1998 EAMT WORKSHOP

Geneva, 2-3 April 1998

EUROPEAN
ASSOCIATION
FOR MACHINE
TRANSLATION

Translation technology: integration in the workflow environment

Proceedings

2-3 April 1998
WHO, Geneva

EAMT Workshop, Geneva, 2-3 April 1998

Translation technology:
integration in the workflow environment

This workshop was the third of a series, following two previous workshops in Vienna in August 1996 and in Copenhagen in May 1997. Its theme was the practical use and integration of machine translation and computer-based translation tools in the workflow of a variety of organisations.

Held by the Committee of EAMT (European Association for Machine Translation)
President: John Hutchins
Secretary: Viggo Hansen
Treasurer: Doris Marty-Albisser
Members: Colin Brace
Bente Maegaard
Dimitrios Theologitis
Jörg Schütz

Venue:
Salle B, Third floor, Main Building,
World Health Organization
20 avenue Appia, CH-1211 Geneva 27

Local Organiser: Olivier Pasteur, Computer-Assisted Translation and Terminology Unit, Room 3029, World Health Organization (Tel (41-22) 791 2317; Fax (41-22) 791 3995; Email: pasteuro@who.ch)

Table of Contents

Preface

John Hutchins

The EAMT workshops held in the past three years have been devoted to exploring the practicalities of using machine translation software and other computer-based translation tools in organisations. In these environments, translation has to be seen by management as a positive enhancement to the promotion and sales of company products. For those involved in the provison of company translations the cost-effective exploitation of the most appropriate translation software and the careful design of operational workflows are of crucial importance. At the 1997 workshop in Copenhagen, it was decided that the meeting in Geneva should focus on the integration of MT and translation tools in the overall documentation workflow.

The proceedings published in this volume include all the papers (with one exception) presented at the workshop in Geneva in April 1998, plus one paper submitted but not presented (the description of PARS by Mikhael Blekhman) and three presentations not originally included in the programme. These are the three papers from Paul Kaeser, Jacqueline van Wees and Thorsten Mehnert. Our thanks to all the speakers for making this workshop as successful and informative as the preceding workshops, and to Olivier Pasteur and his colleagues at WHO for the local arrangements in Geneva.

An important feature of all the EAMT Workshops has been the extensive and fruitful discussion sessions. At Geneva there were two discussion periods, one led by Viggo Hansen and the second by Dimitrios Theologitis. We would have liked to have included reports on both, but in the event it was possible only to record the final session.

It has been decided that the next workshop in this series should take place in Prague, and that it should concentrate on the needs and problems of translation services in Central and Eastern Europe, particularly in the light of the forthcoming entry of countries into the European Union. Arrangements for this Workshop are now well in hand. It will take place at the Hotel Krystal in Prague on 22nd and 23rd April 1999, and is being organised by Eva Hajicova and colleagues at the Charles University. It promises to be a successful and enjoyable meeting, and we hope that many of those who have attended previous EAMT workshops will be able to be in Prague.

Workflow using linguistic technology at the Translation Service of the European Commission

Achim Blatt

1. INTRODUCTION

The following describes the different translation technologies which are currently available in the Translation Service (SdT) of the European Commission. In order to exemplify their usage, they are described in a (maximalistic) workflow.

For almost any of the technologies described below, it is essential to have the source document available in electronic form. In order to give requesters an incentive to use electronic mail, the SdT has developed a simple and user-friendly interface, known as POETRY (Processing of Electronic Translation Requests), which allows users to send a translation request together with the document to be translated and, if possible, reference material. This request is then passed on to individual translators via WINSUIVI, a management tool which makes it possible for allowing the work to be allocated according to the target languages and products required.

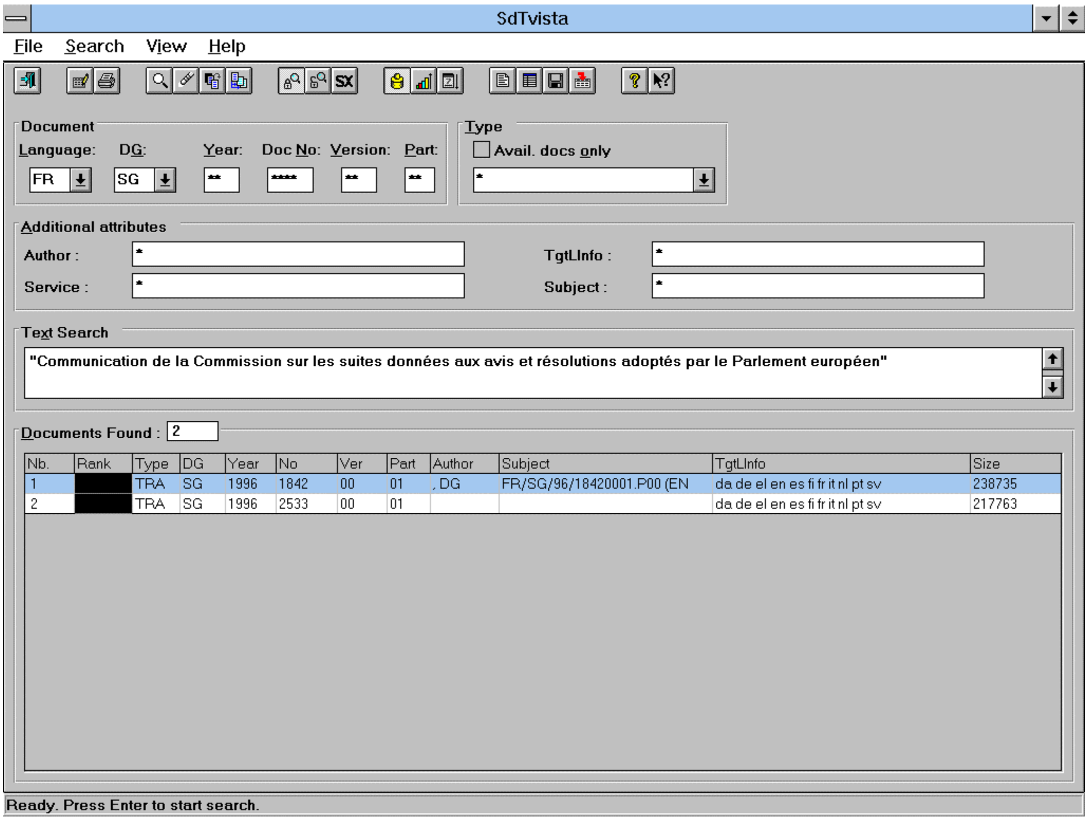
If one abstracts away from organisational aspects (which parts are done by secretaries, which parts are done by translators etc), a difference has to be made between the preparation of a translation, and the translation itself. In the worst case, a translation memory has to be created from zero.

