

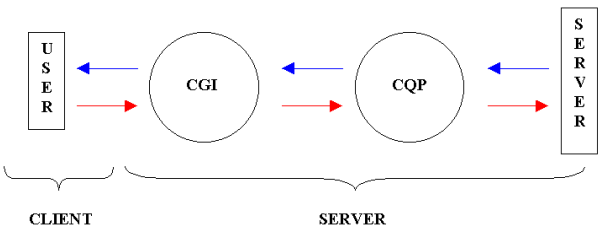


Figure 3: Query routing through the client/server architecture (query from left to right, results the other way round)

a) The GUI for query input

The GUI is intended to be adaptable to the user expertise, to have open access and to be platform independent. For our GUI to accomplish the two last features, an HTML-based interface seemed to be the best option. To qualify for the first one, the interface had to offer at least three search possibilities: common, intermediate and expert mode (see next section for details).

b) The external program interface

This is the module of the architecture that actually makes the query processing. It interprets the user's query, it searches for it in the corpora and gives the result back. The program that does the work is commonly called a cgi (Common Gateway Interface, term whose original sense has been extended to mean “external program interface”). Our cgi is composed of the following packages:

i) Common Gateway Interface (CGI)

The CGI (properly so named) is a standard device to interface with information servers (such as HTTP servers). It passes a web user's request on to an application program and gives the resulting data back to the user. Herewith the server interprets the user’s query.

ii) HTML::Entities

This formatting package ensures that special characters (tildes, cedillas, etc.) are properly transferred during the client/server session.

iii) WebCqp::Query, a web adapted version of the CQP

This package was designed by the creators of the CWB (see above) to let it interact with the web. It can perform the same kind of queries that CQP performs in its PC-Linux version. It thus allows a powerful query setting through regular expressions, access to linguistic tags (through the defined number of features in the corpora) and aligned corpus querying.

5. Exploiting BancTrad

This section outlines different ways in which to exploit BancTrad, from two different but related perspectives regarding its potential users. It describes the search possibilities that BancTrad offers (section 5.1), which relates to the user's level of expertise. Besides, it sketches some possible applications for which BancTrad is indicated (section 5.2), which relates to the user's professional or academic profile.

5.1. Search possibilities

5.1.1. Three levels of expertise

The web interface of BancTrad had to enable the users to access the corpora without having to be experts neither on linguistics nor on regular expressions. Moreover it had to offer the possibility of exploiting the full-fledged regular expressions that CQP allows, as well as the chance of profiting from the quite detailed linguistic annotation of the corpora. Therefore, BancTrad offers three different search modes (corresponding to levels of query expertise):

a) **basic mode:** allows searching for sequences of specific word forms (with possibly their equivalence in a target language).

b) **intermediate mode:** allows searching for sequences of five

[2] See also the web page of the CWB: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

quadruples (form, lemma, morphosyntactic tag, and syntactic function), including the iteration of identical elements

Fig. 4 is a screenshot of a search in this mode: it searches for causative constructions from Catalan into English, that is, for the causative verb *fer* followed by any verb (see next section for the results).

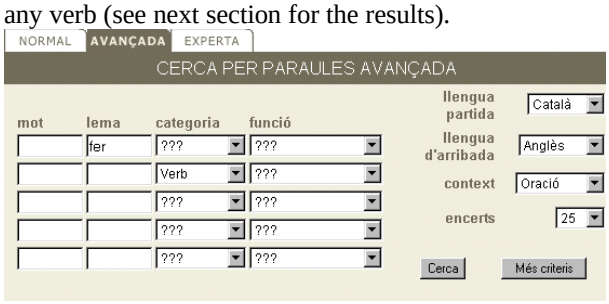


Figure 4: Screen shot of the intermediate query mode of BancTrad

expert mode: to set queries expressed in the full regular language provided by CQP.

5.1.2. Restrictions on extralinguistic features

Additionally to the word units searched for, the user can place restrictions on extra-linguistic features of the texts containing them. This is possible through the initial mark-up stage (see section 2) while formatting the corpora. Thus, through an extended web-form, the user can restrict the occurrences of e.g. the word "bank" to appear in economic texts.

This kind of mark-up gives rise to a different search possibility, planned for the original purpose of BancTrad (which was being useful for teaching purposes at the FTI): the full text query, which allows the user to search for complete texts and their translation, restricting them by the extra-linguistic features mentioned above. Fig. 5 shows a text query in which the user wants to retrieve essays (*Assaig*) on Arts originally written in German (*Alemany*) and translated into Spanish (*Castellà*).

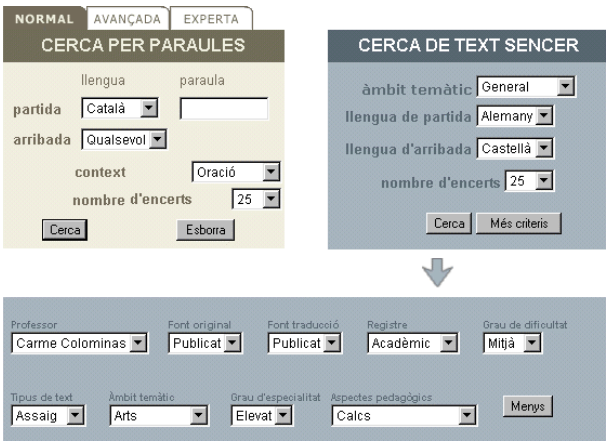
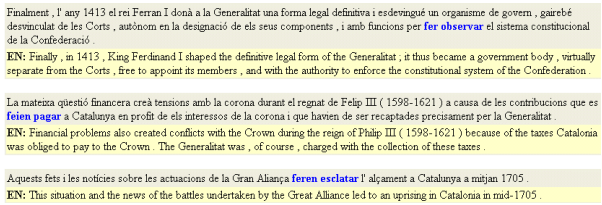


Figure 5: Screenshot of the text query mode of BancTrad

5.1.3. Showing the results

As for the presentation of the results, they are shown by default as aligned full sentences, although it is foreseen that the user can switch to other presentation forms: a full paragraph or just some words to the left and/or right sides of the query target. Of course all the capabilities listed so far are indebted to the Corpus Query Processor that we use as a searching engine.

Fig. 6 shows some of the results for the query on causative constructions made on section 5.1.1:


Finalment , l' any 1413 el rei Ferran I donà a la Generalitat una forma legal definitiva i esdevingué un organisme de govern , gairebé desvinculat de les Corts , autònom en la designació de els seus components , i amb funcions per **fer observar** el sistema constitucional de la Confederació .
EN: Finally , in 1413 , King Ferdinand I shaped the definitive legal form of the Generalitat ; it thus became a government body , virtually separate from the Corts , free to appoint its members , and with the authority to enforce the constitutional system of the Confederation .

La mateixa qüestió financera creà tensions amb la corona durant el regnat de Felip III (1598-1621) a causa de les contribucions que es **feien pagar** a Catalunya en profit de els interessos de la corona i que havien de ser recaptades precisament per la Generalitat .
EN: Financial problems also created conflicts with the Crown during the reign of Philip III (1598-1621) because of the taxes Catalonia was obliged to pay to the Crown . The Generalitat was , of course , charged with the collection of these taxes .

Aquests fets i les notícies sobre les actuacions de la Gran Aliança **feren esclatar** l' alçament a Catalunya a mitjan 1705 .
EN: This situation and the news of the battles undertaken by the Great Alliance led to an uprising in Catalonia in mid-1705 .

Figure 4: Screen shot of the intermediate query mode of BancTrad

5.2. Applications of BancTrad

There are several uses one can think of for BancTrad. Of course, the most direct and obvious one is the one for which the parallel databank was thought: educational use. But there are at least two other kinds of applications that were held in mind while developing the project: research and professional applications. The three of them are outlined, with some examples, in this section.

5.2.1. Teaching

For educational purposes, all of the search modes (be it string or text queries) outlined in the previous subsection are relevant. However, as the full text query has already been exemplified, we will concentrate on the first one. The string equivalence query, which we foresee to be the most significant application for the corpora included in BancTrad, is the search of bilingual equivalences among language pairs. This includes the search of word equivalence, restricted by its form in one of the languages, by its lemma, or by its form or lemma and its morphosyntactic tag. Thus typical searches (which demand different levels of expertise in the search mode) could be:

a) translation of the English form 'stores' into Catalan. Result: *botigues* (noun), *guarda* (verb).
b) translation of the English lemma 'store' into Catalan. Result: *botiga*, *botigues* (noun), and the whole paradigm of the verb *guardar*.
c) translation of the lemma 'store' with part-of-speech 'verb' into Catalan. Result: the whole paradigm of the verb *guardar*.

Note that as in standard corpus search engines, word forms and lemmata can be searched for in specific contexts, as well as particular combinations of forms, lemmata or part-of-speech tags. For example:

d) translation of the gerundive form of the verb 'indicate' right after a colon.

In addition, a specific search condition on the aligned text can be set. For example:

e) translation of the gerundive form of the verb 'indicate' just after a colon provided that in the translated sentence into Catalan no gerundive is present; alternatively, provided that the verb 'indicar' is used.

5.2.2. Professional and research applications

In fact, these kind of applications just follow from the examples described above and the characteristics of the corpora in BancTrad. On the one hand, as far as the corpora are real translated texts (see section 2), and provided the search possibilities sketched above, BancTrad appears to be a useful tool for professional translators. They could look for evidence of previous translation decisions and even have the information of the person in charge for that translation.

On the other hand, linguists and translation theorists (see work done by Baker, M. and Teubert, W.) could also take advantage of this search engine. In fact, this is something we have already been doing with the grammar-developing task we have been carrying on for the last three years. We can retrieve data such as most frequent readings, syntactic structures, etc. This helps us concentrate on problems arising when dealing with written text and develop more data-driven linguistic-based grammars. It is also interesting to note that searches can be made on a sole language, that is, they must not be bilingual.

Other possible applications for BancTrad include creating further Language Resources, such as multilingual dictionaries, chunkers, stochastic-based machine translation systems, etc.

5.2.3. An added value

Finally, it is important to note that an added value to BancTrad's web interface is the fact that it can incorporate other corpora (also monolingual ones) with little amount of work. This would enable our users to query on several corpora, not only the ones prepared at the FTI, in a user-friendly and familiar web interface. For instance, we already have the British National Corpus as part of our searchable corpora and we are planning to integrate the Frankfurter Rundschau corpus soon as well.

6. Conclusions and future work

We have presented a parallel-annotated corpora web interface that integrates several linguistic tools, both for exploiting linguistic information and for exploiting the linguistically enriched texts. It was originally thought to be a translation teaching help tool, but its possibilities have been so extended that it can be of use to both common public and professional users.

Technically speaking, BancTrad integrates tools from different techniques and fields. On the one hand, we use parsing tools developed at our centre, which have been developed with linguistic techniques. Moreover, we are planning to use parsers developed with stochastic techniques (TreeTagger, see above). On the other hand, we have been taking advantage of several ready-to-use packages for client/server interaction. Thus, we feel our project provides evidence of the necessity of academic co-operation to produce tools for the exploitation of linguistic information.

7. Acknowledgments

Thanks all teachers of the FTI for their collaboration. Feedback from the anonymous reviewers was also very useful.

8. References

Badia, T., À. Egea & T. Tuells (1997) CATMORF: Multi-two level steps for Catalan morphology. In *Demo Proceedings of the Conference on Applied Natural Language Processing*. Washington

Badia, T., Boleda, G., Bofias, E. & Quixal, M. (2001) A modular architecture for the processing of free text. *Proceedings of the Workshop on 'Modular Programming applied to Natural Language Processing'* at *EUROLAN 2001*. Iasi, Romania.

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

Christ, Oliver, Schulze, Bruno M. and König, Esther (1999) *Corpus Query Processor (CQP). User's Manual*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart

Karlsson, F. et al. (1995) *Constraint Grammar: a Language-Independent Formalism for Parsing Unrestricted Text*, Mouton De Gruyter: Berlin/New York

Schmid, Helmut (1995) Improvements in Part-of-Speech Tagging with an Application to German, in *Proceedings of the ACL SIGDAT-Workshop*, pp. 47-50

Schmid, Helmut (1997) Probabilistic Part-of-Speech Tagging Using Decision Trees, in Daniel Jones and Harold Somers, editors, *New Methods in Language Processing Studies in Computational Linguistics*, UCL Press, London, pp. 154-164

Tapanainen, P. (1996) *The Constraint Grammar Parser CG-2*, Department of General Linguistics, University of Helsinki, Helsinki, Publications, number 27.

ParaConc: Concordance Software for Multilingual Parallel Corpora

Michael Barlow

Rice University
Dept. of Linguistics
Houston, TX 77005
barlow@rice.edu

Abstract

Parallel concordance software provides a general purpose tool that permits a wide range of investigations of translated texts, from the analysis of bilingual terminology and phraseology to the study of alternative translations of a single text. This paper outlines the main features of a Windows concordancer, ParaConc, focussing on alignment of parallel (translated) texts, general search procedures, identification of translation equivalents, and the furnishing of basic frequency information. ParaConc accepts up to four parallel texts, which might be four different languages or an original text plus three different translations. A semi-automatic alignment utility is included in the program to prepare texts that are not already pre-aligned. Simple text searches for words or phrases can be performed and the resulting concordance lines can be sorted according to the alphabetical order of the words surrounding the searchword. More complex searches are also possible, including context searches, searches based on regular expressions, and word/part-of-speech searches (assuming that the corpus is tagged for POS). Corpus frequency and collocate frequency information can be obtained. The program includes features for highlighting potential translations, including an automatic component "Hot words," which uses frequency information to provide information about possible translations of the searchword.

Keywords: alignment, parallel texts, concordance software

ParaConc is a tool designed for linguists and other researchers who wish to work with translated texts in order to carry out contrastive language studies or to investigate the translation process itself.

1. Alignment

The successful searching and analysis of parallel texts depends on the presence of aligned text segments in each language corpus (and, of course, on the availability of parallel corpora). The alignment, an indication of equivalent text segments in the two languages, typically uses the sentence unit as the basic alignment segment, although naturally such an alignment is not one in which each sentence of Language A is always aligned with a sentence of Language B throughout the texts, since occasionally a sentence in Language A may, for example, be equivalent to two sentences in Language B, or perhaps absent from Language B altogether. (More difficult problems arise in cases where the translation of one sentence in Language A is distributed over several sentences in Language B.) The size of the aligned segments is not set by the software, however. It would be possible to work with paragraphs as the basic alignment unit, but then the results of a search will be more cumbersome because the translation of a word or phrase will be embedded within a large amount of text, which is especially difficult in cases in which the language is not well-known.

The alignment utility in *ParaConc* is semi-automatic. When files are loaded, the user enters information about the format of the files either through reference to SGML tags or via specifications of patterns. The user specifies the form of headings and the form of paragraphs. *ParaConc* uses the information to align the documents at this level and the user can make adjustments by merging/splitting units, as appropriate. Sentence level alignment, if it is not indicated by SGML tags, is performed using the Gale-Church algorithm (Gale and Church, 1993). The alignment information is saved to a file as part of the workspace, as described in Section 6.

No use is made of bilingual dictionaries or of any kind of language-particular information, but the user can enter pairs of anchors, such as cognates, numerals and dates, which the program will track. These anchors are not used in the alignment process itself, but aligned units which do not contain the appropriate corresponding anchors are highlighted for manual checking by the user.

If the parallel texts are pre-aligned, then it is simply necessary to indicate the manner in which the alignment is marked.

2. Loading the Parallel Corpus

When the LOAD CORPUS FILE(S) command is given, a dialogue box appears, enabling particular parallel files to be loaded, as shown in Figure 1.

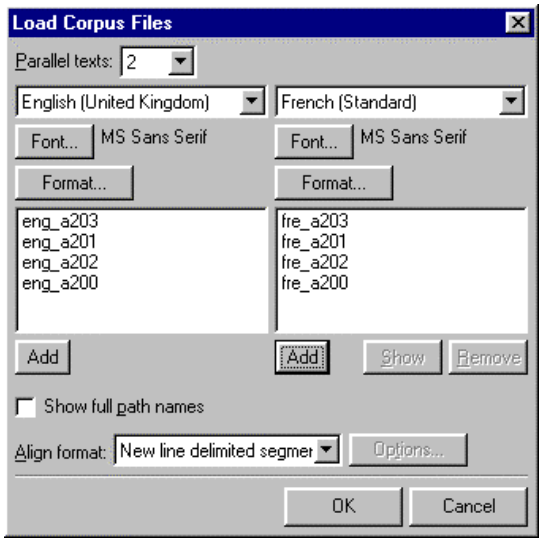


Figure 1. Loading Corpus Files

The heading PARALLEL TEXTS at the top of the dialogue box is followed by a number in the range 2-4 (i.e, two to four different languages). The FORMAT buttons allow the user to describe the form of headings, paragraphs, and sentences, as discussed above. Filenames can be reordered by dragging them to the appropriate position.

3. Searching and Analysing Parallel Texts

The program processes the files as they are loaded, counting words, recording the position of alignment indicators, and processing other format information.

Once a corpus is loaded, some new menu items related to the analysis and display of the text appear on the menu bar. These are FILE, SEARCH, FREQUENCY, and INFO. In addition we can obtain information in the lower left corner of the window relating to the number of the files loaded and in the lower right corner a word count for the two corpora is provided.

Selecting SEARCH from the SEARCH menu initiates the search process and the program starts to work though the loaded files looking for the search string. The search can be based on any of the languages represented: either English or French in this example. (The basic search is fairly simple: a word or a phrase can be entered, including simple wildcard characters if necessary. The symbols acting as wildcards are user-defined, but the default symbols are ? for one character; % for zero or one characters; and * for zero or more characters. The symbol @ covers a specified range of words. Information on the span covered by @ and other information such as a list of characters that act as word delimiters is available in SEARCH OPTIONS.)

Below the results of a search for *head* are illustrated. The instances of *head* are displayed in a KWIC format in the upper window. Clicking on one particular example of *head* in English highlights both the English and French lines. (Double-clicking on a particular line evokes a context window, which provides an enlarged context for the particular instance of the searchword.)