2. BUILDING-UP A TRANSLATION MEMORY (TM)

In the SdT, translation memories are increasingly often used in order to improve productivity as well as quality and coherence. If users want to use TM technology, possibly combined with machine translation or replacement tools, and if translation memories do not yet contain enough suitable data, they have to look for reference material which can be imported.

2.1. Reference documents

The Commission's translators have a number of on-line full-text databases, which can be searched for reference material. SDTVISTA contains almost all the translations from 1994 onwards (except for confidential documents) plus many source texts and a number of reference documents. This allows users to check whether or not a document, or part of it, has already been translated and to retrieve pertinent source documents as well as their translations. Queries are typically made on the basis of search strings, but they can be refined by means of additional filters (e.g. requesting service, year, translation type etc.):



Document

Language: FR DG: SG Year: ** Doc No: **** Version: ** Part: **

Type

□ Avail. docs only

*

Additional attributes

Author : *

Service : *

TgtLInfo : *

Subject : *

Text Search

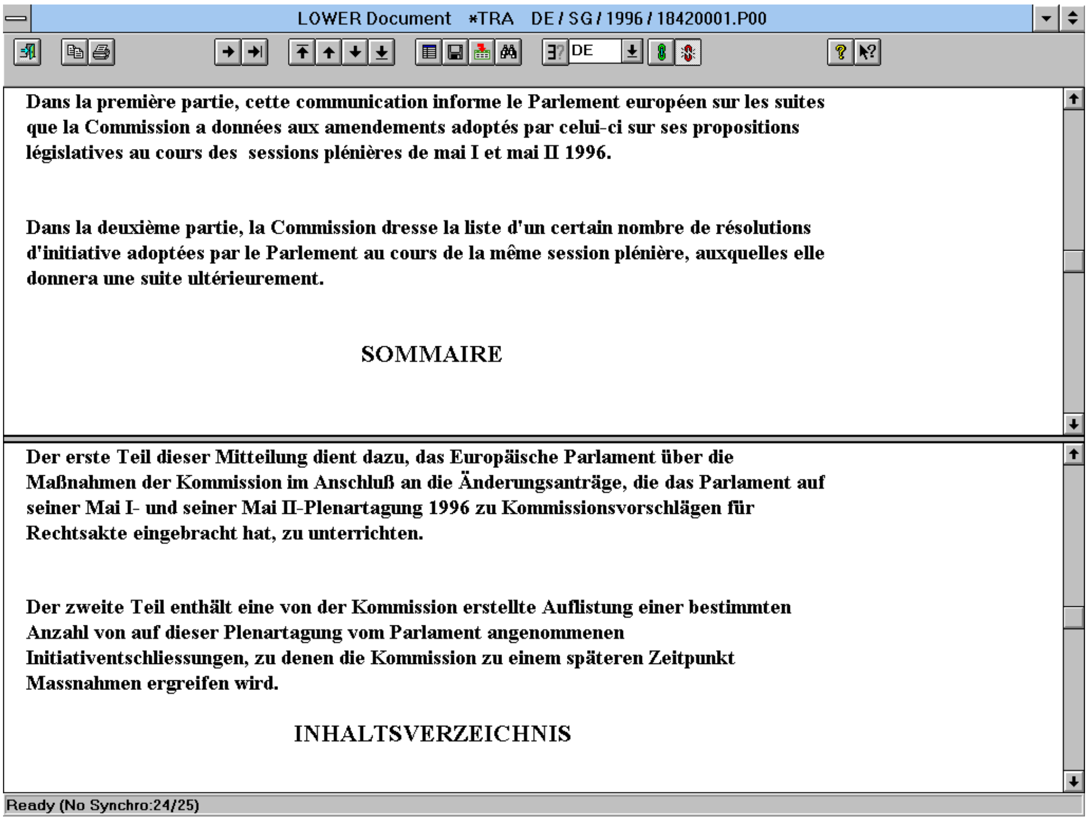
"Communication de la Commission sur les suites données aux avis et résolutions adoptés par le Parlement européen"

Documents Found : 2

Nb.	Rank	Type	DG	Year	No	Ver	Part	Author	Subject	TgtLInfo	Size
1	■	TRA	SG	1996	1842	00	01	, DG	FR/SG/96/18420001.P00 (EN	da de el en es fi fr it nl pt sv	238735
2	■	TRA	SG	1996	2533	00	01			da de el en es fi fr it nl pt sv	217763

Ready. Press Enter to start search.

If interesting material can be found, it can be viewed or downloaded for further treatment. During the translation process, translators might also consult SDTVISTA in order to solve terminological problems:



Dans la première partie, cette communication informe le Parlement européen sur les suites que la Commission a données aux amendements adoptés par celui-ci sur ses propositions législatives au cours des sessions plénières de mai I et mai II 1996.

Dans la deuxième partie, la Commission dresse la liste d'un certain nombre de résolutions d'initiative adoptées par le Parlement au cours de la même session plénière, auxquelles elle donnera une suite ultérieurement.

SOMMAIRE

Der erste Teil dieser Mitteilung dient dazu, das Europäische Parlament über die Maßnahmen der Kommission im Anschluß an die Änderungsanträge, die das Parlament auf seiner Mai I- und seiner Mai II-Plenartagung 1996 zu Kommissionsvorschlägen für Rechtsakte eingebracht hat, zu unterrichten.

Der zweite Teil enthält eine von der Kommission erstellte Auflistung einer bestimmten Anzahl von auf dieser Plenartagung vom Parlament angenommenen Initiativentschliessungen, zu denen die Kommission zu einem späteren Zeitpunkt Massnahmen ergreifen wird.

INHALTSVERZEICHNIS

Ready (No Synchro:24/25)

CELEX is the full text database of the Office for Official Publications of the European Communities. It offers multilingual coverage of a wide range of legal documents. CELEX offers a user-friendly Internet access with a large number of different query types.

The SdT provides an automatic batch retrieval where on the basis of references found in a document, CELEX identifiers are calculated and the corresponding documents are returned to the user. For example, the German sentence

(1) In der Verordnung (EWG) Nr. 210/69 der Kommission zuletzt geändert durch die Verordnung (EG) Nr. 1171/96, sind die Informationen zur Verwaltung des Marktes für Milcherzeugnisse festgelegt, die der Kommission regelmäßig mitzuteilen sind

yields the identifiers 369R0210 and 396R1171, which are then used for remote CELEX queries of titles or complete documents. The example above leads to the following German and English titles:

(2a) Verordnung (EWG) Nr. 210/69 der Kommission vom 31. Januar 1969 über die gegenseitigen Mitteilungen der Mitgliedstaaten und der Kommission im Sektor Milch und Milcherzeugnisse

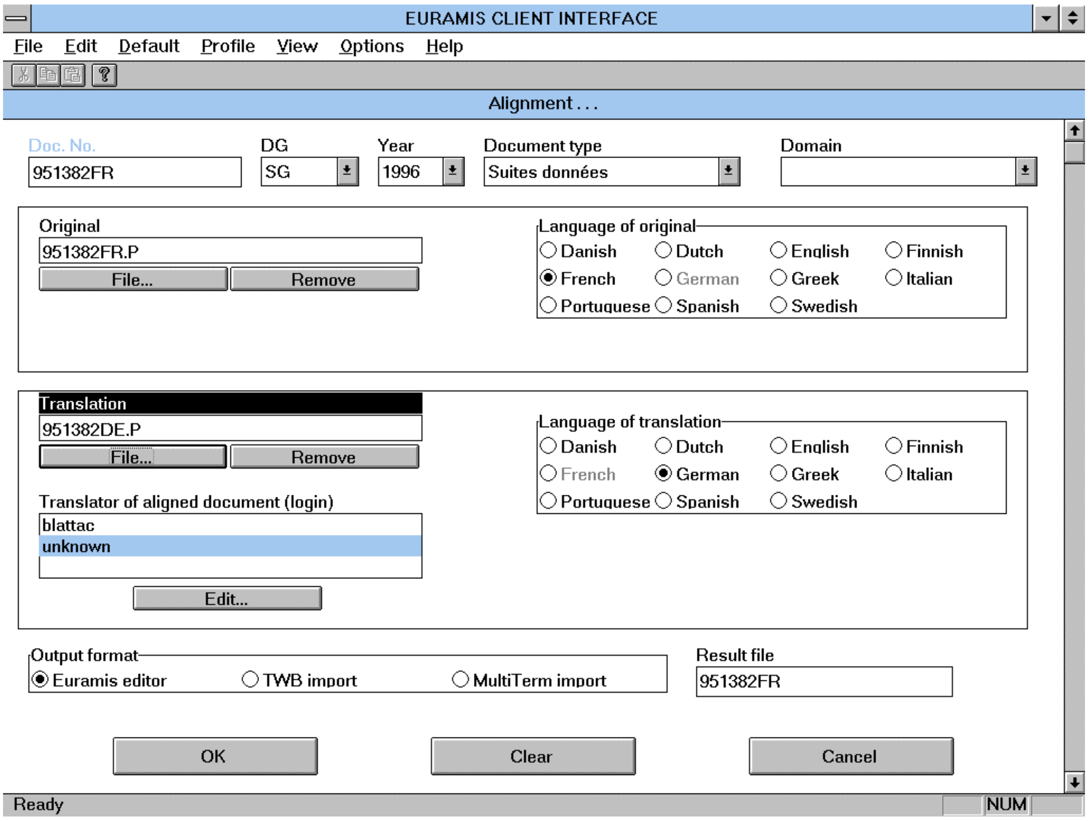
(2b) Regulation (EEC) No 210/69 of the Commission of 31 January 1969 on communications between Member States and the Commission with regard to milk and milk products

(3a) Verordnung (EG) Nr. 1171/96 der Kommission vom 27. Juni 1996 zur Änderung der Verordnung (EWG) Nr. 210/69 über die gegenseitigen Mitteilungen der Mitgliedstaaten und der Kommission im Sektor Milch und Milcherzeugnisse

(3b) Commission Regulation (EC) No 1171/96 of 27 June 1996 amending Regulation (EEC) No 210/69 on communications between Member States and the Commission with regard to milk and milk products

2.2. Alignment

Interesting reference documents and their translations can be downloaded to the user’s PC, and a sentence alignment request can be launched:



Doc. No. 951382FR

DG SG

Year 1996

Document type Suites données

Domain

Original 951382FR.P

Language of original

○ Danish ○ Dutch ○ English ○ Finnish
● French ○ German ○ Greek ○ Italian
○ Portuguese ○ Spanish ○ Swedish

Translation 951382DE.P

Translator of aligned document (login)
blattac
unknown

Language of translation

○ Danish ○ Dutch ○ English ○ Finnish
○ French ● German ○ Greek ○ Italian
○ Portuguese ○ Spanish ○ Swedish

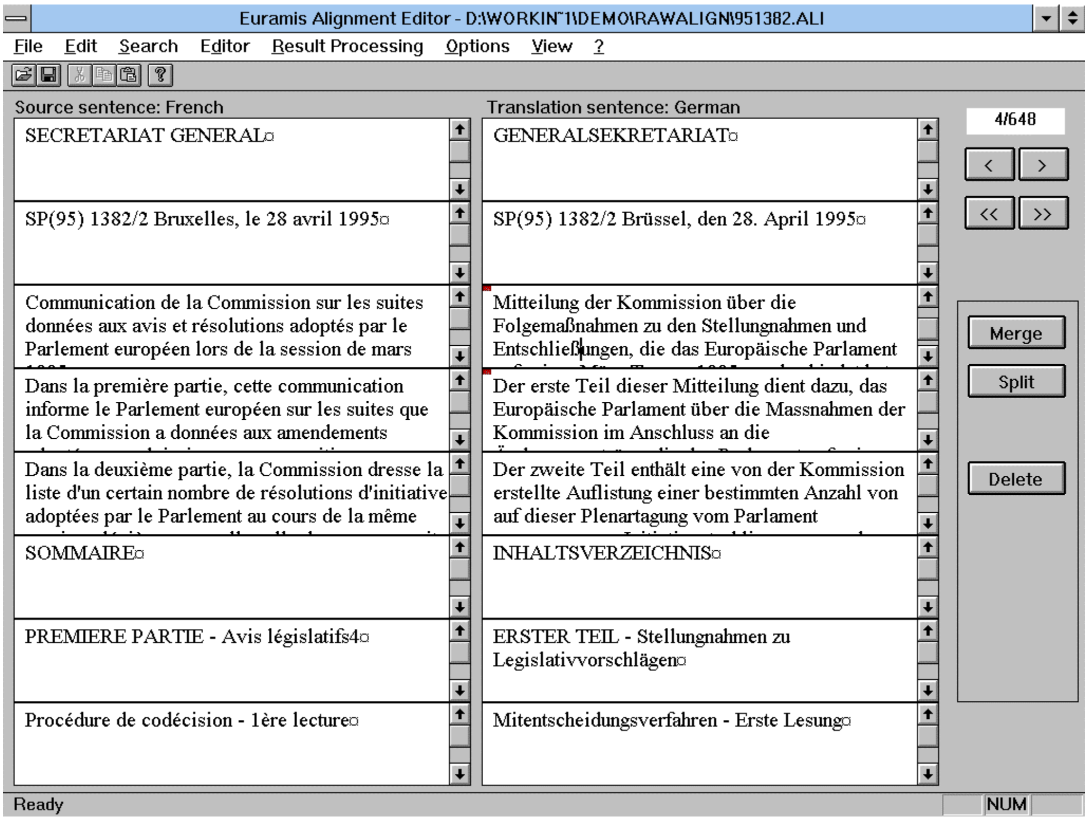
Output format
● Euramis editor ○ TWB import ○ MultiTerm import

Result file 951382FR

The SdT uses its own aligner which is highly customised and therefore produces markedly better results than commercially available applications.