The lower part of the window contains the French sentences (or text segments) that are aligned with the hits displayed in the top window. This display of equivalent units in the two languages is, of course, a consequence of the alignment process. Thus if the first instance of *head* occurred in segment 342 of the English text, then the program simply throws segment 342 of the French text into the lower window, and this process is repeated for all instances of *head*.

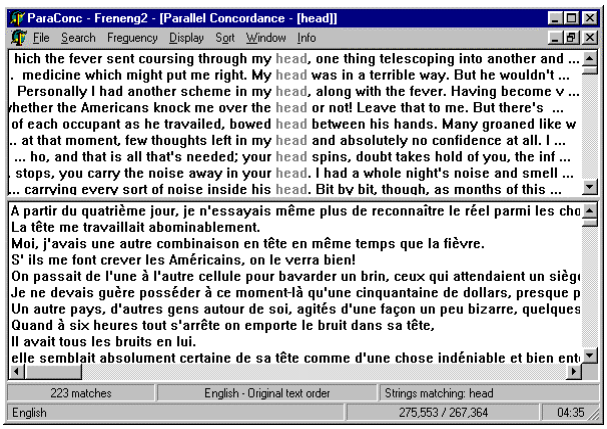


Figure 2: The Results of a Simple Search

Let's follow this example further. Once the search is ended, we can bring to bear the usual advantages of concordance software to reveal patterns in the results data. One may be interested, for example, in different uses (and translations) involving *head: big head, company head, shower head*, etc. One way to find out which English words are associated with *head* is to sort the concordance lines so that they are in alphabetical order of the word preceding the search term. The advantage of performing this 'left sort' is that the modifiers (adjectives) of *head* that are the same will occur together. One easy way to achieve this ordering is to select 1ST LEFT, 1ST RIGHT, from the SORT menu.

It can perhaps be seen from Figure 2. that while all the instances of *head* are clearly displayed, it is difficult to look through the equivalent French segments in order to locate possible French translations of *head* within each segment. To alleviate this, we can highlight suggested translations for English *head* by positioning the cursor in the lower French results window and clicking on the right mouse button. A menu pops up and we can select SEARCH QUERY which gives access to the usual search commands and hence allows us to enter a possible translation of *head* such as *tête*. The program then simply highlights all instances of *tête* in the French results window.

We can now change the context for the French results so that the results in the lower window are transformed into a KWIC layout (at least for those segments containing *tête.)* First, we make sure that the lower window is active. Next we choose CONTEXT TYPE from the DISPLAY menu and select WORDS. Finally, we rearrange the lines to bring those segments containing *tête* together at the top of the French results window. To achieve this, we choose SORT and sort the lines by searchword, and 1st left. The sorting procedure will then rearrange the results in lower window. (The SORT and DISPLAY commands are applied to whichever window is active.) The two text windows then appear as shown in Figure 3. Naturally, only those words in the French text that have been selected and highlighted can be displayed in this way. By sorting on the searchword, all the KWIC lines are grouped together at the top of the text window; the residue can be found by scrolling through towards the bottom of the window. This is a revealing display, but we have to be careful and not be misled by this dual KWIC display. There is no guarantee that for any particular line, the instance of *tête* is in fact

the translation of *head*. It could simply be accidental that *tête* is found in the French sentence corresponding to the English sentence containing *head.*

The idea behind dual KWIC display is to let the user move from English to French and back again, sorting and resorting the concordance lines, and inspecting the results to get a sense of the connections between the two languages at whatever level of granularity is relevant for a particular analysis.

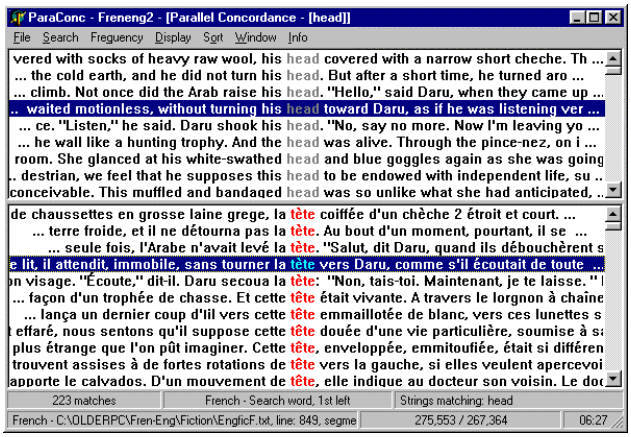


Figure 3: Parallel KWIC displays

4. Hot Words

In the previous section, we described the use of SEARCH QUERY to locate possible translations in the second window. In this section we will look at a utility in which possible translations and other associated words (collocates) are suggested by the program itself. We will refer to these words as *hot words*. First we position the cursor in the lower (French) half of the results window and click using the right mouse button. If we used SEARCH QUERY earlier, we need to select CLEAR SEARCH QUERY and then choose HOT WORDS, which invokes a procedure which calculates the frequency of all the words in the French results window and then brings up a dialogue box containing the ranked list of hot words. The ranked list of candidates for hot words based on *head* are displayed as shown in Figure 4.

To select words as hot words, the program looks at the frequency of each word in the results window and ranks the words according to the extent to which the observed frequency deviates from the expected frequency, based on the original corpus. The words at the top of the list might include translations of the searchword, translations of the collocates of the searchword, and collocations of translation of the searchword.

In addition to the basic display of hotwords, a paradigm option (if selected) promotes to a higher ranking those words whose form resembles other words in the ranked list. This is a simple attempt to deal with morphological variation without resorting to language-particular resources.

Some or all the hot words can be selected. Clicking on OK will highlight the selected words in the results window, and again the words can be sorted in various ways.

Figure 4: Hot Word List

5. Frequency information

ParaConc furnishes a variety of frequency statistics, but the two main kinds are corpus frequency and collocate frequency. The command CORPUS FREQUENCY DATA in the FREQUENCY menu creates a word list for the whole corpus (or parallel corpora), according to the settings in FREQUENCY OPTIONS. The results can be displayed in alphabetical or frequency order and the usual options (such as stop lists) are available.

Choosing COLLOCATE FREQUENCY DATA from the FREQUENCY menu displays the collocates of the search term ranked in terms of frequency. In *ParaConc,* the collocate frequency calculations are tied to a particular search word and so the frequency menu only appears once a search has been performed. The collocation data produced by the COLLOCATE FREQUENCY DATA command is organised in four columns, spanning the word positions 2nd left to 2nd right. The columns show the collocates in descending order of raw frequency.

One disadvantage of the simple collocate frequency table is that it is not possible to gauge the frequency of collocations consisting of three or more words. To calculate the frequency of three word collocations, it is necessary to choose ADVANCED COLLOCATION from the FREQUENCY menu and select one or more languages. The top part of the dialogue box associated with ADVANCED COLLOCATION allows the user to choose from up to three word positions, for example, SEARCHWORD 1ST RIGHT, 2ND RIGHT. The program counts and displays the three-word collocations based on the selected pattern.

6. Workspace

The loading and processing of a parallel corpus in particular can take some time since the program has to process alignment and annotation data before searching and analysis can begin. Since the same sets of corpus files are often loaded each time *ParaConc* is started, it makes sense to freeze the current state of the program, at will, and return to that state at any time, rather than starting *ParaConc* and reloading the parallel corpora afresh. This is the idea behind a workspace. A workspace is saved as a special (potentially large) ParaConc Workspace file (.pws), which can then be opened at any time to restore

ParaConc to its previous state, with the corpus loaded ready for searching. Searches and frequency data are, however, not included in the saved workspace. (Only the search histories are saved.)

A workspace can be saved at any time by selecting the command SAVE WORKSPACE or SAVE WORKSPACE AS from the FILE menu. The usual dialogue box appears and the name and location of the workspace file can be specified in the normal way. Once a filename for the saved workspace has been entered, the user is asked to choose some different workspace options. The line/page and the tracked tag info can be saved as part of the workspace. (The saved workspace consists of a saved file and an associated folder of the same name.)

7. Advanced Search

The simple searches described in Section 3 will suffice for many purposes and are especially useful for exploratory searches. The basic TEXT SEARCH is also very useful when used in conjunction with a sort-and-delete strategy. Particular sort configurations can be chosen to cluster unwanted examples (words preceded by *a* and *the* perhaps), which can then be selected and deleted. For more complex searches, however, we need to use the ADVANCED SEARCH command. This command brings up a more intricate dialogue box (displayed in Figure 5), which at the top contains the text box in which the search query is entered.

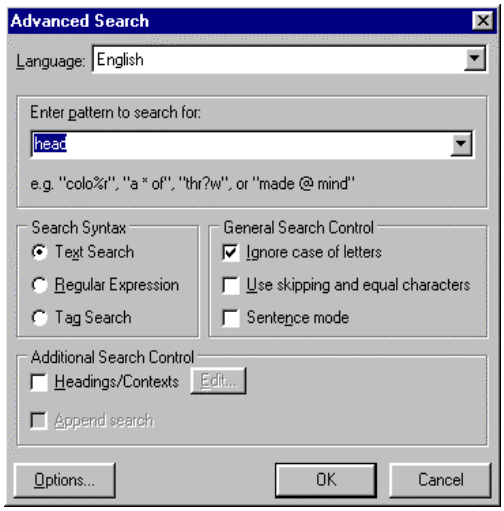


Figure 5: Advanced Search

The most important part of the ADVANCED SEARCH dialogue box is labelled SEARCH SYNTAX. The three radio buttons allow users to specify the kind of search they wish to perform. The first, TEXT SEARCH refers to the basic searches described in the section above.

The REGULAR EXPRESSION search allows for search queries containing boolean operators (AND, OR and NOT). For example, a regular expression to capture the *speak* lemma might be given as **sp[eo]a?k[se]?n?**. This expression will match the string *sp* followed by *e* or *o*, an optional *a*, a *k*., an optional *s* or *e*, followed by an optional *n*. (Word boundaries or spaces would also have to be specified in order to eliminate words such as *bespoke*.) The software also supports the expanded set of regex metacharacters: \d, \w, \s, \S, etc.

The third option in the advanced search dialogue box is TAG SEARCH, which allows the user to specify a search query consisting of a combination of words and part-of-speech tags, with the special symbol **&** being used to separate words from tags in the search query. This search syntax is used whatever particular tag symbols are used in the corpus. (Thus it is necessary to enter the form of the tags in TAG SETTINGS before a tag search can be performed.) To give an example: the search string **that&DD** finds instances of *that* tagged as a demonstrative pronoun, which may appear in the corpus as *that<w DD>*. Similarly, a tag search for **&JJ of&** will find all instances of adjectives followed by the word *off*. (The dialogue box in Figure 5 contains a variety of other options controlling the search function, which will not be discussed in this paper.)

Finally, one kind of search tailored for use with parallel texts is a parallel search, which is one of the options within the SEARCH menu. This type of search, shown in Figure 6, allows a search to be constrained based on the occurrence of particular strings in the different parallel texts.

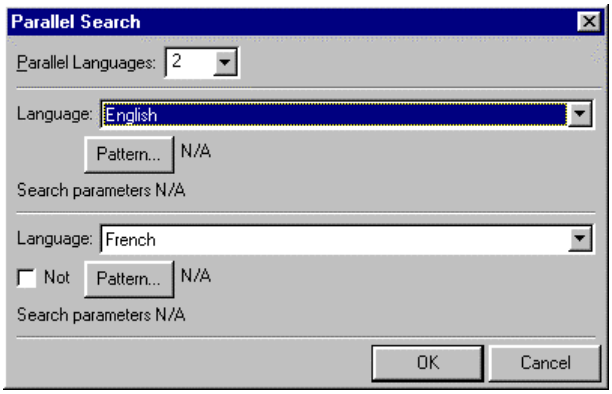


Figure 6: Parallel Search

Clicking on the Pattern box under Language: English brings up the normal advanced search dialogue box and a search query can be entered. In this case, the search term **head** has been entered. Moving to Language: French and again clicking on Pattern, it is possible to enter another search string such as **tête**. Clicking OK initiates the search routine and the software locates examples in which head occurs in the English text segment and tête is also found in the corresponding French segment. If the NOT box (under Language: French) is selected, then the search routine will display head only if tête does not occur in the equivalent French segment.

8. Summary

This paper has provided a brief overview of a Windows parallel concordance program which can be used by a variety of researchers working on the analysis of multilingual texts for translation or linguistic purposes. This article has focussed on the overall design and operation of the software and no linguistic analyses have been presented here, but the potential for cross-linguistic analyses and for the investigation of the translation process is, we hope, reasonably clear.

The main factor impinging on the usefulness of the software is probably the availability of aligned parallel corpora and of parallel corpora in general.

9. References

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

Corpora for Terminology Extraction – the Differing Perspectives and Objectives of Researchers, Teachers and Language Services Providers

Belinda Maia
Faculdade de Letras
Universidade do Porto
Via Panorâmica s/n
4150-563 Porto
Portugal
bmaia@mail.telepac.pt

Abstract

Using corpora to find correct terminology is an activity that is interpreted rather differently according to the final objectives of those involved. This paper will try to show how the perspectives and objectives of researchers, teachers and language services providers do not always coincide, and how this lack of mutual appreciation and understanding can sometimes cause confusion. We shall first look at the more speculative aspects of current terminology research for the possibilities they offer in the future, even though some of this work is not directly related to translation, and consider the reasons why correct terminology is growing in importance in the lives of both domain specialists and language services providers. We shall then briefly consider both the older prescriptive notions of standardisation and the descriptive approach made feasible by technology and corpora today. Corpora in the broadest sense – from formally constructed and officially approved collections of texts to the disposable, do-it-yourself corpora anyone can now collect off the Internet for information on a specific subject – come as part of the information revolution provided by technology. They provide possibilities for any user of language and knowledge that were unthinkable a few years ago, but there are also problems and drawbacks.

1. Introduction

The compilation of terminology used to consist largely of collecting the words and phrases considered to be specific to a certain domain and bringing them together to form glossaries, with or without definitions or information on how or where the information was gathered. Since translators often had a vested interest in finding, or providing recognised equivalents in several languages, these glossaries would often become bi- or multilingual at a later stage. With the increase in availability of electronic text, the advantages of using corpora for term extraction are now generally recognised, particularly since the prescriptive view of terminology work has given way to a more descriptive approach, and the storage of definitions and other information on the terms has been made possible by relational databases.

This paper assumes that there are three classes of people with a particular interest in this terminology work. First there are the researchers in various areas of linguistics in general, as well as more specific terminology research. Many, but not all of these people, are also the teachers who try to train the professional language services providers needed today. The word 'linguist' as someone proficient in two or more languages has become ambiguous since the advent of 'linguistics' as an academic discipline, and the tasks required of someone with a good knowledge of languages are increasingly varied. I have therefore chosen the term 'language services provider' to refer to those who not only provide traditional translation and interpreting services, but also those who write and revise texts professionally, specialise in localisation, sub-titling, dubbing and making web pages, create terminological databases and translation memories, work with machine translation, and both use and take advantage of the information technology now available for a wide variety of projects and customers.

2. Terminology research

Those involved in this workshop on translation work and research will tend to see terminology research as primarily interested in supplying the needs of the translator for specialised terminology, but this is only one aspect of the overall picture. A good deal of terminology research is monolingual in nature and directed at the standardisation and categorisation of the relationship between concepts belonging to certain domains of knowledge and the terms used to describe them. This type of work is typically carried out by the domain experts, with or without the assistance of linguists, and, more often than not, in major languages like English, French and German. The subsequent translation of these standardised terms into other languages is by no means as simple or as well organised as it might be, despite official efforts to the contrary.

Standardisation of terminology has a long history, and its objectives have typically been to prevent confusion in the transmission of knowledge, with all the economic, social, legal and political consequences involved. Some areas of knowledge, like engineering, have a long-standing tradition in producing standardised terminology, but even they find it difficult to keep up with technical and scientific developments. Many other domains have little or no organised terminology resources and what exists is often 'local' in nature, in the sense that it is the property of certain organisations, companies and other entities, of varying size and importance.

The information revolution caused by the Internet, however, has led to demands for better systematisation of knowledge and improved accessibility. For this reason, the computational side of terminology research today is increasingly orientated towards facilitating information retrieval and knowledge engineering (see Budin, 1996, and Charlet et al, 2001). Traditional terminology work tends to be painstaking and slow, and is not adapted to coping with the exploding need for retrieving knowledge. For this reason, efforts are being made by computational linguists and computer scientists to speed up the process of identifying, extracting and processing terminology (see Bourigault et al (Eds.) 2001, and Veronis (Ed). 2000).

3. Computational terminology

So much information is now processed in computer-readable form that there are obvious advantages to be drawn from this for machine (assisted) translation, translation memories and their related terminology databases. The corpora required for this type of research need to consist of texts that are not just well written, in the sense that they represent texts normally produced in a particular domain of knowledge: they need to use terms that are generally accepted in the community that works in that domain. When translations exist of these texts, they, too, need to conform to the same standards of text and terminology in the target language if one is to produce good aligned parallel corpora.

The experimental work done in computational terminology usually involves standardised texts in which both originals and translations are considered to be of high quality. Some of these texts have been provided by organisations like XEROX (see Bourigault 1994). The texts are often chosen for their linear compatibility (See Blank, 2001), which allows for easy alignment at, at least, sentence level, and the standardisation of their technical terminology. This is understandable, since it will only be possible to proceed with the analysis of a wider variety of texts when some sort of procedure has been worked out on the basis of these controlled corpora – rather as machine translation is better at translating controlled language than Shakespeare.

There is, of course, a lot of textual material that apparently conforms to the needs of this type of research. The European Commission has worked hard at making as many of its multilingual texts available as possible. In order to do this, the translation services have effectively created enormous translation memories full of texts translated by themselves, and one can presume that the terminology used is usually supported by the EURODICAUTOM database, which is itself the result of many years of effort by a large number of people. The large multinational companies that have invested heavily in translation memory software and terminology databases could also provide a vast amount of material. Organisations like the International Standards Organisation could provide invaluable material once its standards are efficiently translated in other languages. After all, not only do these standards and their translations represent ideal parallel corpora, but the very purpose of the texts themselves is to standardise the terminology used.