For CELEX documents, a pre-processing is carried out (currently still on the client side so that users can still modify texts before they are aligned). This pre-processing makes it possible to iron out the language-specific differences which affect alignment quality. Example: in a number of languages (English, French, Spanish etc.), several “whereas”-clauses are frequently put in one sentence (separated by a semi-colon); for German, Danish, Swedish and Greek, new sentences are created in these cases; if many such differences occur in a short piece of text, this can lead to mis-alignment. The solution is to insert a paragraph marker if a semicolon is followed by “whereas” (or its equivalent in another language).

Alignment results can be corrected by a special editor and later imported into a translation memory:



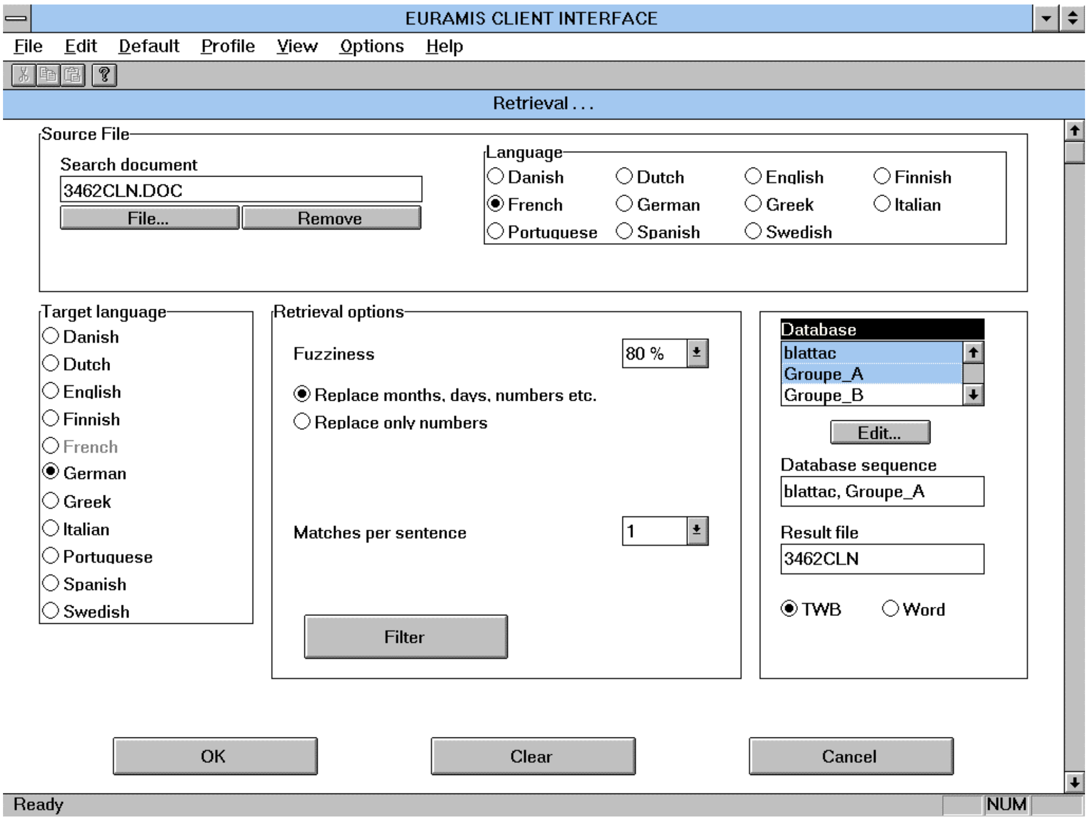
Source sentence: French	Translation sentence: German
SECRETARIAT GENERAL¤	GENERALSEKRETARIAT¤
SP(95) 1382/2 Bruxelles, le 28 avril 1995¤	SP(95) 1382/2 Brüssel, den 28. April 1995¤
Communication de la Commission sur les suites données aux avis et résolutions adoptés par le Parlement européen lors de la session de mars	Mitteilung der Kommission über die Folgemaßnahmen zu den Stellungnahmen und Entschließungen, die das Europäische Parlament
Dans la première partie, cette communication informe le Parlement européen sur les suites que la Commission a données aux amendements	Der erste Teil dieser Mitteilung dient dazu, das Europäische Parlament über die Massnahmen der Kommission im Anschluss an die
Dans la deuxième partie, la Commission dresse la liste d'un certain nombre de résolutions d'initiative adoptées par le Parlement au cours de la même	Der zweite Teil enthält eine von der Kommission erstellte Auflistung einer bestimmten Anzahl von auf dieser Plenartagung vom Parlament
SOMMAIRE¤	INHALTSVERZEICHNIS¤
PREMIERE PARTIE - Avis législatifs4¤	ERSTER TEIL - Stellungnahmen zu Legislativvorschlägen¤
Procédure de codécision - 1ère lecture¤	Mitentscheidungsverfahren - Erste Lesung¤

3. Preparation of work

If users want to work exclusively with their own local translation memory, they can start translating at this point. They can however request additional information coming from other sources via the EURAMIS client interface which gives access to a number of batch services.

3.1. TM

The SdT has developed server translation memories whose design facilitates data sharing and the combination with other products. Central translation memories are provided for each of the seven thematic groups of the SdT. In addition, there are a number of consolidated topical translation memories (e.g. most important European legislation and Court rulings, the monthly Bulletin etc.). For each of these translation memories, a coordinator grants write access, sees to a coherent use of attributes etc. In addition to common TMs, there are personal TMs for every translator: there are no restrictions concerning read access, everybody can retrieve from everybody’s TM. The following shows a typical TM retrieval request:



Source File

Search document: 3462CLN.DOC

Language: ○ Danish ○ Dutch ○ English ○ Finnish ● French ○ German ○ Greek ○ Italian ○ Portuguese ○ Spanish ○ Swedish

Target language: ○ Danish ○ Dutch ○ English ○ Finnish ○ French ● German ○ Greek ○ Italian ○ Portuguese ○ Spanish ○ Swedish

Retrieval options

Fuzziness: 80 %

● Replace months, days, numbers etc.

○ Replace only numbers

Matches per sentence: 1

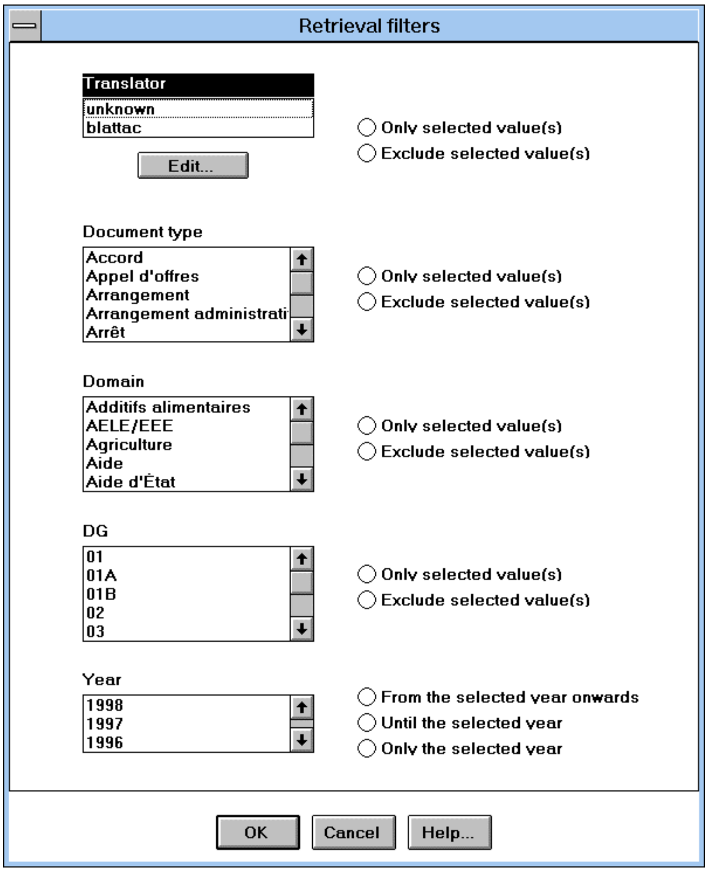
Database: blattac, Groupe_A, Groupe_B

Database sequence: blattac, Groupe_A

Result file: 3462CLN

● TWB ○ Word

EURAMIS TM queries can be fine-tuned by means of positive and negative filters ("only these" or "all but these"). The following filters are possible: requesting service (e.g. Directorate-General XXII), domain (e.g. agriculture), document type (e.g. letter) and individual translator (unique login).



Translator: unknown, blattac

○ Only selected value(s)
○ Exclude selected value(s)

Document type: Accord, Appel d'offres, Arrangement, Arrangement administrati, Arrêt

○ Only selected value(s)
○ Exclude selected value(s)

Domain: Additifs alimentaires, AELE/EEE, Agriculture, Aide, Aide d'État

○ Only selected value(s)
○ Exclude selected value(s)

DG: 01, 01A, 01B, 02, 03

○ Only selected value(s)
○ Exclude selected value(s)

Year: 1998, 1997, 1996

○ From the selected year onwards
○ Until the selected year
○ Only the selected year

3.2. Machine translation (MT)

MT is currently used in the Commission for 17 language pairs with a quality which varies considerably between the different languages: French-Spanish offers the best quality, followed by French-English, French-Italian and English-French; language pairs with German produce less acceptable results.

MT is available not only to translators, but to all Commission officials who are equipped with either a terminal or a PC connected to the internal network. The number of pages translated by the system has increased considerably in recent years: 170 000 in 1995, 231 000 in 1996, 260 000 estimated for 1997. MT is mainly used for

- the fast translation of short, repetitive texts with a standardised structure and terminology (mail, minutes of meetings, parliamentary questions, reports etc.);
- the browsing of texts written in a language the user does not know;
- drafting purposes: users write a text in their mother tongue and request a machine translation in order to have a document drafted in something other than their native or main language.

MT is appreciated by translators because of its speed, its capacity to keep the original format, and as a terminology aid.

3.3. TMan

TMan replaces pre-defined strings (from words up to paragraphs) in the source document so that the resulting document is a mix of source and target language items. TMan replacements are based on a repetition analysis of the document type in question. This means that the use of TMan is only worthwhile if a document type is quite frequent and if it contains a large number of repetitive elements.

This approach is taken for a number of master documents and regular publications (e.g. the Bulletin) where many expressions can be found in every issue, and consequently the analysis of previous texts yields large expression databases to be used for subsequent issues of the same publications. For example, the following French text (from a prototype contract)

(4a) La Communauté européenne, représentée par la Commission des Communautés européennes, ici représentée par Monsieur ..., Directeur général,d'une part, et la firme "...", dont le siège social est à ... ci-après dénommée "le fournisseur", représentée par Monsieur ...
en vertu de la délégation lui conférée par ladite société

would be translated as (4b) where colours (put in bold here) show what remains to be translated:

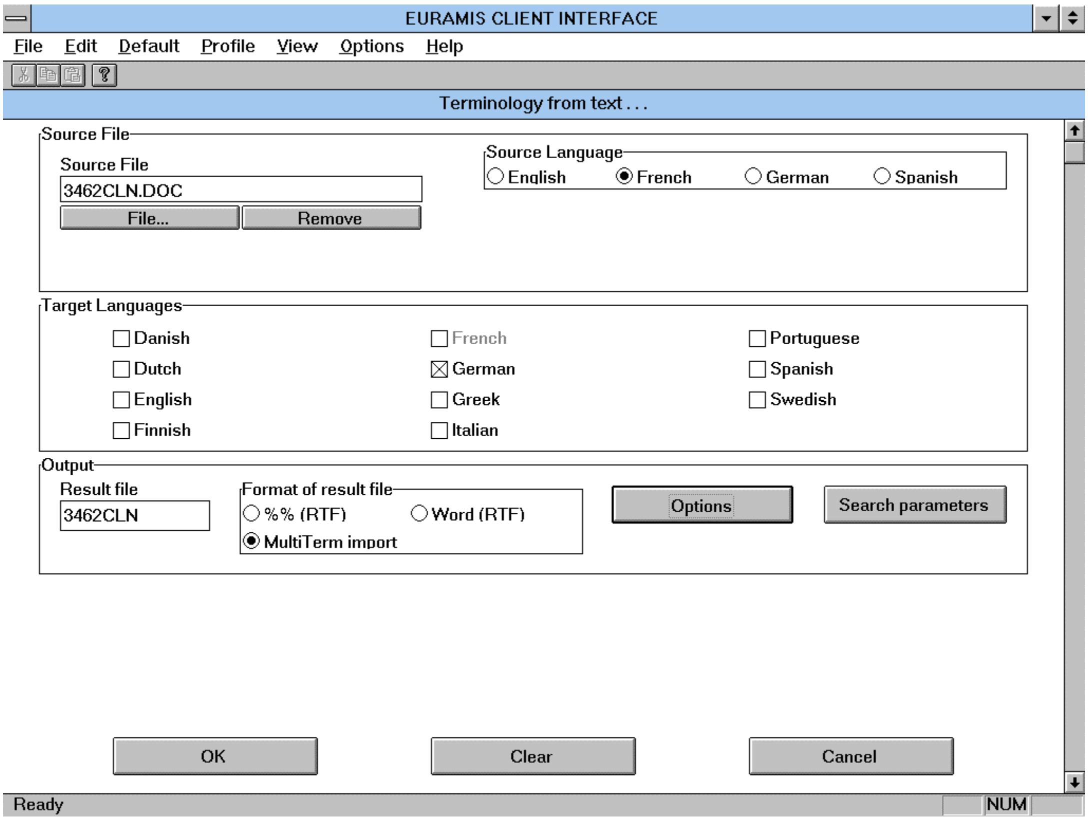
(4b) **La Communauté européenne, représentée par la Commission** de las Comunidades Europeas**, ici représentée par Monsieur ...,** Director General**,** y, de otra, y **la firme "...",** con domicilio social en ... denominada en lo sucesivo "el proveedor", representada por el Sr **...**
según el poder otorgado por la citada sociedad,

3.4. Combination of products

It is possible to combine TM results with MT (the same integration with TMan is foreseen): on the basis of the user requirements, a EURAMIS TM retrieval is carried out, and the remaining gaps are filled with MT results.