4. 'Real-life' terminology

There can be no doubt that a lot of the work to which we have just referred is impressive and of high quality and, therefore, a reliable source of information for the most necessary function of all these texts – the communication of knowledge. However, anyone who has worked seriously on producing terminology with the collaboration of experts will realise that the notion of 'one concept = one term' is an ideal, not a reality. International classifications that do exist have sometimes tried to escape the problems of normal language in different ways, as when natural species are classified in Latin, or chemical and mathematical concepts use formulas and symbols.

There are various reasons why the 'one concept = one term' notion is an ideal. It is easy enough for the linguist to understand the fluidity of the lexicon. After all, one of the perennial problems of general linguistics is how to deal with it in an easily classifiable way, hence all the work with projects like Wordnet (at: http://www.cogsci.princeton.edu/~wn/). On the other hand, experts in any particular domain are also aware of the fluidity of concepts and probably spend a good deal of time arguing about how to stabilise them for practical purposes - and stable terminology is only one aspect of this problem. In practice, they often resort to diagrams, images and other pictorial representations in order to circumvent or supplement the limitations of language. The general public, however, likes to believe in the stability of both language and concepts, and, for the practical purposes of communication, we all accept that there has to be some sort of 'social contract' whereby we agree to this stability in order to understand each other.

Prescriptive terminology has usually aimed at providing this stability in an organised fashion and most specialised dictionaries and glossaries are the result. The technology of databases, however, allows for a more descriptive approach, with all the implications this has for including all the information terminologists collect in the course of their work. When one is no longer limited by space on paper – a major factor in previous lexicographical work – the prospects of including all the information available and/or prescribed by international standards for terminological databases are, to say the least, tempting. These prospects may seem unnecessary to the more immediate problems of communication, but they contribute in no small way to various visions of the systematisation and documentation of knowledge.

Terminology is not the simple accumulation of words, their equivalents in other languages, definitions and a certain amount of grammatical information. Nor is it the simple matching of term to concept. One has to deal with all the usual problems of language - social, geographical, historical, political, and other aspects of style and register. At the level of standardisation, one can even become involved in authentic battles between academics or commercial companies who want to see the words they use to describe their particular theories or products prevail.

5. 'Real-life' corpora

When one is not working for the interests of computational terminology, one will probably not have access to the type of standardised corpora already described, except for the online documentation of the European Commission. Besides this, these standardised texts, no matter how well written or translated, tend to reflect a degree of deliberate homogenisation of style and register across languages. In the more routine terminology work carried out in universities and other institutions, every terminology project will come up against a different situation, and circumstances will play an important role.

First of all, one has to find what texts are available in the domain one is studying and it is more than likely that the most important ones will not be in digital form. We have found that this is often the case when one wants to use first-class academic texts published by well-known publishers. Working with industrial or commercial institutions or companies is one way of obtaining texts, but we have not yet tried this, partly because it will

require careful negotiation, and partly because we have found several academic partners interested in cooperating on a serious and more unbiased basis.

One can always scan texts, and there are, of course, plenty of texts already in digital form. It is often easy enough to obtain permission to use these texts if one explains why one needs them and what one intends to do with them, as there is plenty of interest among domain experts to see their terminology systematized. The Internet, as we all know, can provide an enormous amount of material in certain areas, but is less useful in others. For example, we have found it of limited interest for certain engineering terminology projects because both the high level expert-to-expert type of academic article and the more didactically orientated teaching text are not freely available to the general public. Too often one ends up with commercial sites trying to sell certain types of engineering equipment, and the information thus obtained is not necessarily very reliable. In the area of population geography, however, where one is dealing with a subject that cuts across the disciplines of geography, sociology and demography, one project group was able to find a sizeable amount of material in several languages, of both a parallel and comparable nature, precisely because there are plenty of official or governmental institutions who want to publish such material on-line. The other interesting aspect of this area is that the subject is relatively new and the relative instability of the terminology was observable in the texts found.

As our projects must have a Portuguese component, one of the problems we have found is that some languages are more equal than others. If the languages involved are English, French or German, there is a chance that one will be able to find reliable texts of a parallel or comparable nature, but the same will not be true of less used languages. We have found this to be true at all levels of text we look for. We have also found that the translations of websites - whatever the original language - are often of poor quality and cannot be used as parallel corpora.

6. Teaching and Project work

The type of project work we have done over the years started as a typical translation exercise in vocabulary research that owed much of its dynamics to the fact that the translation classroom contained PCs connected to the Internet. Our curriculum had been formulated by believers in the notion that 'general translation', together with six months placement at the end of the course, was sufficient for training Modern Languages students to become translators. Our experience, and that of our graduates, soon told us that this was far from enough and we developed specialised subject project work as a way of training students in LSP (see Maia, 1997 and Maia, 2000) within the limitations of the curriculum. We have now moved on to interdisciplinary postgraduate training in terminology and translation work, working with professors from the Engineering Faculty and History and Geography departments. Our early wordlists processed in Word have now developed into more sophisticated terminology work in Excel and Multiterm, and include definitions, sources, images and other data fields. We soon hope to have our own database system and make it available online.

Corpora have always been obligatory elements of our project work but, although we have collected quite a lot of specialised mini-corpora over the years, we admit that they have not always been the most successful part of the projects. There are various reasons for this. On the one hand, perhaps the biggest enemy of terminology related corpora work is the large number of existing on-line glossaries on everything under the sun that our students soon discover from each other. One can, of course, argue that these glossaries, which are often easy to copy or download, are in themselves language resources of the type we are discussing here. However, they are usually monolingual, largely in English, often rather general in scope, and infrequently backed up by any form of official recognition. When the glossaries are good, complete, and officially recognised, adding Portuguese terminology to them is usually beyond the scope of an undergraduate project. Of course, one might argue that beginners could do worse than discover how to convert them into their own languages.

The big problem here is that such work merely encourages the idea that finding the 'right word' is enough. This means they miss out on the didactic strengths of making mini corpora - the understanding of the subject itself, brought about by having to find and read texts, the appreciation of different types and styles of text gained while doing this, and the extraction of terms in context. Although students are encouraged to use software like Wordsmith to look for keywords and to study concordances of both general language words and specialised terminology, there is always a preliminary stage when the actual reading of the texts is necessary – at least from a pedagogical point of view. If they are lucky, they will also find definitions in the texts, although these are not as frequent, or as reliable, as the literature on the subject would have us believe.

There are successful types of glossary work that do not require corpora, such as some excellent ones our students have done on tools of various types – e.g. carpentry and gardening tools - in which the 'corpora' were largely catalogues with images, and students had to work hard to make the words in both languages match the pictures provided, a process that involved plenty of questioning of individuals, but little text work.

7. Conclusions

Corpora and terminology research can work well together, but they are not always equal partners. Ideally, students should be able to find good texts and extract terms, definitions and other information from them. When mini-corpora form the basis for terminology work, the process of producing the terminology project is didactically more valuable, and it is an easy step from collecting and aligning texts, and then using concordancing, to understanding the theory behind translation memories and other software and making them work in practice. As we have said, however, valuable terminology work can be done without resort to corpora. Perhaps the most important attitude to adopt towards project work is flexibility, since each domain brings its own circumstances and problems. If at the end of the experience our undergraduate students have learned how to take special languages seriously, the main objective has been achieved. Our postgraduate students already know

how important they are and need to learn how to progress further, and perhaps even join the process of research into computational processes that will speed up the accumulation of valuable resources for all of us who do not want to see the world speaking only one language.

8. References

Austermühl, Frank, 2001 *Electronic Tools for Translators*, Manchester: St. Jerome Publishing.

Blank, I., 2001. Terminology extraction from parallel technical corpora. In D. Bourigault, C. Jacquemin and M-C. L'Homme. 237-252.

Bourigault, D., 1994. *LEXTER, un Logiciel d'Extraction de TERminologie, Application à l'acquisition des connaissances à partir de textes.* PhD thesis. Paris: École des Hautes Études en Sciences Sociales.

Bourigault, Didier, Christian Jacquemin, & Marie-Claude L'Homme, (Eds.) 2001. *Recent Advances in Computational Terminology*. Amsterdam & Philadelphia: John Benjamins Publishing Co.

Budin, G., 1996. *Wissensorganisation und Terminologie*. Tübingen: Gunter Narr.

Charlet, J., M.Zacklad G.Kassel D.Bourigault, 2001. *Ingénierie des connaissances.* Paris: Éditions Eyrolles.

Maia, B. 1997. Do-it-yourself corpora ... with a little bit of help from your friends! In Barbara Lewandowska-Tomaszczyk and Patrick James Melia (Eds.) *PALC '97 Practical Applications in Language Corpora.* Lodz: Lodz University Press. 403-410.

Maia, B. 2000, Making corpora – a learning process. In Bernardini, S. & F. Zanettin, (eds). 2000: *I corpora nella didattica della traduzione*. Bologna: CLUEB pp.47-6.

Maia, B., (forthcoming),'Comparable and parallel corpora – and their relationship to terminology work and training', paper presented at the *CULT - Corpus Use And Learning To Translate*. Bertinoro, Italy, November 3-4, 2000.

Maia, B. (forthcoming). ' Terminology – where to find it, and how to keep it', Proceedings of *III Jornadas sobre la formación del traductor e intérprete,* Universidad Europea de Madrid 7 -10 March 2001.

Veronis, Jean (Ed). 2000. Parallel Text Processing – Alignment and Use of Translation Corpora. Dordrecht: Kluwer Academic Publishers.

Wright, Sue Ellen and Gerhard Budin, 1997. Handbook of Terminology Management – Volume 1: Basic Aspects of Terminology Management. Amsterdam & Philadelphia: John Benjamins Publishing Co.

2001. Handbook of Terminology Management – Volume II: Applications-oriented Terminology Management. Amsterdam & Philadelphia: John Benjamins Publishing Co.

Working Together: A Collaborative Approach to DIY Corpora

Lynne Bowker

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

Abstract

Corpora can be invaluable resources for translation students, but creating DIY corpora on a frequent basis can be a time-consuming exercise. This paper describes an experiment whereby the students in a translation class worked in collaboration to build corpora for use in their technical translation course. The guidelines used for this collaborative approach are outlined, and the results of the experiment are discussed. A general discussion on the value of the World Wide Web as a resource for building DIY corpora is also include.

1. Introduction

Researchers such as Zanettin (1998), Yuste (2000), and Bowker and Pearson (2002) have amply demonstrated the value of using corpora as translation resources in the context of translator training. However, there are relatively few "ready-made" or "off-the-shelf" corpora available for use in specialized domains, so translator trainers and/or students typically need to construct their own. This paper outlines an experiment that was conducted with 4th-year undergraduate students in a French-to-English technical translation course. The purpose of this experiment was to see if it was possible for the class to collectively build "DIY" or "disposable corpora" (Varantola, forthcoming) that could be used as resources for their translation course work.

My previous experiments with corpus building had proceeding following either a teacher-centred approach or a learner-centred approach. Both of these approaches had a number of drawbacks. In the case of the teacher-centred approach, the translator trainer was responsible for constructing all the corpora – a job which proved to be very time consuming (resulting in relatively small corpora) and which excluded the students from the design phase of the corpus building process. In the case of the learner-centred approach, each student was individually responsible for building his or her own corpora. This approach also proved to be inefficient, with students building corpora that were often small and generally poorly designed.

It was hoped that by adopting what Kiraly (1999 and 2000) and Yuste (2001) refer to as a learning-centred and collaborative approach, the resulting corpora would be larger and more useful, and the students would engage in active discussions with the trainer and with each other and would move towards becoming empowered critical thinkers and more independent learners.

2. Setting the parameters

In order to ensure that things ran smoothly during the collaborative exercise, it was necessary to first establish a number of guidelines or ground rules. The following strategy was developed and refined based on our experience over the academic year. It addresses the following issues: a) coordinators, b) number of texts contributed by each student per corpus, c) quality of texts, d) time frame, and e) file format.

2.1. Coordinators

For each corpus, two students would act as coordinators. When students were acting as coordinators, they did not have to contribute texts to the corpus (but they still had to do the actual translation homework). Essentially, the coordinators were to act as a sort of clearing house. Students in the class would e-mail their texts to a special account set up for the coordinators, who would 1) evaluate these texts for relevance, and 2) eliminate duplications (i.e., cases where the same text had been submitted by multiple students). The remaining texts would then be collated into a single corpus that would be posted on the class Web site.

2.2. Number of texts contributed by each student per corpus

Each student (with the exception of the coordinators) would try to identify three relevant texts that would make a good addition to the corpus. Given a class of between 20 and 30 students (this class had 22 students), this number was considered to be a reasonable goal; however, it was not an absolute. If a student could only identify two suitable texts, these would still be welcome; likewise, if a student located four or five relevant texts, they could all be submitted.

2.3. Quality of texts

The students agreed to put some time and care into selecting their three texts. It was noted that if everyone were to simply submit the texts corresponding to the first three hits that came up using a Web search engine, then there would be a lot of duplication and the texts may not be pertinent, which would limit the value of the corpus.

2.4. Time frame

In order for the process to run smoothly, a reasonable amount of time had to be given for both the contributions and the coordination. It was agreed that each target text would be distributed three weeks in advance. Students would have one week to identify suitable texts and e-mail them to the coordinators. The coordinators would have one week to check the texts for relevance and for

duplication, to amalgamate the texts into a corpus, and to post this corpus on the class Web site. All the students would then have one week to consult the corpus.

2.5. File format

Students e-mailed their contributions to the coordinators as attachments in plain text (ASCII) format. This simplified the job of the coordinators as it meant that they did not have to worry about having access to different types of computers or software packages and they did not have to manipulate different file formats. It also ensured that the corpus would be in a format that could be manipulated by the corpus analysis software to which the students had access (i.e., WordSmith Tools). In addition, it reduced the chances of spreading viruses.

3. Results of the Collaborative Corpus Building Exercise

In order to give some coherence to the course, the theme of "computer security" was selected and seven different source texts – each of a different text type and each focusing on a different subject relating to computer security – were chosen. Table 1 summarizes the corresponding comparable corpora that were compiled as part of the exercise.

4. Discussion

This section will outline strategies used by the students in selecting the texts and compiling the corpora; difficulties that were encountered and solutions used to overcome them will also be discussed. In addition, some general comments will be made on the suitability of the World Wide Web as a resource for building comparable corpora. Specific details about techniques used to extract translation-related information from the corpora have been detailed elsewhere in the literature (e.g., Bowker, 2000; Bowker and Pearson, 2002) and so will not be repeated here.

The first corpus to be constructed was on the subject "passwords", and the text type was a FAQ, which is a list of *Frequently Asked Questions* (and answers) about a given subject. In total, the students submitted 58 texts for possible inclusion in the corpus; however, there was a high degree of duplication and the final corpus ended up containing only 23 texts.

A class discussion following the creation of this first corpus revealed that most students preferred to use the Web to identify comparable texts. Other resources, such as CD-ROMs and online databases, were available in the university library; however, many students had Internet access from home and found it more convenient to work from there. Their preferred method of identifying texts for inclusion in the corpus was to read the source text and then select potential subject key words to enter into a search engine. In the course of the discussion, it was revealed that most students used the Alta Vista search engine, and many of them had not been very discerning when it came to selecting the three texts that they contributed – they often simply took the first three hits that came up. In order to identify a wider selection of texts for future corpora, students agreed to make an effort to look beyond the first three hits. Moreover, we discussed the fact that different search engines index different Web sites, which means that the hits returned by one search engine may be different than those returned by another. Students agreed to use a wider range of search engines (and meta search engines) when looking for comparable texts, and it was hoped that by doing this, there would be less duplication in future corpora.

The next three corpora were intended to help translate an instructional text on "antivirus programs", a popularized informative text about "encryption", and a buyer's guide for "firewalls". In the world of computer security, these are all popular subjects and common text types, so there was a lot of information available. In particular, popularized informative texts are among the most common type of text on the Web, and many of the texts identified by the students were quite long, which elevated the word count of the encryption corpus considerably. Given that there were many texts to choose from, a number of students submitted more than three texts each. Moreover, the degree of duplication for these three corpora was reduced as a result of the students' efforts to use different search engines and to look beyond the first three hits.

The corpus on "steganography" was supposed to be used to help students translate a product description. Steganography is much less common than other security measures and there are a limited number of products on the market. Consequently, the students found that there were fewer texts to choose from with the result that only 35 texts were submitted, and of these, only 14 were retained. Of the texts that were rejected, many were duplicates; however, the coordinators also rejected a number of texts that were not of an appropriate text type. Given the relative scarcity of comparable texts, some of the students had submitted texts that were about steganography, but which were not product descriptions. Similar behaviour has been observed by Pearson (2000), who notes that translation students sometimes show poor judgment when sourcing terminology and phraseology from comparable texts. For example, they are often primarily concerned with identifying texts that deal with the subject matter in question, but they do not ensure that the texts they choose are comparable to the source text with respect to its other features, such as register, technicality and text type. In a class discussion, the matter was raised and it was emphasized that in order for a text to be "comparable", it had to take into account text type as well as subject matter.

The source text on biometrics was an extract from a research article. There were 29 comparable texts submitted, but only 12 of these were retained. However, since research articles tend to be long, the word count was still reasonably high. The main problem that the students had was in finding the relevant text type on the Web. Although there were a number of hits that looked promising, many of these links led to Web sites that required a paid subscription in order to gain full access to the contents of the site (e.g., online journals). This led to a discussion about other non-Web resources that may be useful for building corpora, including the *Computer Select* CD-ROM, INSPEC abstracts and a variety of online

journals that were part of the university's library collection. It was noted that although students would rather work from home (hence their preference for consulting the Web rather than the library databases), it was not unreasonable to expect them to make a trip to the library in order to consult more appropriate resources.

Subject	Text type	Texts submitted	Texts rejected	Number of texts / words in corpus
Passwords	FAQ Web page	58	35	23 texts / 40,600 words
Antivirus programs	Instructional	78	22	56 texts / 170,919 words
Encryption	Informative/popularized	74	19	55 texts / 216,522 words
Firewalls	Buyer's guide	63	18	45 texts / 136,017 words
Steganography	Product description	35	21	14 texts / 7,401 words
Biometrics	Research article	29	17	12 texts / 69,651 words
Cookies	Technical encyclopedia entry	41	19	22 texts / 11,754 words

Table 1: A brief description of the corpora produced as part of the collaborative corpus building exercise.