3.5. Document-related terminology

EURAMIS provides automatic terminology extraction of pertinent EURODICAUTOM entries, together with indirect access to TIS and EUTERPE, the terminology databases of the European Council of Ministers and the European Parliament. Since the Commission's MT system is used as a supporting application, this service is restricted to documents in English, French, German and Spanish. The user can choose between the following output formats: EURODICAUTOM import format (for data correction), a sequential RTF file, and MULTITERM input format.



Source File

Source File: 3462CLN.DOC

Source Language: ○ English ● French ○ German ○ Spanish

Target Languages

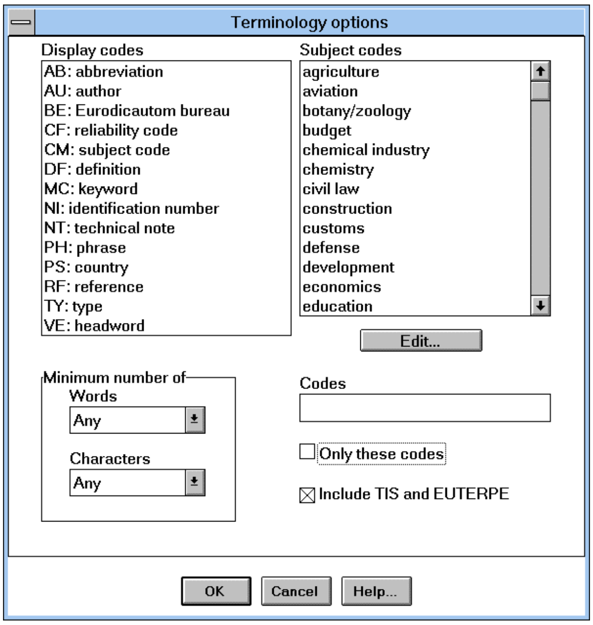
□ Danish □ French □ Portuguese
□ Dutch ☒ German □ Spanish
□ English □ Greek □ Swedish
□ Finnish □ Italian

Output

Result file: 3462CLN

Format of result file: ○ %% (RTF) ○ Word (RTF) ● MultiTerm import

Queries can be restricted in a number of ways (e.g. by indicating domain). Output can be restricted by selecting specific EURODICAUTOM fields:



Display codes

AB: abbreviation
AU: author
BE: Eurodicautom bureau
CF: reliability code
CM: subject code
DF: definition
MC: keyword
NI: identification number
NT: technical note
PH: phrase
PS: country
RF: reference
TY: type
VE: headword

Subject codes

agriculture
aviation
botany/zoology
budget
chemical industry
chemistry
civil law
construction
customs
defense
development
economics
education

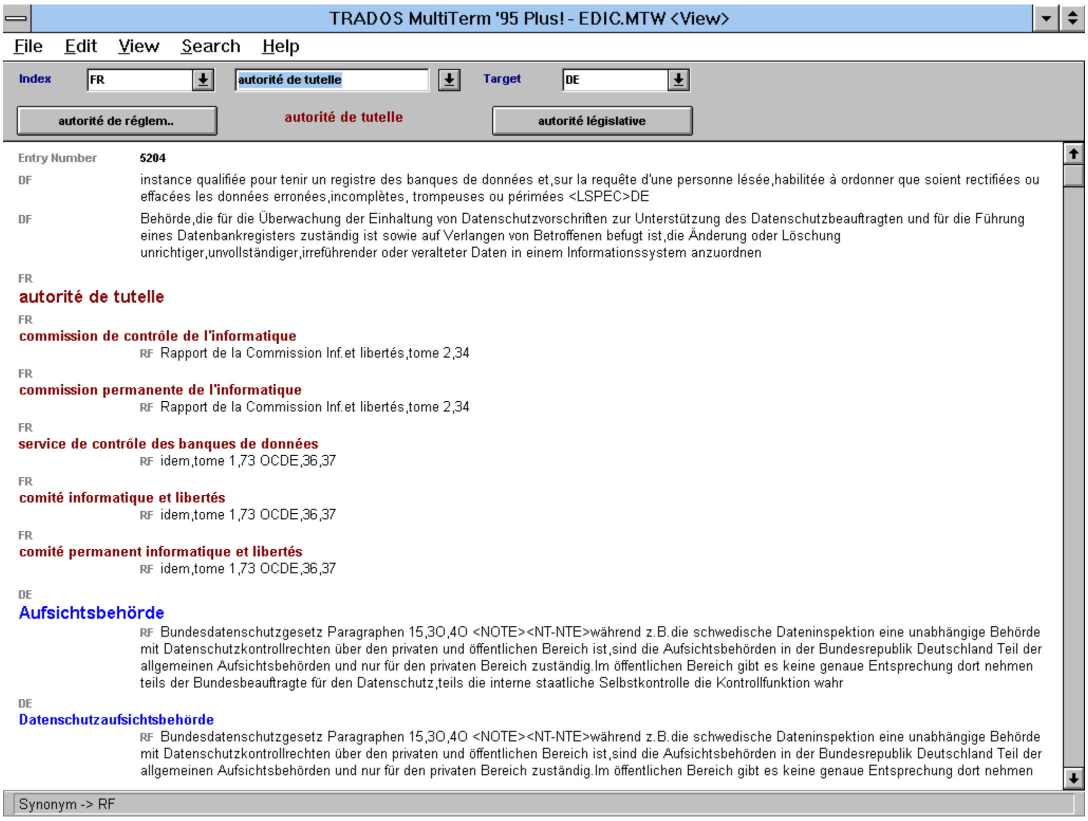
Minimum number of
Words: Any
Characters: Any

Codes

□ Only these codes

☒ Include TIS and EUTERPE

As a first step towards migrating EURODICAUTOM to a relational database management system, a new database structure has been designed recently: among other things, formal restrictions are imposed on values and synonyms are put in separate fields (the latter sometimes leads to some overgeneration, which can be seen in the notes of the German entries of the example below). A conversion to this new data model (including some clean-up and normalisation) is already used for output in MULTITERM:


TRADOS MultiTerm '95 Plus! - EDIC.MTW <View>
File Edit View Search Help
autorité de réglem..
autorité législative

Index: FR | autorité de tutelle | Target: DE

autorité de tutelle

Entry Number 5204

DF instance qualifiée pour tenir un registre des banques de données et,sur la requête d'une personne lésée,habilitée à ordonner que soient rectifiées ou effacées les données erronées,incomplètes, trompeuses ou périmées <LSPEC>DE

DF Behörde,die für die Überwachung der Einhaltung von Datenschutzvorschriften zur Unterstützung des Datenschutzbeauftragten und für die Führung eines Datenbankregisters zuständig ist sowie auf Verlangen von Betroffenen befugt ist,die Änderung oder Löschung unrichtiger,unvollständiger,irreführender oder veralteter Daten in einem Informationssystem anzuordnen

FR
autorité de tutelle

FR
commission de contrôle de l'informatique
RF Rapport de la Commission Inf.et libertés,tome 2,34

FR
commission permanente de l'informatique
RF Rapport de la Commission Inf.et libertés,tome 2,34

FR
service de contrôle des banques de données
RF idem,tome 1,73 OCDE,36,37

FR
comité informatique et libertés
RF idem,tome 1,73 OCDE,36,37

FR
comité permanent informatique et libertés
RF idem,tome 1,73 OCDE,36,37

DE
Aufsichtsbehörde
RF Bundesdatenschutzgesetz Paragraphen 15,3O,4O <NOTE><NT-NTE>während z.B.die schwedische Dateninspektion eine unabhängige Behörde mit Datenschutzkontrollrechten über den privaten und öffentlichen Bereich ist,sind die Aufsichtsbehörden in der Bundesrepublik Deutschland Teil der allgemeinen Aufsichtsbehörden und nur für den privaten Bereich zuständig.Im öffentlichen Bereich gibt es keine genaue Entsprechung dort nehmen teils der Bundesbeauftragte für den Datenschutz,teils die interne staatliche Selbstkontrolle die Kontrollfunktion wahr

DE
Datenschutzaufsichtsbehörde
RF Bundesdatenschutzgesetz Paragraphen 15,3O,4O <NOTE><NT-NTE>während z.B.die schwedische Dateninspektion eine unabhängige Behörde mit Datenschutzkontrollrechten über den privaten und öffentlichen Bereich ist,sind die Aufsichtsbehörden in der Bundesrepublik Deutschland Teil der allgemeinen Aufsichtsbehörden und nur für den privaten Bereich zuständig.Im öffentlichen Bereich gibt es keine genaue Entsprechung dort nehmen

Synonym -> RF

4. TRANSLATION PROPER

In principle, users can translate in two different working environments; depending on their own preferences and on the type of the document to be translated, they can choose between TRADOS' Translator's Workbench and a simpler treatment completely inside word processing.

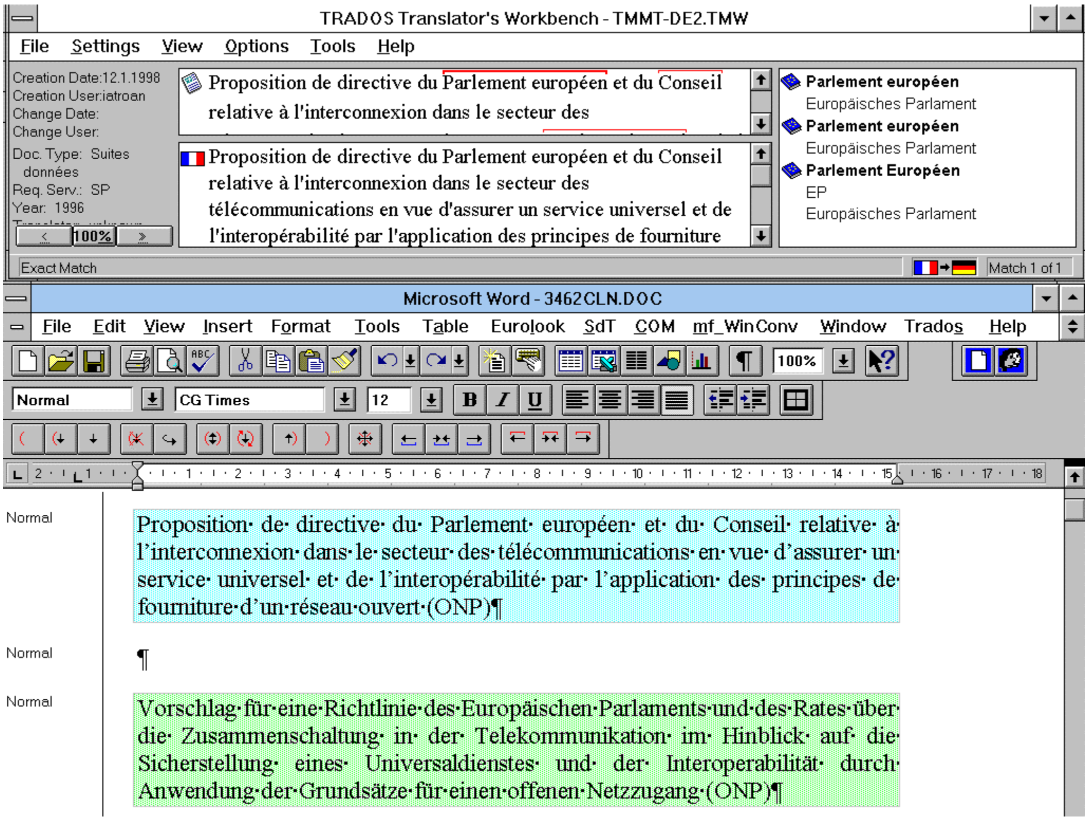
4.1. Translator's Workbench

With TWB, the products mentioned above can be used in parallel: retrieval from central TM is imported together with MT output; the latter receives a special attribute in order to warn users. Terminology retrieval can be imported directly to MULTITERM.

The main advantages of Translator's Workbench (as compared to a word processing approach) are the following:

- it is interactive, i.e. document-internal repetition can be exploited immediately;
- it is fully integrated with MULTITERM so that terminology can be consulted on the fly;
- it provides a number of additional features such as concordance and coverage analysis.