Finally, the source text on "cookies" consisted of an entry taken from a technical encyclopedia. Once again, there were relatively few submissions (41 texts), coupled with a high degree of duplication (only 22 texts were retained). This was because there are a limited number of electronic technical encyclopedias that could serve as comparable texts. Furthermore, it was observed that the entries in such encyclopedias tend to consist of short texts, which resulted in a relatively low word count for the corpus as a whole.

5. General observations about using the Web as a resource for building DIY corpora

In addition to discussing particular problems that came up when creating specific corpora, the class also discussed a number of more general points, many of which concerned the nature of the Web and its suitability as a resource for building DIY translation corpora. For example, it was noted that there are many texts on the Web that are of poor quality and which therefore do not make good translation resources. When asked to reflect on potential reasons for this poor quality, students came up with the following possibilities. Firstly, they noted that anyone can post information on the Web, including non-subject field experts and non-native speakers, and that Web documents are not always subject to an editing process in the same way that printed documents usually are. Furthermore, the Web is seen by many as an ephemeral resource; people are interested in communicating information, but unlike the case with printed documents, this information may not be preserved for long (i.e., a Web page can be revised, updated or removed very easily) and so people are less willing to invest much time or effort in formulating that information. In other words, many people feel that a Web page does not need to be elegant (or even grammatically correct!) as long as it adequately conveys the essential information.

Another comment focused on the types of texts that are commonly found on the Web. Given that the Web is most often used as a means of disseminating information to a non-expert audience, it contains primarily informative or instructional texts that are popularized. More specialized material and different text types can be accessed via the Web, but such information is often available only by paid subscription. This means that while the Web can a valuable resource for constructing corpora that deal with popularized informative texts, it may prove less helpful for constructing corpora that must comprise other types of texts.

A similar observation was made about the languages of texts available on the Web. The students in this class were attempting to compile comparable corpora containing English-language texts, of which there are many on the Web; however, they noted that for translators working in less widely-used languages, there may be fewer texts available (at least for the present, though hopefully this will change over time).

The very nature of the Web gave rise to two other observations. Firstly, the idea behind hypertext is that people can jump from page to page to view associated information. Good Web design dictates that there should be a limited amount of information on each page so that people are not required to scroll unnecessarily; related pieces of information should be provided on separate pages with relevant links between them. When compiling a corpus from the Web, each page must be copied/saved separately and then later amalgamated into a corpus. Therefore, from a corpus builder's point of view, it would be preferable to have a single page containing a lot of information, as this page could be copied/saved in one operation, rather than having that same information spread over several pages, which would then need to be copied/saved separately. This basically means that good Web design is not conducive to easy corpus building! Secondly, the multimedia nature of the Web is another characteristic that is not always conducive to building text-based corpora. On a number of occasions, students rejected Web pages that would have been extremely useful sources of information but which could not easily be incorporated into a text-based corpus because their primary value resided in their graphical or audio content. This raises an important point: a corpus can be an invaluable resource, but it is not a panacea. There are many other complementary types of resources that can

also provide helpful information, and these should not be ignored.

Finally, the sheer volume of information that is available on the Web made students aware of the importance of formulating search queries carefully in order to be able to focus in on relevant material. As previously mentioned, students tended to read the source text first in order to get ideas for potential key words. These words were then entered into a search engine, and the resulting hits were examined for relevancy as well as for ideas for other key words that could be used for further searches. In addition to key words that dealt with the subject matter, students also found that it could be useful to enter key words relating to the text type. For instance, a search using only the subject key word "cookie" returned many irrelevant texts such as recipes; however, a more carefully formulated search that combined subject and text type key words, such as *+cookie +computer +encyclopedia*, returned hits for entries for "cookie" in resources such as *The Grand Encyclopedia of Computer Terminology*, *TechEncyclopedia* and *PC Webopedia*. Other tricks, such as remembering to search for alternate spellings (e.g., encyclopedia/encyclopaedia) also helped to increase the number of relevant hits. In addition, as mentioned previously, the students also found it useful to conduct a search using a variety of different search engines or a meta-search engine. Bergeron and Larsson (1999) provide additional tips for effective Internet search strategies for translators.

6. Concluding Remarks

Overall, the collaborative corpus building exercise proved to be a worthwhile experience. The students demonstrated that they were eminently capable of working together to construct valuable translation resources, which they could then consult to identify relevant lexical, phraseological and stylistic information. Not surprisingly, of the seven collective corpora that were built, the larger ones, such as those on antivirus programs and encryption, tended to contain a greater number of examples. Of more interest, however, is the fact that even the small corpora, such as those on steganography and cookies, contained useful information. This supports the point made by researchers such as Rogers and Ahmad (1994), who note that when working in specialized fields, it is not necessary to have the sort of multimillion word corpora that are typically required for general language work.

In addition to furnishing students with an opportunity to explore the merit of corpora as translation resources, this exercise also provided a valuable opportunity for a shift in pedagogical strategy. The collaborative corpus building exercise made it relatively easy for the trainer to take on the role of facilitator (rather than information provider), which in turn allowed the students to become independent learners and critical thinkers, who were encouraged to reflect on the characteristics of different text types and on the suitability of the World Wide Web as a translation resource. Acting as both contributors and coordinators, students learned to identify relevant features of texts and to be more discerning with regard to the appropriateness of a text (e.g., in terms of quality, text type, nature) for use as a resource for the translation at hand.

7. Acknowledgements

The work described here has been partially funded by grants awarded to Lynne Bowker by the Faculty of Arts of the University of Ottawa and the University of Ottawa Research Fund.

8. References

Bergeron, Manon and Susan Larsson, 1999. Internet Search Strategies for Translators. *The ATA Chronicle* 28(7): 22-25.

Bowker, Lynne, 2000. Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources. *International Journal of Corpus Linguistics* 5(1): 17-52.

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

Kiraly, Don, 1999. From Teacher-centered to Learning-centered classrooms in translator education: control, chaos or collaboration? In *Innovation in Translator and Interpreter Training (ITIT)* – an online symposium held from January 17-25, 2000) http://www.fut.es/~apym/symp/kiraly.html

Kiraly, Don, 2000. *A Social Constructivist Approach to Translator Education*. Manchester: St. Jerome.

Pearson, Jennifer, 2000. Surfing the Internet: Teaching Students to Choose their Texts Wisely. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.

Rogers, Margaret and Khurshid Ahmad, 1994. Computerised Terminology for Translators: The Role of Text. In M. Brekke, O. Andersen, T. Dahl and J. Myking (eds), *Applications and Implications of Current LSP Research, Vol. II*. Norway: Fagbokforlaget.

Varantola, Krista, Forthcoming. Translators and disposable corpora. In F. Zanettin, S. Bernardini and D. Stewart (eds), *Corpora in Translator Education*, Manchester: St. Jerome.

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

Yuste, Elia. 2000. Translation Instruction in the Y2K – Electronic Corpora, Internet and Translation Technology. In *CD-ROM Proceedings of the Seventh Conference of the International Society for the Study of European Ideas (ISSEI 2000)*, Workshop 501 - Teaching Translation in the Information Age. University of Bergen, Norway.

Yuste, Elia. 2001. Technology-Aided Translation Training. *Hieronymous* (3). Bern, Switzerland: ASTTI.

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

Language resources and the language professional

Elia Yuste

Computerlinguistik (CL)
Institut für Informatik (IfI) der Universität Zürich
Winterthurerstrasse 190, CH-8057 ZÜRICH, Switzerland
yuste@ifi.unizh.ch

Abstract

This paper aims at raising awareness about electronic language resources (henceforth LR) in the translation community at large. Examining how technological advances in the profession have transformed the notion of translating itself and what is expected from a qualified translator today, the paper goes on to focus on resources, rather than tools. It then discusses what type of LR should feature in the training of professional translators, and how these should be tackled in various translation-training settings. It contains several useful pointers throughout the article and an extensive bibliography covering the various issues addressed herewith.

Keywords: translation profession, language professional, qualified translator, translation training, tools, resourceful, resources, language resources (LR), corpora, translation technology and HLT, academic training, vocational training, collaborative approach, real-life scenarios, translation workflow, multi-user access, corporate language, content management, resource creation / maintenance / evaluation / validation / exchange, exchange standards

1. Introduction

Traditionally speaking, translation has been regarded as a craft, a fairly unusual gift that, for some, did not even require formal academic training, let alone continuous education on (technological) advancements in the profession. From that standpoint, the translator's major asset, and only utensil, is his or her own competence for translating, that is, some special ability to transpose meaning from one language to another. But even if natural linguistic talent is always desirable, translators cannot solely rely on it to succeed as language professionals today. Translating has become a complex and permeable professional activity, which among other things requires plenty of intercultural sensitivity and disposition to adapt to new work patterns.

In fact, professional and qualified translators (against the unqualified intruders that slip in the translation profession) do usually gain respect and recognition (and in practical terms, are more employable) for being *resourceful* and acquainted with the *tools of the trade*. But what do we mean by 'resourceful' here? 'Resourceful' in that they are expected to be capable of resolving linguistic problems (and/or cultural misinterpretations) efficiently and at once? Or perhaps, 'resourceful' in that they ought to be familiar with resources that allow them to find the right information at a mouse click? What 'tools of the trade' do we refer to? Are commercial translation memory[1] packages the hot tools for translators, the one and only?

1.1. Tools …AND resources, please!

Up to the late 20th century's information revolution, heavily characterized by the advent of the personal computer (PC), the so-called ICT[2], and the Internet, the translator's *tools* had primarily been pen and paper (without forgetting about the now old typewriter and the Dictaphone®). Of course, *paper* understood in its broad sense (different sizes, textures, colours…) as a means to **manually** catalogue, archive and, hopefully, retrieve – throughout the years – translation notes, bibliographical references, and laborious samples of terminographic work. Undoubtedly, these were extremely valuable (and praiseworthy) *self-made resources* under a not very convenient support.

Other conventional translators' *resources*, linguistic and non-linguistic, consist of printed dictionaries and reference materials (such as voluminous encyclopedias –now online[3]), as well as certain cultural and/or domain-specific knowledge, gradually acquired through reading, visits to libraries, travelling, life experience and, sometimes, long discussions with fellow translators and subject experts over a cup of coffee.

Although the latter still works for some translators to some extent, the newer generations normally resort to other (quick-access) information sources and data processing applications, usually computer (e.g. on CD-ROM or DVD) or Web based, in order to accomplish their translations. Not surprisingly, 'tools' and 'resources' often get listed as *useful links* in Web sites and other publications for the translator, without making much of a distinction between them. However, I would like to see these two concepts differentiated (despite their undeniable affinity – and even interdependence[4] – in today's translation workflow),

[1] *Translation tools* have become the buzzword in translation educational and work contexts. By and large, they are usually identified with translation memory (TM) packages, the apparently ideal solution for a cost-effective and consistent translation. Yet, apart from these tools managing and reusing previously translated repetitive input, translators also ought to get to know about tools that allow them to create, retrieve, exploit, interconnect, and exchange…*language resources* (LR) – simply because LR are their most precious resources.

[2] Acronym of 'Information and Communication Technologies'.

[3] E.g. *Encyclopaedia Britannica* (http://britannica.com).

[4] If a translator uses a terminology management program to manage their terminology records, then the program itself would be the *tool* whereas the resulting records would be the

since this paper will concentrate upon *resources*, rather than *tools*.

In essence, *tools* should refer to those instruments or equipment (e.g. ball-pen, computer, printer, software program, etc.) that translators use in their daily work or that they need for a particular job assignment (e.g. a concordancer[5] for automatic term extraction, the comment utility of a word-processing software for proof-reading, etc). But equally important are *resources* (e.g. corpora, dictionaries and reference materials, glossaries and terminological databases, etc.), especially *language resources* (henceforth LR), since these are useful elements in the translation process and contribute to enhancing the translator's professional profile.

1.2. LR and HLT applications – something to equip the language professional, too

Moreover, in the area of HLT[6], where translation technology indisputably has its place, LR can be essential *components*. Without them, many research and real life systems would not see the light. Godfrey, J. J. and A. Zampolli (1996) thus define LR as '…(usually large) sets of language data and descriptions in machine readable form, […] used in building, improving, or evaluating natural language (NL) and speech algorithms or systems. Examples of linguistic resources are written and spoken corpora, lexical databases, grammars, and terminologies, although the term may be extended to include basic software tools for the preparation, collection, management, or use of other resources.'

Apart from offering us an overview of LR, Godfrey & Zampolli stress the fact[7] that LR may be extended and used to elaborate other resources, then including or interacting with tools. This is an important aspect for translation work and research. LR are usually conceived with a purpose in mind, but they may serve other purposes later, by being expanded, tailored to the needs of another user-group or integrated in a system. For instance, a paper-based glossary is linguistically-enriched (i.e. annotated or marked-up) and transformed into electronic form to become available in an organization's intranet; a few navigation and edition tools are added to allow for rapid cross-referencing and regular content updates. Some time later, this and other LR are part of a new terminological workbench, also accessed by translators and domain expert validators working for the same organization. Since the time this resource gets digitized, its lifecycle varies dramatically according to its functions and targeted user-groups.

The *resourceful* language professional[8], interested in the advances of the profession, should thus be able to create, use, and evaluate those LR serving their job or area of specialisation needs better. Translation training programmes should then prioritize topics related to LR creation, manipulation, and evaluation.

2. Goal of the paper

This paper therefore aims at discussing the importance of resources, in particular LR, shaping every facet of translation (the training of translators, the profession itself, translation as part of global content production, etc.). Ideally, our translator will be conceived as an eclectically evolving, and qualified language professional, rather than as a word artist exclusively.

3. LR in the training of translators

In order to response to revolutionized translation work patterns, most translation training institutions have incorporated some technology-related elements within their syllabuses, but it still remains unclear whether they are sufficient and efficient enough. Whereas at the beginning much emphasis was given to introductory modules on IT[9], most recently some commercial translation memory packages seem to be getting all the attention.

In 1999, the LETRAC[10] commission reported that in the surveyed translation training institutions[11] across Europe, 'LE/IT [not expliciting the concept of LR, though] in translator curricula vary from nothing but basics in word processing to a broad range of sophisticated software tools (terminology management, translation memory, machine translation, Telecommunications / Internet, CD-ROM-based information systems...).' Also of interest are their

resource. But, obviously, given this interdependence between tool and resource, one might argue that there is a very fine line between the two.

[5] A *concordancer* is a software application aimed at retrieving *concordances* (an automatic display of a word or phrase occurrence/s, known as KWIC – key word in context, surrounded by left and/or right accompanying words) from a text or corpus previously loaded. As this tool allows for rapid linguistic insight of any word, it is of great value for the linguist, lexicographer, or translator. This is why most translator workbenches include now a concordancer among their growing panoply of utilities.

[6] Acronym of 'Human Language Technologies'.

[7] Fact also reported by OVUM (1995): 'In order to provide users with a working system adapted to their environments, […] linguistic resources may also include the ability to create other bi-lingual, multi-lingual or reversible dictionaries to provide terminology quickly in other language pairs'. The potential multi-user access is also highlighted.

[8] The term *language professional* is normally applied to translators, who do not perceive their professional activity restricted to translation in its traditional sense. It may also be applied to other professionals working with language, such as terminologists, proof-readers, cross-cultural multilingual advisers, content managers, etc. They all are key *language industry* players.

[9] Acronym of 'Information Technology'.

[10] LETRAC - *Language Engineering for Translators Curricula*. EU-funded research project that run from 1998 to 1999, whose aim was to survey best practices in the training of translators enhanced by language engineering (LE) components.
http://www.iai.uni-sb.de/LETRAC/home.html

[11] Reuther, U. (ed.). April 1999. 'LETRAC survey findings in the Educational Context', Deliverable D1.2.

observations[12] on how (the type of) training has a determining effect on translators' professional success:

- 'A translator does not only perform translation.
- Training in IT should be obligatory.
- Translators do not feel well prepared by their institutions for the real world of work.
- Translators gained their present LE/IT knowledge mainly from work experience, by means of "learning by doing".
- Among freelancers, two extremes can be observed: those translators who follow the principle *as little IT as possible*, and those who can cope with virtually all aspects of new technologies. The latter are those who do better economically.
- Most translated texts are LSP[13] texts; therefore specialised translation and terminology should be an essential element in curricula.
- There is a lack of qualified IT-specialists on the translation market. Translators with LE/IT-skills have far better professional prospects.'

These reflections show the big challenge for translation training institutions posed by global language market needs, described by Shreve (1998:5) as 'an evolution in fast-forward', highly dominated by the areas of multilingual technical communication and software/web localization.

3.1. LR in academic training

Plenty of translation scholars and researchers have advocated the use of corpora in the classroom, presented them as invaluable analytical resources in TS[14] (among others, see Austermühl 2001, chapter #8, Baker 1992/3/6/9, Bernardini & Zanettin 2000, Bowker 2000a/b and 2001, Kenny 1998, Laviosa 1997, Pearson 1996/8, and 2000, Tognini-Bonelli 2000, Ulrych 1997, Yuste 2000/1, Zanettin 2000/1 and forthc., as well as Hansen and Teich, Olohan, Zanettin, Maia, and Bowker, in order of appearance in this vol.), and also created tools for their exploitation or access (see Badia et al., and Barlow, this vol.). However, it appears that a generalized systematic inclusion of LR, mainly corpora, in translation training curricula still remains a necessity (Yuste, forthcoming), especially in places where English is not an official or a tuition language.

It is beyond the scope of this paper to advocate again for corpora in translation training scenarios. Yet, it is relevant to bear in mind that translators 'need, above all, to acquire a sound knowledge of the raw material with which they work: to understand what language is and how it comes to function for its users' (Baker, 1992: 4). This is better achieved through meaningful training activities whereby future translators look into authentic (against pre-fabricated) language instances in context. Besides, corpora allow the translator trainer to keep a steady balance between theoretical linguistic insights and practical applications.

Most importantly, one should not forget that many aspects of corpus linguistics (e.g. concordancing, alignment, parallel corpora) are present in current and future language/translation technology applications. Future translators should be made aware of the fact that the commercial TM package available in their lab contains such and such corpus linguistics features. It is only when modern tools for the translator are presented comprehensively and, if necessary, theoretically backed-up, that the translator can fully understand the mechanisms behind the tool. He or she is then also empowered to make the most out of the tools or applications at hand.

Tools such as translation and localization workbenches, knowledge and content management systems, to name but a few, are usually solutions which get constantly fed with linguistic data, i.e. LR such as corpora. Under such circumstances, it is important to promote research linked to market needs, e.g. fostering of LR *exchange*[15] *standards* or *reusability* (see Kübler, this vol.). An ideal first step is to get future language professionals involved in the creation and maintenance of resources, such as (DIY) corpora (see Zanettin forthc. and Zanettin, Maia, and Bowker, this vol).

In this line of work, it is important to follow *collaborative* (see Kiraly 1999/2000 and Yuste 1999/2001) and *project-based* training approaches, whereby future translators do not only learn about how to create or exploit shared LR but also get used to teamwork, project management, etc. – skills so highly appreciated in corporate and institution settings where cross-site language work is crucial.

3.2. LR in vocational or continuous training

Most technology-related vocational or continuous training courses on offer for future and practicing translators deal with TM systems or localization tools, sometimes with little reference to LR, such as corpora. Software tools producers (usually their marketing-oriented training departments), translation training academic departments (often postgraduate course modules devoted to translation technology, which may be opened to an external audience), and occasionally translators' societies or bodies organize these courses, whose training quality and price can vary considerably.

Their merit is mainly to aim at compensating for the lack of up-to-date technology-aided translation training in formal academic settings. These courses, heavily market-oriented, should nevertheless employ solid

[12] Reuther, U. (ed.). April 1999. 'LETRAC survey findings in the Industrial Context', Deliverable D2.2.

[13] Acronym of 'Language for Specific Purposes'.

[14] Acronym of 'Translation Studies'. Note that the impact of corpora in TS has lead to *Corpus-based Translation Studies* (CTS), with M. Baker as one of the main precursors.

[15] In that corpora, terminologies, language ontologies, output from TM systems, etc. may represent valuable LR not for the resource creator or first intended user-group only, LR have to be conformant to formats so that they can be exchanged, made accessible to other user-groups or integrated into other applications. For more information on recently agreed standards, such as TMX and TBX, see specifications drawn from the SALT Initiative and Abaitua (2001), Budin et al (1999), Budin & Melby (2000), Budin (2002), and Zerfass (this vol.).

training principles (see previous section) and real-life application scenarios (i.e. full description of interrelated components, usefulness of the tool within overall workflow, satisfaction and benefits for the translator, etc. instead of a mere exposition of reduction of costs).

3.2.1. Training on LR at the workplace

When the course takes place at the workplace, it is of utmost importance to analyze what the needs for LR (and any form of translation/language technology) are, not only for translators or linguists, but also for other staff members, such as resource evaluators and domain experts.

Similarly, it is advisable to look at LR from the *corporate language* (or even institution-wide language) perspective, and see how they may contribute to optimizing (global) multilingual documentation production. For example, to learn how to create corporate databases (product names, enterprise-wide terminology) helps reinforce a company's image, promoting clear, consistent communication and aiding cross-cultural understanding. *Controlled language* (see Fankhauser 2000) schemes (e.g. multilingual corporate style guides for written documents of all kinds) and *content management* (see Budin, this vol.) strategies may have to be implemented.

Finally, similar initiatives/solutions developed by other language industry players and organizations of the same sector will have to be carefully examined, not to reinvent the wheel. Ideally, language professionals (and other LR user-groups) will have to be able to maintain, customize and tailor existing LR, as budget controls may prevent them to create their own. Sharing and exchanging LR with other partners will be essential, and so it will be training focused upon LR exchange standards (see footnote #15).

4. Conclusion

Despite the length limit of the paper, we have attempted to discuss the relevance of language resources (LR) for the translator and the rapidly evolving translation profession in a comprehensive and up-to-date manner.

LR are crucial to transform the qualified translator into a *resourceful* language professional, able to respond to any challenge, and enhance any translation-related workflow. But, of course, nothing of this is possible without adequate and tailored LR training be it in an academic setting or at the workplace.

5. References

Abaitua, J. 2000. Tratamiento de corpora bilingües. Paper presented at the *La ingeniería lingüística en la sociedad de la información* Seminar, hosted by the Fundación Duques de Soria, Soria, Spain. 17th - 21st July 2000. http://www.serv-inf.deusto.es/abaitua/konzeptu/ta/soria00.htm and http://www.serv-inf.deusto.es/abaitua/konzeptu/ta/sorefs00.htm (paper references).

Abaitua, J. 2001. Memorias de traducción en TMX compartidas por Internet. In *Revista Tradumática*. Número 0 – October 2001. http://www.fti.uab.es/tradumatica/revista

Austermühl, F. 2001. *Electronic Tools for Translators*. Translation Practices Explained series (A. Pym, series editor). Manchester: St. Jerome.

Baker, M. 1992. *In other words – a coursebook on translation.* London / New York: Routledge.

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

Baker, M. 1995. Corpora in Translation Studies. An Overview and Suggestions for Future Research. In *Target* 7(2): 223-43.

Baker, M. 1996. Corpus-based Translation Studies: The Challenges that Lie Ahead. In Sommers, H. (ed.). *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, Amsterdam / Philadelphia: John Benjamins.

Baker, M. (ed.). 1998. *Routledge Encyclopedia of Translation Studies*. London / New York: Routledge.

Baker, M. 1999. The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators. In *International Journal of Corpus Linguistics* 4(2): 281-298.

Bernardini, S. & F. Zanettin (eds.). 2000. *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate*. Bologna, Italy: CLUEB.

Bowker, L. 2000a. The translator as LSP learner: Using an electronic LSP corpus as a translation resource. In M. Ruane and D.P. Ó Baoill (eds). *Integrating Theory and Practice in LSP and LAP*. Dublin: IRAAL, 85-91.

Bowker, L. 2000b. A Corpus-based Approach to Evaluating Student Translations. In *The Translator*, Vol. 6(2), 183-210.

Bowker, L. 2001. Towards a Methodology for a Corpus-Based Approach to Translation Evaluation. In *Meta*, Vol. 46(2), 345-364.

Budin, G. et al. 1999. Integrating Translation Technologies Using SALT. In *Proceedings of the 21st International Conference on Translating and the Computer*. London, 10th – 11th November. London: ASLIB.

Budin, G. and A. Melby. 2000. Accessibility of Multilingual Terminological Resources – Current Problems and Prospects for the Future. In Zampolli et al. *Proceedings of LREC*. Athens, June 2000. 837 ff.

Budin, G. 2002. Der Zugang zu mehrsprachigen terminologischen Ressourcen – Probleme und Lösungsmöglichkeiten. In Mayer, F., K.-D. Schmitz, and J. Zeumer (eds.). *eTerminology* – Akten des Symposions. Deutscher Terminologie Tag e.V. Cologne, 12th – 13th April, 2002.

Chriss, R. 2000. *Translation as a Profession*. Available online at:

http://www.foreignword.com/Articles/Rogers/default.htm

Esselink, B. 2000. *A Practical Guide to Localization*. Amsterdam / Philadelphia: John Benjamins.

Godfrey, J. J. and A. Zampolli. 1996. Overview [of Language Resources]. Subsection 12.1 of Chapter 12, *Language Resources*, by Cole, R. (ed.). In G. B. Varile and A. Zampolli (managing eds.). *Survey of the State of the Art in Human Language Technology*. Sponsored by the National Science Foundation and the European Commission. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

Fankhauser, R. 2000. Corporate Language Management. Paper presented at the *tecom Schweiz* Seminar with the same title, organised by Iris Jahnke (Trados Switzerland). Hotel Mövenpick, Zurich. 31st August, 2000. Paper available in German from *tecom Schweiz* (Swiss Society for Technical Communication) at: http://www.tecom.ch ('Publikationen' section).

Kenny, D. 1998. Corpora in Translation Studies. In Baker, M. (ed) *Routledge Encyclopedia of Translation Studies*, 50-53.

Kiraly, D. 1999. From teacher-centered to learning-centered classrooms in translator education: Control, chaos or collaboration? In *Innovation in Translation and Interpreting Training - ITIT*, an online symposium (17th – 25th January, 2000) organised by Anthony Pym. Intercultural Studies Group. Universitat Rovira i Virgili, Tarragona, Spain. http://www.fut.es/~apym/symp/kiraly.html

Kiraly, D. 2000. *A Social Constructivist Approach to Translator Education – Empowerment from Theory to Practice*. Manchester: St. Jerome.

Laviosa, S. 1997. How Comparable Can 'Comparable Corpora' Be? In *Target* 9(2): 289-319.

OVUM Report. 1995. Mason, J. and A. Rinsche. *Translation Technology Products*. OVUM Ltd., London.

Pearson, J. 1996. Electronic texts and concordances in the translation classroom. In *Teanga* 16. Dublin: IRAAL. 86-96.

Pearson, J. 1998. Teaching terminology using electronic resources. In S. Botley, A. McEnery and A. Wilson (eds.) *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi. 92-105.

Pearson, J. 2000. Surfing the Internet: Teaching students to choose their texts wisely. In Burnard, L. and McEnery, T. (eds). *Rethinking language pedagogy from a corpus perspective*. Papers from the 3rd. International Conference on Teaching and Language Corpora. (Lodz Studies in Language). Hamburg: Peter Lang, 2000.

Robinson, D. 1997. *Becoming a Translator – An Accelerated Course*. London / New York: Routledge.

Sager, J.C. 1994. *Language Engineering and Translation: Consequences of Automation*. Amsterdam / Philadelphia: John Benjamins.

SALT: Standards-based Access service to multilingual Lexicons and Terminologies. More information about this project is available at the TTT – Translation, Theory and Technology site (http://www.ttt.org/salt/), hosted by A. Melby, and at the project's site (http://www.loria.fr/projets/SALT/)

Shreve, G. M. 1998. The Ecology of the Language Industry: Prospects and Problems. Keynote Address of the Language in Business / Language as Business Conference. Institute for Applied Linguistics, Kent State University, Kent, Ohio, USA. October 8, 1998. Available as a .pdf document at: http//appling.kent.edu/ResourcePages/ConferencesandWorkshopsPast/LanguageinBusiness/Thursday/OI-Shreve.PDF

Theologitis, D. 2000. Language Technology in EU Institutions. In Proceedings of the EAMT Machine Translation Workshop "Harvesting existing resources". Ljubljiana, Slovenia. May 2000. Online presentation at: http://www.eamt.org/archive/ljubljana/Theologitis.ppt

Tognini Bonelli, E. 2000. *Corpus Linguistics at Work*. Amsterdam / Philadelphia: John Benjamins.

Yuste, E. 2000. Translation instruction in the Y2K - Electronic Corpora, Internet and Translation Instruction. In CD-ROM *Proceedings of the 7th Conference of the International Society for the Study of European Ideas (ISSEI)*. Section V, Workshop No. 501 "Teaching Translation at the Information Age". Bergen, Norway: University of Bergen HIT Centre, 14-18 August 2000.

Yuste, E. 2001. Technology-aided translation training. In *Hieronymous* (3/2001) Bern: Switzerland: ASTTI (Swiss Association of Translators, Terminologists, and Interpreters).

Yuste, E. forthcoming. *Translation technology understanding and use in Switzerland*. Internal project report.

Ulrych, M. 1997. The impact of multilingual parallel concordancing on translation. In Lewandowska-Tomaszczyk, B. and P. J. Melia (eds) *PALC '97. Practical Applications in Language Corpora*, Lodz: Lodz University Press, 421-436.

Zanettin, F. 2000. Parallel Corpora in Translation Studies. In Olohan, M. *Intercultural faultlines*. Manchester: St. Jerome. 105-118.

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

Zanettin, F. forthcoming. DIY corpora: the WWW and the translator. In *Proceedings of the Conference on Training the Language Services Provider for the New Millenium.* Porto. May 2001. http://www.federicozanettin.net/DIYcorpora.htm

NOTE.

All online bibliographical references were last ckecked in April, 02.

Textual and terminological bridgeheads for traversing the language gap

Marita Kristiansen, Magnar Brekke

Norwegian School of Economics
and Business Administration
Department of Languages, Helleveien 30, N-5045 Bergen, NORWAY
Marita.Kristiansen@nhh.no Magnar.Brekke@nhh.no

Abstract

We describe here the basic modules of a concept-oriented bilingual text-and-term-based knowledge management system (KB-NHH) to which students, teachers, researchers, domain experts, terminologists, linguists, translators and writers of various categories can turn for content learning, reference and documentation. The aim is to ensure that the interface between English and Norwegian is being handled with efficiency and consistency.

Primary user context of the implementation described here is an on-campus e-learning system.The aim is to facilitate the representation, learning, teaching and dissemination of relevant domain knowledge, to monitor changes in and development of the subdomain languages and to document all through authentic citations. Conceptual linkage of terms and authentic segments in the text bank allow source inspection and evaluation by user. Focus is on corpus-based term extraction, definitions, terminological representations, Norwegian-English equivalence problems and contrastive phraseology.

This paper makes a distinct contribution by proposing the integration of a conceptual knowledge-base with the textual manifestation of its underlying domain knowledge and its terminological representation in one or more languages, all in the context of a standard e-learning system. This should greatly facilitate learning by bridging the language gap experienced by native and non-native students alike in approaching a new knowledge domain.

1. The general problem

Communication in very specific domains of activity is crucially dependent on possession of specific domain knowledge and mastery of the specific domain language through which such knowledge is conventionally represented and transmitted. Whereas translation of general text between two national languages remains a general challenge for both human and machine translation, the translation of special domain text presupposes far greater proficiency in handling the content and representation of that domain knowledge.

Thus translation work undertaken along the interface between two special domain languages, each of which being entrenched in its respective national language, puts heavy demands on the translator's ability to control content and expression on both sides of the gap. Similarly a student of a specific domain faced with teaching or textbooks in a foreign language will have a dual problem: He or she will need to connect the technical terms and concepts encountered on the far side to equivalent concepts and terms on the near side, which in principle involves learning new content also in the native language. The need for a content and terminology management system at this point should be obvious, while the practical solution is not.

2. Specific obstacles

The potential problems arising in the contrastive language situation just described can be further aggravated if there are marked asymmetries between the two languages involved. In the domains referred to above English tends to be the source language for the overwhelming majority of communication involving bilingual text and terminology, and the pace at which new concepts and terms are created and disseminated can be quite hectic. This puts under considerable pressure a number of "lesser spoken languages", and particularly their cultural resilience and readiness for "terminological self defense". This makes it all the more important to compensate for the asymmetry by providing efficient and user friendly tools for managing the relevant language resources.

Fortunately the rapid development of information technology has placed tools within our reach which may enable even a lesser-spoken language such as Norwegian with 4.5m speakers to cope, partly at least, with such major challenges. We will describe here the basic modules of a concept-oriented text-and-term-based knowledge management system (KB-NHH) to which students, domain experts, terminologists, linguists, translators and writers of various categories can turn for content learning, reference and documentation. The aim is to ensure that the interface with English is being handled with efficiency and consistency.

The project described here is being developed in the context of the TERMINEC project, a three-year effort to establish the foundations of such a resource database for Norwegian and English special language as used in economic-administrative domains. What follows below is a description of a particular implementation of tools for bilingual data capture, terminology handling and application in a research and teaching environment dependent on economic-administrative communication. Due to space limitations the focus will be on modules involved in a web-based e-learning system.

3. The building blocks[1]

3.1. Data capture.

The foundations of the TERMINEC database are being implemented in the form of two parallel text corpora, one English, one Norwegian, of representative text from about

[1] "Modules" referred to in this section are shown in appended diagram

30 economic-administrative subdomains (see table 1), and a parallel term database whose contents are largely being derived from and dynamically linked to the text corpora.

One of the modules is thus a textbank (module 4 in appended diagram), a representative corpus of indexed full texts in the chief genres associated with didactic, expository and popularizing text types drawn from textual representations of the universe of subdomain knowledge (module 1).

Accounting and Costing, Capital markets, Corporate analysis, Corporate strategy and Ethics, Economic geography, Economic history, Economics, Economy systems and management, Finance, Investment, Information systems and management, Law (Corporate law, EU/EEA law, Tax law, etc), Macroeconomics, Management, Market communication, Market economics, Market research, Marketing, Mathematics and statistics, Microeconomics, Organization and management, Organizational behavior, Public economy, Quality management, etc.

Table 1: Tentative economic-administrative domains/subdomains

Typologically the text bank will contain English and Norwegian parallel texts both in the sense that they are original texts which share subdomain and genre, as well as in the sense of aligned translation pairs of source text and target text, which will increase the research value of the collections considerably.

3.2. Knowledge representation.

Terminological research is normally based on the onomasiological principle, the grouping of terms according to their conceptual meaning. Thus any knowledge subdomain can be characterized by a (partially) structured set of basic concepts which are represented linguistically through domain-focal terms (cf. Brekke, 2000). Establishing or extending conceptual systems (cf. module 8) becomes essential in achieving authentic representations of the knowledge which constitutes a given subdomain. This activity presupposes close cooperation between a domain expert and a trained terminologist (cf. "module" 2 & 3) in identifying and delimiting what the basic concepts are, conventional term usage, acceptable synonymy etc. The repository for their work is a termbank (cf. module 9) holding terminological units defined, classified as to subdomains, and mapped to their respective key concepts and conceptual hierarchies in module 8. Using the concept as a term record pivot (as is done in e.g. Trados MultiTerm, which is employed in the pilot project) facilitates the inclusion of other language equivalents (French, German and Spanish are obvious candidates for inclusion later on).

3.3. Term extraction (cf. module 5).

The slow time-honored techniques of "excerption" has long since been supplemented by increasingly sophisticated computational methods. Many of the results are impressive but have not allowed us to dispense entirely with the services of the domain expert in tandem with the terminologist. Given that the selection of input texts has yielded a representative corpus, it remains a sampling and thus very far from being exhaustive of the knowledge constituting a given subdomain. The problem is twofold: On the one hand, any automatically generated list of term candidates (cf. module 6) will reflect massive overgeneration of spurious combinations which will need to be pruned. On the other, no automatic term extractor will point out which basic terms are NOT represented in the sampled text, which takes an alert and knowledgeable human being.

The TERMINEC pilot project has allowed room for experimentation along these lines using SystemQuirk's suite of terminology tools. The point of departure is frequency lists followed up by selective concordance work. A typical subcorpus (of about 17000 words) yields the following (table 2):

90 internet	41 countries	31 investment
81 america	40 prices	30 companies
73 new	39 funds	30 capital
71 economy	37 years	30 fund
63 growth	37 economic	28 markets
60 productivity	32 technology	28 japan
45 firms	32 high	26 big
44 business	32 year	26 commerce
44 investors	32 risk	26 shares
41 market	31 share	26 pension

Table 2: Top of System Quirk's standard frequency list.

Some of these one-word units of fairly general scope can be identified as Economics terms, which is useful but of limited value. SystemQuirk provides two different functions for enhancing frequency lists to improve on our term enquiry.

3.3.1. Weirdness.

SQ exploits a "weirdness"-function based on a comparative ratio which expresses the likely occurrence of a given item in the text being scrutinized compared to the same for a large general corpus. Where the latter occurrence is zero the ratio will of course be infinite, indicating either a typo, a nonce word, or in fact a technical term, which is also indicated by a very high ratio. As a result a number of items occurring only once in a given text will be brought to the top of the frequency list, and such lists usually give significant inputs to the ensuing frequency studies. Table 3 (over) reveals a typical situation. It should be noted in table 3 that of the top 30 items on the list, 2/3 of them occur only once, which would effectively drown them out of the investigator's attention had not the "weirdness"-function been active (cp table 2).

While both tables contain terms which are immediately recognizable by an economist they only share one (*investment*), and those on the "Weirdness"-list are clearly of a more specific domain-related scope (and presumably less recognizable by a nonexpert). Table 3 has

12 inf!-terms, i.e. items not occurring in a large corpus of general English, while the remainder occur between 151 and 3 times more often than they would in that corpus. Thus their degree of specialization is approaching general usage.

3.3.2. Terms as strings of content words.

The other tool offered by SQ for sniffing out potential multi-word terms, aptly named Ferret, is based on a very simple algorithm: It takes a general list of function words as boundary signals and proceeds to identify any string of content words uninterrupted by such boundary signals as a term candidate. Table 4 displays the results obtained from examining the same text as above.

Frq	Match	SL/GL Ratio
10	capital markets	inf!
3	business cycle	inf!
2	annual report	inf!
2	central bank	inf!
1	new york stock exchange	inf!
1	dow jones industrial average	inf!
1	cost of capital	inf!
1	capital stock	inf!
1	european union	inf!
1	institutional investor	inf!
1	balance sheet	inf!
1	fiscal policy	inf!
1	solvency	151.2382
6	equity	88.5297
1	annuity	75.6191
1	takeover	75.6191
1	futures	50.4127
31	investment	35.1849
2	premium	32.7001
2	inventory	31.8396
1	liquidity	30.2476
1	diversification	27.4978
1	downstream	19.5146
1	revenues	10.2534
3	yield	8.0303
4	bond	7.8565
1	float	6.8745
1	commodity	5.5500
1	options	4.4482
1	margin	3.2700

Table 3: Top of System Quirk's frequency list with "weirdness"-function active.

For reasons which are unclear Ferret missed two of the occurrences of *capital markets*, and it does seem to invite some obvious refinements of its list of boundary signals, but otherwise the high end of the frequency list does throw up some promising term candidates.

3.3.3. Equivalence checking: Plugging the terminological holes.

In economic domains the terminological pressure from English has increased in proportion to the rapid globalization processes seen through the nineties and continuing unabated, while the readiness to invest in professional means for handling the textual interface has been lacking. Most of the recent efforts have gone into developing a speech interface, and the systematic monitoring and creation of suitable terminology for use in translating economic texts has been left to private initiative. Some subdomains thus appear well looked after, while others tend to end up with haphazard and ad hoc equivalents for newly formed concepts and terms from English-speaking

8 capital markets	7 pension funds	7 mutual funds
5 see chart	5 less than	5 life insurers
5 past decade	5 information technology	4 institutional investors
4 share prices	4 s economy	3 this year
3 on average	3 recent years	3 since america
3 point out	3 but there	3 this survey
3 other countries	3 retail sales	3 but even
3 cost savings	3 poorest countries	3 world bank
3 foreign aid	3 emerging economies	3 supply chain
3 short term	3 b2b e	3 s gdp
3 hedge funds	3 state street	3 an annual average

Table 4: Ferreted strings

cultures. Since the two languages have very close historical and lexical affinities, one should not be surprised to encounter a variety of terminological misfits, from simple (and humorous) "folk translations" through cognate shifts to serious "false friends" which may create hazardous and expensive mistakes.

Cognates constitute a rich quarry for terminological misfits. Consider the following examples:

*1. Federal Reserve Bank of Minneapolis President Gary Stern warned on Friday against the ``**moral hazard**'' that may prompt banks to undertake too much risk amid excessive confidence of government safety nets.*

Anyone connected professionally with hedging and insurance will recognize the special term (in bold). While each member of the phrase has a cognate with several

meanings in Norwegian, it is rather obvious that the connotations they bring along are quite different from the English ones. Nevertheless the temptation to use the "direct method" is clearly irresistible, as the following sample (from a sizable collection) will show:

*2. Kombinasjonen av usikrede lokale banker, **moralsk hasard** i utlandet, av kortsiktige utenlandske kapitalplasseringer og Pengefondets innstrammingspolitikk, ga kraftige negative utslag.*

A linguistically sensitive person familiar with the concept underlying the original expression in 1 (including their use as separate English words) will realize that the "calque" in 2 creates undesirable connotations. Unfortunately many will fail to see the problem, which allows the emergence of a Norwenglish (quasi-Norwegian) terminology lacking professional and cultural quality assurance. Arriving at the Norwegian equivalent "**åtferdsrisiko**" requires professional handling, time, and relevant domain knowledge (another subdomain prefers "**subjektiv risiko**"). It also requires an efficient dissemination channel to ensure its adoption and use.

Equivalence checking is thus serious and important business for anyone purporting to traverse the knowledge gap as well as the language gap through translation or related forms of text production. It appears to be one stage of the bridge building which cannot easily dispense with the bilingual human expert/terminologist or their term creation principles, be they linguistically, politically or culturally motivated. In other words, the bridge heads on either side must be anchored in their respective professional context, and the quality of work assured through a content and terminology management system. Only then can our efficient computer-based tools for processing and dissemination come into their own.

4. Dissemination and use.

At the outset the material held in the KB-NHH database will form the basis for student oriented bilingual domain glossaries with definitions, as well as genre-related material for learning and teaching. Both textbank and termbank will be SGML conformant, adhering as far as possible to the TEI guidelines, which allows interactive access via a Web-browser or ftp downloading. In addition all or parts of the termbank may be distributed on CD-ROM. Printed versions are possible, but the main emphasis will be on interactive use via electronic networks. This will take full advantage of the dynamic aspects of electronic media, allowing e.g. fuzzy matching of any search to the nearest form.

The diagram referred to in Appendix outlines the current architecture of KN-NHH, a "proof-of-concept" implementation of the e-learning oriented application of TERMINEC. The student enters the e-learning system (cf. module 11, a "Blackboard"-type system) via a standard web-browser (cf. module 10), accesses the course catalog and proceeds to the description/presentation of the course content in either English or Norwegian. All domain focal terms have active links to the central conceptual system. At this point the student may follow the link to the relevant term record in the desired source language, study definitions etc. and go from there into the text bank to inspect authentic text samples illustrating usage, phraseology etc. This is particularly useful for a non-native student. Alternatively the student may proceed directly from conceptual system to the text samples, and from there via clickable text-embedded terms across to the full term-bank representation of the desired concepts to study definitions, synonyms, acronyms etc.

Students approaching a new knowledge universe will easily detect concepts not adequately covered or explained. All searches will be logged to allow a study of user behavior and user needs, with a view to enhancing the intuitiveness of the user interface. Following an unsuccessful search the user will be asked (through automatic routines) to report unfound terms and submit a relevant text segment with source reference, and will have a chance to include responses or comments. It will be considered whether users also should be invited to join an «official» discussion group. Success in engaging the user in such dynamic interaction will not only provide a way of monitoring a continuous growth of the collection but may also create greater user identification with the KB-NHH, which in turn may have a standardizing effect.

5. Maintenance and development.

New concepts are constantly being created in the professional community and migrate towards general usage, sometimes even grabbing front page headlines: *unit-link*, *derivatives* and *hedge funds* have recently enjoyed such instant attention. At the time of writing *e-business* is very much in vogue (along with almost any noun with an *e-* prefix), and *creative accounting* is already a cliché in the financial headlines.

This implies that simply registering the constitutive concepts of a given domain, including their manifestation through the terminology of one or more national languages, is not done once and for all. What is required is a more or less continual monitoring of the entire life cycle of any given term, from creation through extension and expansion to disuse and eventual death. The above are random examples of an ongoing process which is in fact quite normal, although the speed and intensity may vary with the times and the subdomain. Ideally the new or altered terms would need to be absorbed by writers, their underlying concepts defined and systematized by domain experts and terminologists, standardized by professional bodies, and their usage documented through carefully vetted citations. At the receiving end of this process would be speakers of other languages (be they experts, journalists or textbook authors) who would ideally have to establish procedures for finding or creating equivalent terms and determine proper usage.

6. Outlook

This paper makes a distinct contribution by proposing the integration of a conceptual knowledge-base with the textual manifestation of its underlying domain knowledge and its terminological representation in one or more languages, all in the context of a standard e-learning system. This should greatly facilitate learning by bridging the language gap experienced by native and non-native

students alike in approaching a new knowledge domain. A well documented and web-accessible clearinghouse for English-Norwegian economics text and terminology as envisaged here would also establish a significant point of reference for empirically based term-formation and possibly standardization, thus providing Norwegian export-oriented corporations with a much needed quality assurance of the linguistic interface. The same would hold for Norway's administrative and political cooperation with the outside world, as well as for the global language industry, which depends on the availability of multilingual databases and some form of translation. The realism in trying to stem the flood of English usage in conducting the professional affairs of people whose normal mode of communication is something other than English is highly debatable, but the virtue of avoiding linguistic domain losses in Norwegian is not.

7. References

Ahmad, K. & M. Rogers (1994). "Computerised terminology for translators: the role of text", in Brekke, Andersen, Dahl & Myking (eds) *Applications and implications of currenct LSP research*. Bergen: Fagbokforlaget.

Brekke, M, J. Myking og K. Ahmad (1996). "Terminology Management and Lesser-Used Living Languages: A Critique of the Corpus-Based Approach", in *Proceedings of 4th International Congress on Terminology and Knowledge Engineering* (TKE '96). Vienna: Indeks Verlag.

Brekke, M. (1998) "When «Empiry» strikes back: A Corporal Confrontation". *Proceedings from Workshop on Adapting Lexical and Corpus Resources*, First International Conference on Language Resources and Evaluation, Granada, Spain.

Brekke, M. (1999). "TERMINEC: The dual linkage of text and terminology", in *Proceedings of 5th International Congress on Terminology and Knowledge Engineering* (TKE '99). Vienna: Indeks Verlag.

Brekke, M. (2000). "On the Lexical Identification of Domain Focal Text and Terminology". *Proceedings of COMLEX 2000*, University of Patras, Greece.

Brekke, M. (forthcoming) "TERMINEC. A Clearinghouse for Economics Text and Terminology", to be published in Proceedings of ICAME 2000, Sydney, Australia, April 2000.

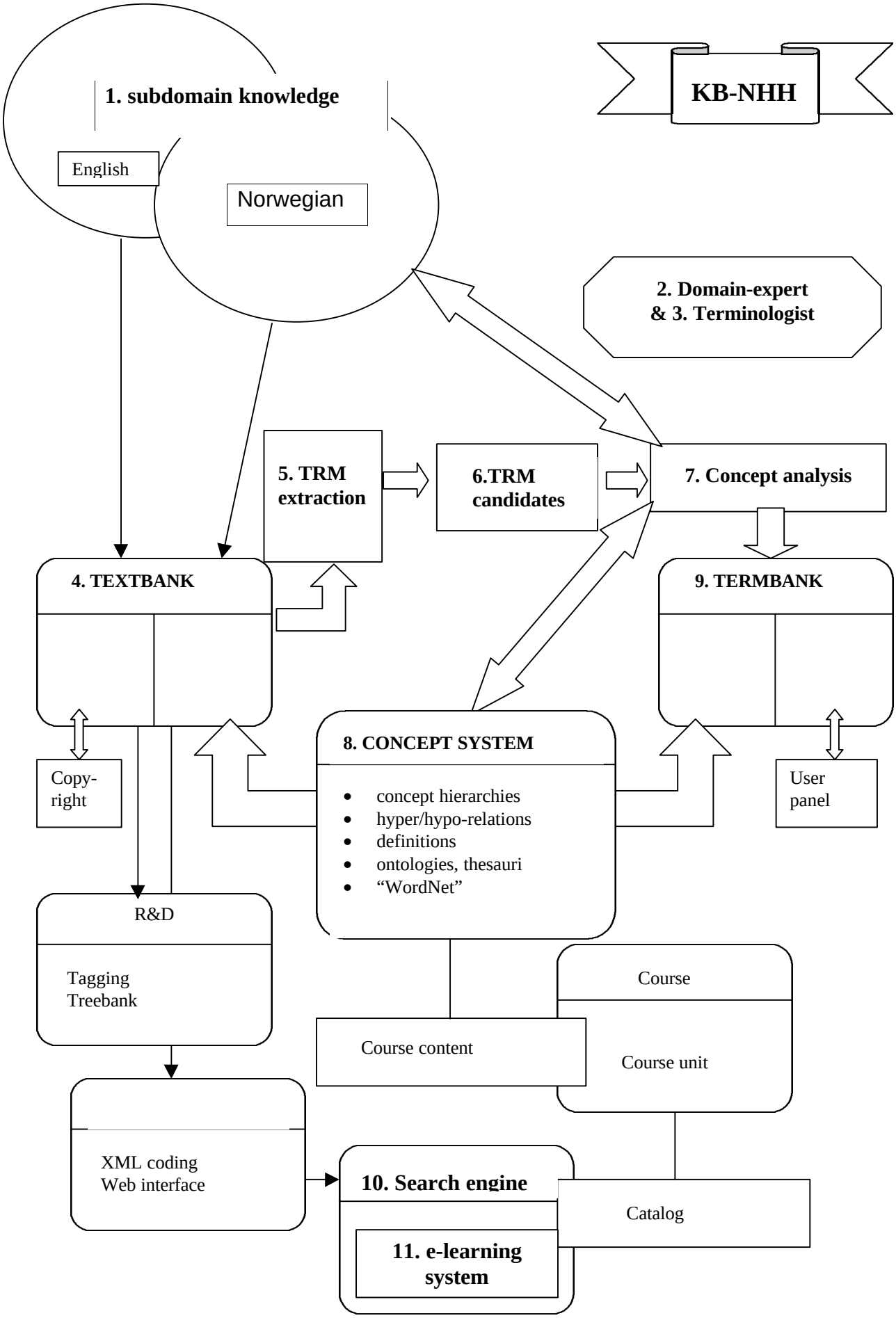
Church, K. et al. (1991). "Using statistics in lexical analysis". In Zernik, U. (ed.) *Lexical Acquistion: Exploiting on-line resources to build a lexicon.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lebart, L., A. Salem and L. Berry (1998). *Exploring textual data*. Dordrecht: Kluwer Academic Publishers.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: EUP.

SystemQuirk: http://www.mcs.surrey.ac.uk/Research/CS/AI/SystemQ

Appendix: **Outline of the KB-NHH architecture** (see next page)


1. subdomain knowledge
English
Norwegian
KB-NHH
2. Domain-expert
& 3. Terminologist
5. TRM extraction
6.TRM candidates
7. Concept analysis
4. TEXTBANK
9. TERMBANK
8. CONCEPT SYSTEM
concept hierarchies
hyper/hypo-relations
definitions
ontologies, thesauri
"WordNet"
Copy-right
User panel
R&D
Tagging
Treebank
Course
Course content
Course unit
XML coding
Web interface
10. Search engine
Catalog
11. e-learning system

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

Natalie Kübler

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

Abstract

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

1. Introduction

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

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

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

2. Resources

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

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

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

2.1. Bilingual glossaries

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

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

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

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

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

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

2.2. Corpora

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

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

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

2.3. The Web

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

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

2.4. Tools

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

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

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

3. Systranet: customisable dictionaries

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

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

The translation results can be improved by modifying the dictionary, a recurrent process which can be continued so long as the modifications produce significant improvement. Systranet offers specific features that allow

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

the user to see which terms have been translated using customised dictionaries, and which terms are not recognised at all. It allows the user to check whether the dictionary entries have really improved the translation results as expected. Another feature used to complete the dictionary is the non-word feature: all the words that have not been recognised by Systran or the user's dictionaries appear in red. They can then be integrated into the user's dictionary.

4. Experiment and methodology

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

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

4.1. Creating the dictionaries

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

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

4.1.1. Step-one dictionaries

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

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

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

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

Figure 1. The noun *README* in context

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

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

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

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

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

4.1.2. Step-two dictionaries

Once a set of dictionaries has been produced for each HOWTO, it must be tested not only to correct possible

[13] http://www.traduc.org

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

[15] http://www.termium.com

errors in the entries, but also to add the new words that are neither in Systran's nor in the customised dictionaries. The more HOWTOs are translated, the fewer words have to be added until the dictionaries are saturated, i.e. no new word can be added to improve translation results.

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

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

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

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

Source Text	This page contains a simple ***cookbook*** for ***setting up Red Hat 6.X*** as an ***internet gateway*** for a ***home network*** or small ***office network.***
Step-one dict.	Cette page contient un **cookbook** simple pour l'établissement ***Red Hat 6.***X en tant que ***passerelle Internet*** pour un ***réseau domestique*** ou un petit ***réseau de bureau***
Step-two dict.	Cette page contient ***des recettes*** simples pour ***l'installation Red Hat 6.***X en tant **que** ***passerelle Internet*** pour un ***réseau domestique*** ou un petit ***réseau de bureau***.

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

4.2. Translation outcome

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

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

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

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

Fig. 5: Conjunction and disjunction in an NP context

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

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

Fig. 6: An example of transfer rule overgeneralisation

4.3. Human vs machine?

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

5. Conclusion

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

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

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

6. References

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

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

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

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

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

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

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

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

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

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

Yuste Rodrigo E. 2001. Making MT Commonplace in Translation Training Curricula –Too Many Misconceptions, So much Potential. In *Proceedings of the MT Summit VII*, Santiago de Compostela, 18-22 September 2001.

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

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

Evaluating Translation Memory Systems

Angelika Zerfass

Freelance Translation Tools Consultant
Holzemer Str. 38
53343 Wachtberg
Germany
azerfass@debitel.net

Abstract

Since the mid 1980s, translation tools have taken over more and more of the daily lives of translators and translation project managers. But a lot of time now has to be spent on evaluation, training and administrative tasks.
Translation tools were designed to make the translator's work easier, faster and more efficient. They range from conversion utilities to terminology management, translation memories, machine translation as well as workflow and project management systems.
They were developed with the aim to reduce repetitive translation work, but on the other hand they add different tasks to the workload, like administrating databases and the like.
This presentation will give an overview of one area of translation tools - the different translation memory systems on the market today and the technologies they use. It includes a comparison of common basic features like word count, analysis/statistics function and pre-translation, some tools' specialities as well as the description of data exchange possibilities between the systems by use of the TMX format.
As there is no "one best tool" for everything, the aim of this workshop is not, to recommend one tool, but to provide some guidelines for evaluating translation memory systems according to individual requirements.

1. Translation Memory Tools - Overview

Translation memory systems, as the name implicates, "memorise" the translations made by a human translator. Most translation memory systems (often also called "TM-systems"), consist of a database that stores the original text along with its translation - a database of segment pairs.

"Segment" here indicates that the units that is being translated and stored to the database can range from a single word (for example a heading or an item in a bulleted list) to phrases, complete sentences or even whole paragraphs. The tools recognise a segment by a set of internal rules that define, for example, that a segment ends with a full stop or a paragraph mark.

During translation itself, the tool will automatically look up every new source language segment to be translated in that bilingual translation memory. If the same segment is found in the database, the system will offer the translation that was saved with this segment as a suggestion to the translator for reuse. If it does not find the very same segment, it will start looking for similar segments. These are the so-called "fuzzy" matches, as the source language segments (in the document and in the database) only match to a certain percentage. When the translator gets such a fuzzy match from the database, they can decide if and how much of it can be reused for the current translation. Usually the translator can even set the level of "fuzziness", that is the percentage of similarity, so that the system will only offer translations that can be reused without having to make too many changes to the suggested translation.

Thus the use of a translation memory system can increase consistency and it cuts the time for writing a translation. This is especially true for the translation of repetitive documents like technical documentation, manuals, instructions and updates of already translated material.

Translation memory tools are usually the main component of a tools' suite. These suits also offer recycling tools, so called alignment systems. These are used to prepare translations made without translation memory systems for reuse in such a translation memory tool. They read in the source and the target language files, display them in parallel and propose connections of the source language segments to the corresponding target language segments. A translator will then review these connections. Then, the segment pairs can be imported into the translation memory. From now on they can be used just as if they had been translated interactively with the system itself. Another component of such a tools' suite is the terminology management system - another database that stores single terms (or phrases) together with their translation(s) into the target language(s). The translation memory database and the terminology database work together during translation. The translator will not only get suggestions for the translation of whole segments but also a list of all the terms within that segment that were found in the terminology database. Other components of such a tools' suite could be workflow or project management systems as well as filters and utilities for file format conversion.

Translation memory systems also start to be customisable for use with document or content management systems and some are even programmable via an API (application programming interface - programming commands that enable the user to call the translation memory system from other applications).

2. Translation Memory Tools Basic Principle

Basically all translation memory tools were developed with the same goal in mind: Something that has been translated before should not have to be translated again from scratch. It should come out of the database or reference material so that the translator only has to decide whether the previous translation can be reused or needs to be modified.

The technologies used to achieve this are different. Some tools use a model of referencing the files of a previous project, The referencing model uses those previously translated files (original source language files and translated files) as the source for suggestions of new translations. This model works especially well for projects with many updates containing a lot of small changes.

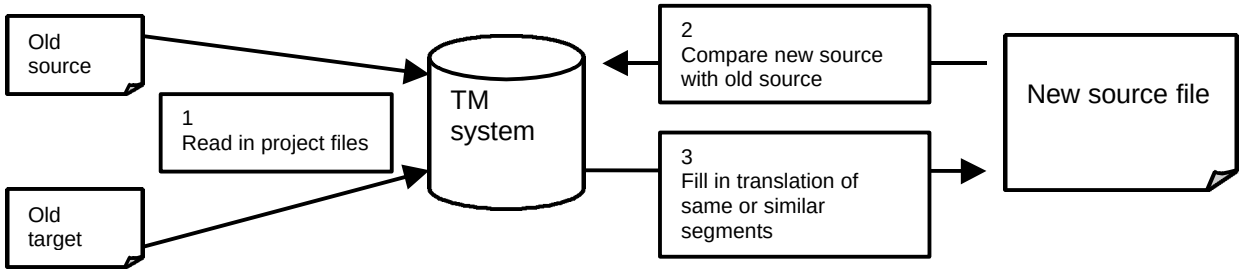


Figure 1. Reference Model

The database model on the other hand stores all translations ever made in one database, independent of context, which is useful if the same or similar segments appear in different projects and document types. Most of the commonly used translation memory systems are able to work with any language installed on the user's machine and they usually also allow the user to add project or user specific information to each translation.

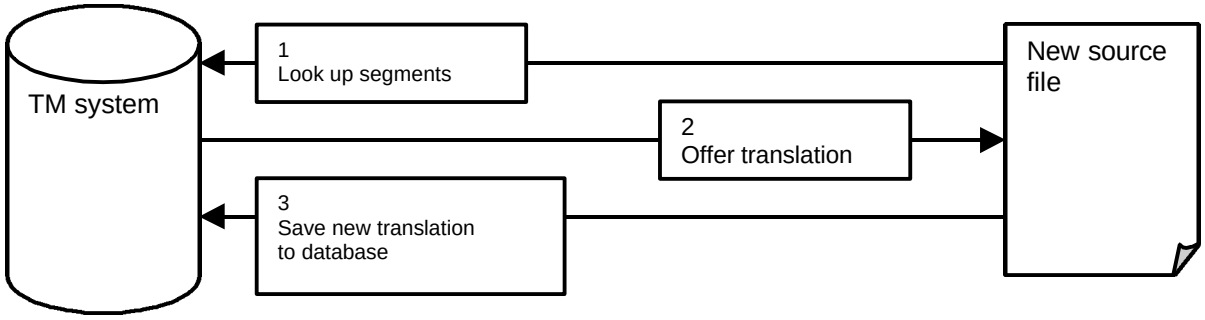


Figure 2. Database Model

3. Translating with Translation Memory Tools

The text to be translated consists of smaller units like headings, sentences, list items, index entries and so on. These text components are called "segments". Translation memory systems are equipped with a set of rules, which enables them to recognise, where a segment starts and where it ends. When translating with a translation memory system, the system goes through the text segment by segment, offering each of them to the translator together with any translation for this or a similar segment that has been stored in the database or can be found in the reference material. The translator decides whether to reuse the proposed translation, to adapt it or to create a new translation and then saves it to the system. Thus the translator builds up a store of segment pairs that can be referenced for future translation.

This store of segment pairs can also be used for analysing new files to determine the rate of recycling that can be achieved. Or it can be used to run a pre-translation, which creates files that contain segments with more or less matching translations already in them. This is very useful when working on a large batch of files or preparing files for other translators who are not working with a translation memory tool.

To be able to use translation memory tools on different file formats, from common Word files to DTP (desktop publishing) files, for example FrameMaker or Interleaf or files for the web in HTML, XML or SGML, some of these formats need to be converted to a format that the translation memory tool can work with. This happens either by use of separate conversion tools or filters integrated into the translation memory systems. Selecting a TM system therefore also depends on what file formats have to be worked on and how much time and effort needs to be spent on preparing and converting them to a usable format for translation and back to the original format afterwards.

Also, when it comes to software localisation for example, different tools have to be used for different parts of the project. The project might consist of text within the software from the user interface (GUI) to dialogs and messages as well as online-help files, documentation, packaging and marketing material and so on. And here different types of text require the use of different tools. GUI, software dialogs and messages are best translated with a software localisation tool, that is a translation memory tool that can read those special software file formats. They usually also contain testing features to check for consistent use of hot keys for example, or length related problems that might arise, if the translated text does not fit the button space it is supposed to appear on. But those systems are mostly specialised on the software itself.

For translation of the documentation, another translation memory tool is needed. And here the question arises how those tools for translating software and documentation interact, because what has been translated for one part might also be reusable in the other (this will be covered in the section about data exchange further down).

Online-Help files for example, could be translated with either a software localisation tool or with a translation memory system for documentation as both system types support this format.

4. Feature Comparison

All translation memory tools offer basic functionalities like word count or an analysis of recycling potential (how many of the segments in the file to be translated are present in the database or reference material as 100% matches or as similar, fuzzy matches). They also provide features for automatic pre-translation, search functionalities within the segment database, as well as access to terminology management components during translation. But each and every tool also has its specialities. These are the features that can influence the choice of tools.

Most translation memory systems read the files to be translated into the system itself, converting them into a table where one column contains the source language segments and another column that will be filled with the respective translation. Others connect to Microsoft Word so that any file that can be opened in Word does not have to be converted before translation and can be worked on in a WYSIWYG (what you see is what you get) mode. The translators can work in an environment that they are used to. Other file formats, for example DTP formats or so called tagged file formats like XML, HTML or SGML, are either converted or displayed in a separate editor. Colours are used to mark text to be translated as well as tags that make up the structure and formatting of the file.

More and more developers are enhancing the functionalities of the translation memory tools by adding new features like context sensitive pre-translation or machine translation-like components (for segments that have no match from the translation memory) as well as project management components.

5. Data Exchange between Translation Memory Systems

For some time, translators did not have the possibility to bring the data from one translation memory system into another system for reuse. A situation that was alleviated to some extent by the tools manufacturers by adding export functionalities for some proprietary formats of other manufacturers. But it was not feasible for each tool to support all export/import formats of all other tools - especially with new tools being developed and marketed all the time.

Now, the tool manufacturers have agreed to use one standard format for representing the data in their systems or at least to offer this format as one of the export formats. This allows an easier transfer of translation memory data from one system to another - even though the results are not always completely satisfactory. This standard is called TMX - Translation Memory Exchange format. It is an XML based representation of the data stored in a translation memory system.

5.1. Example of data representation in TMX format:

Segment pair:

This is a test. (English segment)
Dies ist ein Test. (German segment)

TMX representation:

```
<tu>
    <tuv lang="EN_US">
    <seg>This is a test.</seg>
    </tuv>

    <tuv lang="DE_DE">
```

```
        <seg>Dies ist ein Test.</seg>
        </tuv>
</tu>
```

Each segment pair is represented with a <tu> and </tu> tag that denote beginning and end of the segment pair. ("tu" stands for "translation unit", as those segments pairs are often called.) Then come the individual languages of the segments and the textual contents. This format could be produced and read by any translation memory system that works with TMX.

There are three levels of TMX compliance today. The first level only represents the text itself. The second level is able to represent the formatting information as well. And level three would be used to represent additional tool specific data like user IDs, project names and everything else the user has specified. Today, most tools comply at least to TMX level 1 or even to level 2.

6. Conclusion

Before investing in any translation tool, it is necessary to list the individual user requirements. This includes the file types that are to be translated. As most translation memory tools rely on structural and formatting information in the file, to segment and display the text, it should be tested if the way the files for translation are constructed, work well with this or that translation memory system. It could even mean to adapt the way of writing the documents in the first place, so that, at the translation stage, the tools that are used can handle the files more easily.

Another point is the networkability and the list of supported languages as well as the different supported file types.

Pricing for licenses, training and support should also be taken into consideration.

Then the tools should be tested for some time with real life examples to be able to evaluate, which tools answer the user's requirements best. Most tool manufacturers offer a trial period of about 30 days or a limited demo version of the software or, in case a longer evaluation period is needed, an extended trial with the full version of the software. This usually includes the need to buy a training session as well, to prepare the people who will be evaluating the software in the best possible way.

7. References

Some download sites for demo versions of translation memory tools:

- Trados - Translator's Workbench
  www.trados.com/products/download.htm

- Atril - Déjà Vù
  www.atril.com

- SDL - SDLX
  www.sdlintl.com

- Cypresoft - TransSuite2000
  www.cypresoft.com
  (supports only European languages)

- Star - Transit
  no download, contact Star for a demo CD at www.star-group.net

- Champollion - Wordfast
  Freeware
  www.geocities.com/wordfast/cat.htm

Some download sites for demo versions of software localisation tools:

- Pass Engineering - Passolo
  www.passolo.com

- Alchemy - Catalyst
  www.alchemysoftware.ie/demo4/

More information on TMX:
www.lisa.org/tmx

Language Resources at the Languages Service of the United Nations Office at Geneva

Marie-Josée de Saint Robert

United Nations Office at Geneva
1211 Geneva 10
mjdesaintrobert@unog.ch

Abstract

The language staff at the United Nations makes a very selective use of language technologies. So far no computer-assisted translation software has been installed on translators" workstations even though tests have been conducted for several years on the two major computer-assisted translation (CAT) systems at United Nations Headquarters in New York, for instance. The aim of this paper is twofold : 1) to show why CAT systems are not considered as potential sources of improvement of quality nor quantity in translation work at the United Nations, and 2) to present the kind of language resources that are considered essential for the adequate rendering of content in any of the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish). This paper analyzes the particular linguistic and technical constraints specific to an international setting and argues in favour of a selected number of language resources used at the United Nations other than translation tools readily available on the market. Among such language resources, one finds search engines, government and research institutions" websites, and, in a not too distant future, institutional knowledge bases.

1. Introduction

In an international, multilingual environment such as the United Nations, surprisingly enough, translators and language staff in general are not considered on the same footing as substantive departments, which prepare reports and organize conferences. Wherever technological innovations are designed and developed, the primary concern is the diplomatic community or the international community at large, not the language staff. Although translators do have a major role to play in the preparation of parliamentary documentation, their needs, such as prompting automatic alignment of two language versions of the same document whenever desirable, are very seldom taken into consideration by United Nations designers and developers. This low profile for linguists may well explain why so few technological innovations have made their way through to the translator and the terminologist. More reasons can be found in the very nature of the translation process in multilateral diplomatic settings where linguistic and technical constraints play an important role.

2. Linguistic Constraints

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

2.1. Word Choice

Translation cannot be reduced to the mechanical substitution of one set of terms in one language by a similar set in another language.

2.1.1. Semantic Adequacy

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

(1) the report shows
(2) le rapport montre que
(3) il ressort du rapport que

Also, the correct rendering in French of the English phrase (4) is not (5) but (6):

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

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

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

It is not always clear with CAT whether faulty phrases such as (2) and (5) would be offered by the system, as it may only keep the first instance found and disregard other instances of the same phrases found subsequently, and whether the translator in haste may not accept the phrases in (2) and (5) since both look correct from the grammatical point of view but are incorrect from the semantic point of view[1]. Maybe more accurate information on what CAT systems do is needed. Yet it remains to be seen whether distributed management of translation memories can be efficiently organized on a large scale, with fifty translators having the right to update the translation memory on a permanent basis in each language pair.

2.1.2. Lexical Variety

Translations serve the purpose of a specific communication need and should not be considered as models for translators to replicate across the board. Such is also the case for terminology in any target language. Mere electronic bilingual dictionaries or glossaries cannot

[1]In (2) an inanimate noun is used with an animate verb; in (5) it is as though sexual practices would be divided into two categories: abusive and non-abusive, which is wrong in the case of very young girls.

satisfactorily capture variation, not only in the original language but also in the target language, if based upon the assumption that a notion corresponds to a term in English and one or several terms in French, for instance. Names given to human rights are a case in point. A terminologist would very happily collect the names of all rights, starting with the right to food, to adequate housing, and to education, while a translator would resent it. Such rights are indeed referred to under different names by different speakers, and a too rigid list of rights would miss the needed subtleties while discussions are still under way. Should "adequate housing" be rendered in French by "logement convenable," "logement adéquat," "logement suffisant," or "logement satisfaisant," all four equivalents being found in United Nations legal instruments or resolutions, and not by "bonnes conditions de logement" or "se loger convenablement" when the context allows or requires it? Translators want to preserve flexibility, when present-day translation systems propagate rigidity and, as a lurking consequence, poverty of style and vocabulary. For Fernando Peral (2002), a translator at the International Labour Organization: "The main operational problems of "semi-automatic" translation [i.e., translation with the help of translation memory systems] are linked to the quality of the output and to a process of "de-training" of the translator, who becomes less and less used to the mental process of searching for proper solutions in terms of functional equivalence and relies more and more on the machine"s decisions, which inevitably affects professional development and job satisfaction."

2.2. Linguistic Insecurity

Document originators at the United Nations are nationals from over a hundred and twenty countries. In most cases their native language is not one of the official languages of the Organization, and document drafters erroneously think they have to use English, which may prevent them from using their main language, even when it is an official language, and produce better originals. Documents may also be submitted to the United Nations by officials or experts working for Member States that do not have either any of the official languages of the Organization as their main language. Syntactic, semantic and morphological mistakes are therefore not rare in documents, and in most cases only translators are in a position to detect mistakes and rebuild faulty sentences in the original text. Only they are required to work in their native language that is one of the official languages. Due to lack of resources at the United Nations, only a small portion of all documents is edited prior to being translated (e.g., documents prepared by the Commission on Human Rights). Translators consequently do act as filters for grammatical correctness and language consistency as they work on the texts to be translated. As a result, they often improve original texts whenever the drafters or submitting officers accept their changes in the original documents. A translation memory processing straightforwardly a document to be translated prior to the perusal of a translator may not detect inappropriate use of terms or syntactic errors in the original language. Even when an automatic term-checking system is appended to the translation memory, it may not be as efficient as a human eye either. The fear therefore is that a computer-assisted translation system may add more mistakes to the original ones, which will then be even harder to detect and correct.

2.3. Different Stylistic Rules

Document drafters use a variety of writing rules and styles to convey meaning. For instance, among writing styles one can mention the fact that repetitious words are not considered as poor style in English but are definitely considered poor style in French. The English sentence (7) presents a repetition of the word "aircraft" which the French rendering in (8) would avoid:

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

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

2.4. Functional Adequacy

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

These constraints conflict with the concept of translation reuse for translation purposes on which most commercially available alignment tools and translation memory systems are based, especially when document traceability (i.e., the capacity of retrieving the complete document from which a sentence is extracted by the translation memory system) is not guaranteed.

3. Technical Constraints

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

3.1. Time Constraints

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

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

3.2. Digital Divides

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

3.3. Lack of Preparedness

CAT tools are known to be most efficient with repetitive texts. So far, since at the United Nations not all texts are available in electronic form, it is hard to assess the amount of repetition to be able to ascertain whether or not CAT is an efficient tool in this environment.

Proper training also has to be given to translators to make certain they know how to utilize the tools that they are given. The fear is that translators are no longer assessed only for their linguistic and narrative competence and performance, but by their computer skills.

Finally, equipment used in an international organization has to be compatible with the equipment required by a particular CAT software.

4. Tools for Translators

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

4.1. In-house Glossaries

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

"Significant - Accusé, appréciable, assez grave/long, caractéristique, certain, considérable, de conséquence, d'envergure, de grande/quelque envergure, digne d'intérêt, d'importance, de poids, de premier plan, distinctif, efficace, élevé, éloquent, explicatif, expressif, grand, important, indicatif, instructif, intéressant, large, louable, lourd de sens, manifeste, marquant, marqué, net, non négligeable, notable, palpable, parlant, particulier, pas indifférent, perceptible, plus que symbolique, positif, pour beaucoup, probant, qui compte, qui influe sur, réel, remarquable, représentatif, révélateur, sensible, sérieux, soutenu, significatif, spécial, substantiel, suffisant, symptomatique, tangible, valable, vaste, véritable, vraiment; a significant proportion: une bonne part; in any significant manner: un tant soit peu; not significant: guère; the developments that may be significant for: les événements qui peuvent présenter un intérêt pour; to be significant: ne pas être le fait du hasard." [2]

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

[2] Organisation des Nations Unies (2000).

context in which the document has been drafted, rather than any previous context.

4.2. Web resources

Language resources used by translators also include online dictionaries and government and research institutions" websites that translators have learned to identify and query for information extraction and data mining. Portals have been designed to help translators locate best language and document sources on the Internet.

4.3. Alignment Tools

Additional tools are document alignment tools by language pairs. Indexing of large text corpora for retrieval of precedents are felt preferable to tools that provide text segments, be they paragraphs, sentences or sub-units with their respective translations, but without any indication of date, source, context, originator, name of translator and reviser to assess adequacy and reliability in an environment where many translators are involved.

4.4. Knowledge Base

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

Such knowledge base ideally would reduce the choices offered to the translator rather than list all possibilities. The easier it is for the translator to make the decisions he or she needs the faster he or she delivers.

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

5. Conclusion

In conclusion, United Nations translators are very cognizant of the limitations of automated tools for translation and are more inclined to rely on easily accessible, structured information concerning the history and main issues in a particular subject matter in order to be completely free to choose the best translation equivalents.

6. References

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

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

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

Global Content Management – Challenges and Opportunities for Creating and Using Digital Translation Resources

Gerhard Budin

University of Vienna
Department of Translation and Interpretation
Gymnasiumstraße 50, A-1190 Vienna
gerhard.budin@univie.ac.at

Abstract

In this paper the concepts of content management and cross-cultural communication are combined under the perspective of translation resources. Global content management becomes an integrative paradigm in which specialised translation is taking place.

1. Convergence of content management and cross-cultural communication

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

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

2. Reflections on concurrent trends

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

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

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

3. What is Content?

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

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

data + interpretation = information + cognitive appropriation = knowledge + collective representation and utilization = content

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

Looking at the generic concept behind the word content, we would say: *Content* is what is *contained* in a written document or an electronic medium (or other

containers of such types). We would expect, that any content has been created by humans with certain intentions, with goals or interests in their minds. So we can say that content is usually created for specific *purposes* (such as information, instruction, education, entertainment, arts, etc.).

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

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

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

- *Cognitive appropriation*: knowledge is always the result of cognitive operations, of thinking processes. Yet knowledge is not limited to the personal, individual, subjective level. When people consciously share knowledge on the basis of directed communication processes, it is still knowledge, either referred to as collective or shared knowledge, or as interpersonal, intersubjective, or objective knowledge. In theories of scientific knowledge, the term 'objective knowledge' was mainly explicated by Karl Popper (1972) and is the result of regulated research processes such as hypothesis testing, verification, proof, etc., and that is written down in science communication processes. This is the justification for libraries to talk about their knowledge repositories in the form of books that contain this type of knowledge, i.e. objective knowledge – but as mentioned above, this knowledge has once been subjective knowledge in some persons, in this case scientists, that had thought and communicated about it before.
- *Complexity*: the level of complexity is another factor in the transition from information to knowledge. The same processes as on the previous emergence level, from data to information, are relevant: condensation of information (summarising), analysis and interpretation of information gathered, contextualisation (relating information to concrete problem solving situations, embedding and situating information in historical contexts and drawing conclusions from that, correcting (revision of data collections on the basis of experience) and categorising knowledge accordingly.
- *Life span:* the validity of knowledge has to be checked all the time. Again we are reminded by Karl Popper that all knowledge is unavoidably hypothetical in nature and that no knowledge is certain for eternity. Therefore we constantly have to redefine the criteria by which we evaluate our current knowledge for its validity. Another metaphor from nuclear physics is used for knowledge, especially in scientometrics: the 'half life'of knowledge is constantly decreasing, due to the increase in knowledge *dynamics*, not only in science and technology, also in industry, commerce and trade, even in culture, the arts, government and public sectors, the social sector, etc.

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

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

The focus and the real goal of knowledge management is actually on *content*, i.e. not on the formal aspects of computing, but on what is behind the strings and codes: the concepts and the messages. When knowledge is then packaged as a product for a certain audience, presented in certain media presentation forms, then we can speak about *content*, which also has to be managed in specific repositories and to be processed for publishing purposes, for instance.

As soon as we introduce another dimension, that of culture and cultures, communicating content across cultural boundaries becomes a crucial issue. Since we talk about *localization* as the process of culturally adapting any product to a market belonging to another culture than that of the original market of a product, content also needs to be localized when it should be presented to other cultures. Translation, as a part of the complex process of localization, is one crucial step in this process, but not the only one. Content localization may very well involve more than translation in the traditional sense, i.e. we might have to re-create part of that content for another culture, or at least change fundamentally the way this content is presented to a certain culture.

Since 'content' is a relational concept, we have to ask ourselves, what contains something, i.e. what is the container, and what is in this container. A book (with its table of *contents*), for instance, is such a container, or a database with the information entered in the records as the content. A text or a term can also be containers, with the semantics of sentences and the meaning of the term as the

content. But this distinction between container and content cannot be made in a very clear-cut way. We are faced with a semiotic dilemma. Form and content always interact. The medium we choose to present certain information will have some impact on this information, the structure of the information will also lead us in the choice of an adequate medium. Usually we cannot completely separate the container from the content, the form from the content, the term from the concept, the semantics from the text, the medium from the message, etc. Despite the heuristic validity and necessity of an analytical separation, we need a synthesis in the sense of a dynamic interaction, an interactive complementarity. At the same time we also might want to transform one form of knowledge representation into another one, for certain purposes and tasks, and then have to be sure that the content of each knowledge representation does not change – a difficult task.

Similar to typologies of data, information, and knowledge, we also need a content typology. There are different criteria for distinguishing types of content:

the domain where specific content is created in: any field of scientific knowledge, a business branch, a profession, a form of art, a type of social activity, etc. For this type of distinction, we may also differentiate different degrees of specialisation (highly technical and scientific, monodisciplinary or multidisciplinary, popularised, etc., depending on the audience targeted);

the form of representation: text, picture, personal action, etc. or the medial manifestation: web site content, the 'story' of a film, of a video, a piece of music recorded, a digitized scroll, etc.

Here we see again that the form of representing content and the medium chosen to do this is constitutive for distinguishing types of content.

First of all, the purpose of the content: instruction, education, research, aesthetic and artistic purposes, etc. Secondly, the kind of content product that is designed for a particular target audience (e.g. a multimedia CD-ROM for 6-year old children to learn a foreign language, e.g. English). In addition to a content typology, we also have to look at the structures of content. In this respect, and regardless of the content type, we can make use of terminology engineering, and, more recently, also ontology engineering. Terminologies and ontologies are the intellectual (conceptual) infrastructures of content, both

- implicitly (in the form of personal or subjective knowledge of the content generator), or
- explicitly (as objective knowledge laid down in a specific presentation form).

So we can conclude that concepts are content units (conceptual chunks) and that conceptual structures (the links among concepts) are the structures of concept. Again we have to remember that the multi-dimensional content typology will determine the concrete structures of content that users will encounter in specific products.

4. Global Content Management

After having investigated a little bit into the concept of content, we can now look at content management and how cultural diversity determines this practice. Since the target audience of any content product is always culture-bound, i.e. belonging to one or more cultures, with we can simply state that content management always has to take into account cultural factors in content design and all other processes and tasks of content management. The language(s) spoken by the target audience, social and historical factors, among many others, are examples of criteria for concrete manifestations of content management. Also the meta-level of content management, i.e. those who are content managers, are also culture-bound. Those who have designed and created content products, such as multimedia encyclopedias on CD-ROM, have to be aware that they themselves are belonging to at least one culture (in most cases, there will be one pre-dominant culture in such content management teams), and that this very fact will unavoidably determine the way the content of the product is designed.

Now we look at a list of key processes of content management:

- Design and creation of content
- Processing of content, such as

Analysis of existing content structures, segmentation of content into units, aggregation of content units into structures, condensation of content (summarization, abstracting, etc.), expansion of content into more detailed forms, transformation of content, etc.

- Presentation of content in different media and knowledge representation forms (see above)
- Dissemination of content on intranets or other web structures, on CD-ROMs, but also more traditionally in the form of books, etc.
- Sharing content in collaborative workspaces
- Using content for various purposes

Taking into consideration the differentiation between data, information, knowledge, and content (see above), we can make a parallel distinction between data management, information management, knowledge management, and content management. It is important to note that each management level is based on the one underneath, i.e. information management is impossible without data management, knowledge management needs both, data management and information management, and content management relies on all three levels below. The following figure shows different levels of complexity and levels of integration. As a result of combining these two dimensions, degrees of usability can be differentiated: data management is usually not user-oriented, since it is an internal process at an infrastructural level. Content management, on the other end, is most user-oriented.

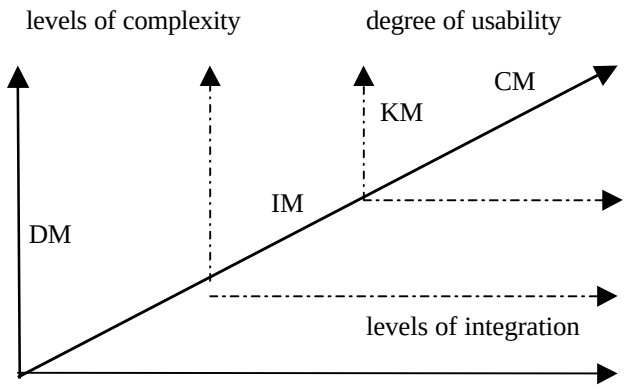


Figure 1: Levels of complexity and levels of integration, and degrees of usability as emergence levels of data management (DM), information management (IM), knowledge management (KM), and content management (CM)

Now we should return to the aspect of cultural diversity and the way it determines content management. Global content design, accordingly, is an activity of designing content for different cultures as target groups and is cognizant of the fact that content design itself is a culture-bound process, as shown above.

From the field of cultural studies we can benefit when looking at definitions of what culture is: a specific mind set, collective thinking and discourse patterns, assumptions, world models, etc.

Examples for types of culture are corporate cultures, professional, scientific cultures, notably going well beyond the national level of distinguishing cultures.

Cultural diversity is both a barrier and at the same time an asset and certainly the raison d'être for translation, localization, etc.

The following model shows the various dimensions of Global Content Management discussed above. The term element 'global' stands for all the cross-cultural activities such as translation, localization, but also customization, etc. 'Content' includes terminologies and ontologies as its infrastructures, products and their design, user documentation, but also pieces of art, etc. And the management component includes all the processes such as markup and modelling, processing, but also quality management, communication at the meta level, etc. Usability engineering is crucial for all these components:

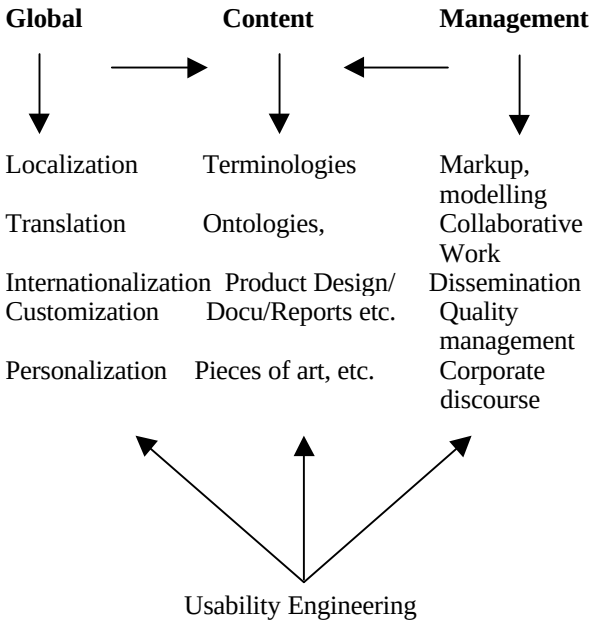


Figure 2: the three components of global content management with individual processes and components, all three nowadays determined by usability engineering imperatives

5. Pragmatic Issues in Global Content Management

Content management processes cannot do without appropriate knowledge organization and content organization. Terminological concept systems are organized into Knowledge Organization Systems (KOS) that can be used for this purpose of content organization:

- Thesauri, Classification Systems, and other KOSs, also conceptualized as (extrinsic) ontologies
- (Intrinsic) Ontologies (language-related, e.g. WordNet), domain-specific (medicine, etc.)

In order to establish and maintain the interoperability among heterogeneous content management systems, federation and networking of different content organization systems are necessary in order to facilitate topic-based content retrieval and exchange of content in B2B interactions.

Global Content Management may have very different manifestations. In the area of Cultural Content Management, for instance, cultural heritage technologies have developed in order to build up digital libraries, digital archives and digital museums.

Other applications of Global Content Management systems are:

- ePublishing (single source methodologies)
- eLearning (managing teaching content
- Cyber Science (Collaborative Content Creation)
- Digital Cities and other Virtual Communities projects.

On the pragmatic level of maintaining content management systems we observe similar problems as on the level of knowledge management, that a corporate

culture of knowledge sharing has to be developed and nurtured, that special communicative and informational skills are needed to share knowledge across cultures and that the dynamic changes in content require a management philosophy that is fully cognizant of the daily implications of these constant changes.

Translation resources such as translation memories and other aligned corpora, multilingual terminological resources, reference resources, etc. are typical examples of content that needs to be managed in such global action spaces.

6. Outlook

On the technological level a number of enabling technologies for global content management have emerged that are converging into Semantic Web technologies. Intelligent information agents are integrated into such systems. They are combined with knowledge organization systems (in particular multilingual ontologies). Semantic interoperability has also become a major field of research and development in this respect.

In the field of the so-called content industry different business models have developed that could not be more diverse: on the one hand open source and open content approaches are rapidly gaining momentum, also facilitated by maturing Linux-based applications. On the other hand national, regional and international legislation concerning intellectual property rights is becoming more and more strict and global players are buying substantial portions of cultural heritage for digitisation and commercial exploitation that might eventually endanger the public nature of cultural heritage.

Epistemological issues of global content management will have to be addressed, as well as best practices to be studied in detail in order to develop advanced methods for these complex management tasks. Managing cultural diversity in a dynamic market with rapidly changing consumer interests and preferences, with new technologies to be integrated, also requires a strategy for sustainable teaching and training initiatives (based on knowledge management teaching and training initiatives) in this fascinating field.

7. References

Davenport, Thomas H./Prusak, Laurence (1998). *Working Knowledge. How Organizations Manage What They Know*. Boston: Harvard Business School Press

Hoffmann, Cornelia/Mehnert, Thorsten (2000). „Multilingual Information Management at Schneider Automation“ Robert C. Sprung (Hrsg.). *Translating Into Success. Cutting-edge strategies for going multilingual in a global age*. (pp. 59-79) Amsterdam/Philadelphia: John Benjamins

Holden, Nigel J. (2002). *Cross-cultural Management. A Knowledge Management Perspective*. Harlow: Pearson

Nonaka, Ikujiro/Takeuchi, Hirotaka (1995). *The Knowledge-Creating Company*. Oxford University Press

Popper, Karl (1972). *Objective Knowledge. An Evolutionary Approach*. London: Routledge

Tiwana, Amrit (2000). *The Knowledge Management Toolkit. Practical Techniques for Building a Knowledge Management System*. Upper Saddle River: Prentice Hall

TFPL (1999). *Skills for Knowledge Management: building a knowledge economy*. London: TFPL

Trompenaars, Fons/Hampden-Turner, Charles (1993/2001). *Riding the Waves of Culture. Understanding Cultural Diversity in Business. 2nd edition.* London: Nicholas Brealey Publishing

Wright, Sue Ellen/Budin, Gerhard (comp.) (1997, 2001). *Handbook of Terminology Management*. 2 volumes. Amsterdam/Philadelphia: John Benjamins