Its main disadvantage is that several applications have to share the limited space on the screen and that many people find it too complicated to use:


TRADOS Translator's Workbench - TMMT-DE2.TMW
File Settings View Options Tools Help
100%
Microsoft Word - 3462CLN.DOC
File Edit View Insert Format Tools Table Eurolook SdT COM mf_WinConv Window Trados Help
100%
Normal
CG Times
12
B I U
2 1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Normal
Normal
Normal

Creation Date:12.1.1998
Creation User:iatroan
Change Date:
Change User:
Doc. Type: Suites données
Req. Serv.: SP
Year: 1996

Proposition de directive du Parlement européen et du Conseil relative à l'interconnexion dans le secteur des

Proposition de directive du Parlement européen et du Conseil relative à l'interconnexion dans le secteur des télécommunications en vue d'assurer un service universel et de l'interopérabilité par l'application des principes de fourniture

Parlement européen
Europäisches Parlament
Parlement européen
Europäisches Parlament
Parlement Européen
EP
Europäisches Parlament

Exact Match
Match 1 of 1

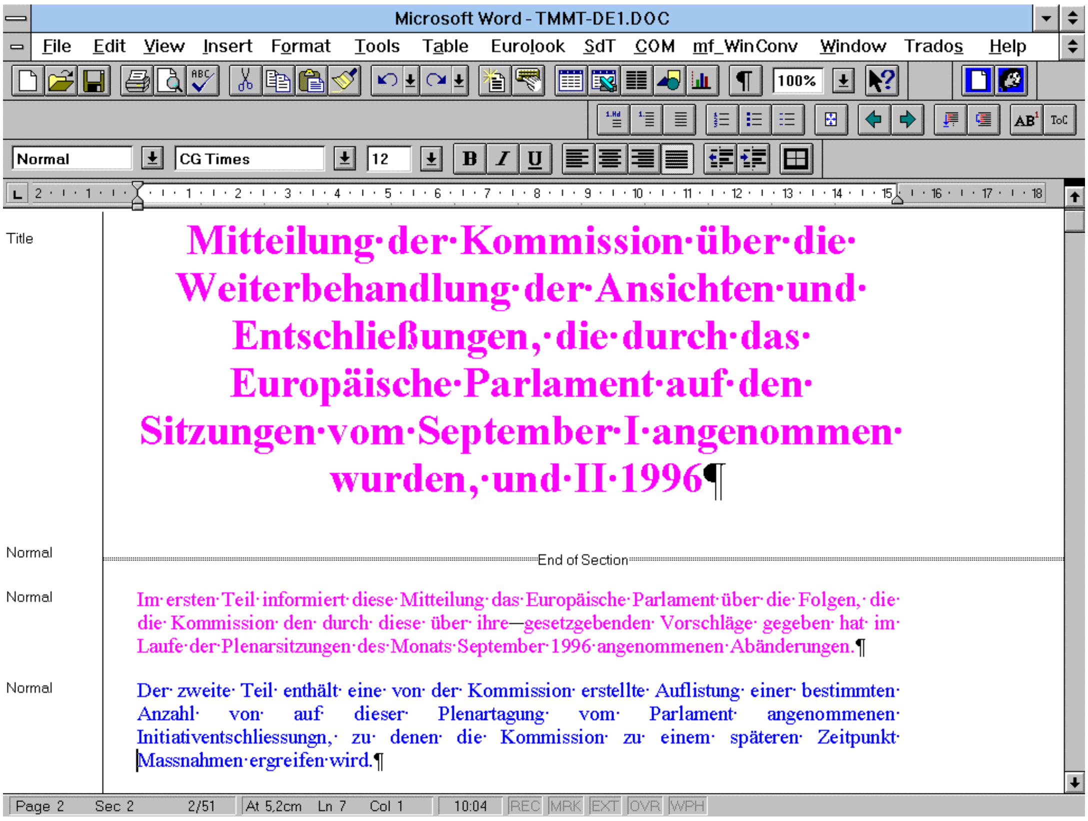
Proposition· de· directive· du· Parlement· européen· et· du· Conseil· relative· à· l'interconnexion· dans· le· secteur· des· télécommunications· en· vue· d'assurer· un· service· universel· et· de· l'interopérabilité· par· l'application· des· principes· de· fourniture·d'un·réseau·ouvert·(ONP)¶

¶

Vorschlag·für·eine·Richtlinie·des·Europäischen·Parlaments·und·des·Rates·über· die· Zusammenschaltung· in· der· Telekommunikation· im· Hinblick· auf· die· Sicherstellung· eines· Universaldienstes· und· der· Interoperabilität· durch· Anwendung·der·Grundsätze·für·einen·offenen·Netzzugang·(ONP)¶

4.2. Word processing only

If the result is requested in native Word format, the formatting of the source text is preserved to a very large extent[1]. Since colours are rare in Commission texts, they are used to convey information on result type and have to be reset after editing, e.g. blue for TM full matches, red for TM fuzzy matches, and magenta for MT.


Microsoft Word - TMMT-DE1.DOC
File Edit View Insert Format Tools Table Eurolook SdT COM mf_WinConv Window Trados Help
100%
Normal
CG Times
12
B I U
Title
Normal
Normal
Normal
End of Section
Page 2 Sec 2 2/51 At 5,2cm Ln 7 Col 1 10:04 REC MRK EXT OVR WPH

Mitteilung·der·Kommission·über·die·Weiterbehandlung·der·Ansichten·und·Entschließungen,·die·durch·das·Europäische·Parlament·auf·den·Sitzungen·vom·September·I·angenommen·wurden,·und·II·1996¶

Im·ersten·Teil·informiert·diese·Mitteilung·das·Europäische·Parlament·über·die·Folgen,·die·die·Kommission·den·durch·diese·über·ihre—gesetzgebenden·Vorschläge·gegeben·hat·im·Laufe·der·Plenarsitzungen·des·Monats·September·1996·angenommenen·Abänderungen.¶

Der·zweite·Teil·enthält·eine·von·der·Kommission·erstellte·Auflistung·einer·bestimmten·Anzahl·von·auf·dieser·Plenartagung·vom·Parlament·angenommenen·Initiativentschliessungn,·zu·denen·die·Kommission·zu·einem·späteren·Zeitpunkt·Massnahmen·ergreifen·wird.¶

5. FUTURE DEVELOPMENTS

Although the applications at hand make it possible to produce more coherent translations in less time, there are still two main disadvantages:

- too much interaction is still needed in order to prepare work (e.g. downloading reference documents, aligning them, importing alignments etc.);
- a certain degree of experience is necessary in order to determine which application(s) should be used under which circumstances.

[1] Contrary to MT (which "knows" what is translated by what), TM cannot establish a 1:1 correspondence between words. It is therefore not possible for TM to reproduce correct character formatting below sentence level (e.g. bold, italic etc.). The following solution has been implemented instead:

- if some character formatting is switched on AND off inside the same sentence, this formatting information is not taken over in the translation
- if some character formatting is only switched on OR off inside the same sentence, the switch is moved to the front of the sentence, i.e. the formatting information is taken up from the beginning of the sentence.

The first problem can partly be solved by combining existing modules. One example for this is the automatic retrieval of CELEX titles and documents, where a longer chain could be implemented:

- calculation of CELEX identifiers from references in source document;
- file transfer of identified documents from CELEX server;
- clean-up and normalisation of CELEX documents at server level;
- alignment (very reliable due to normalisation of CELEX documents);
- creation of an *ad-hoc* TM;
- combined search with source document in *ad-hoc* and other TMs.

The second problem can partly be solved by providing an expert system which suggests the most suitable treatment for a given text. As far as TM treatment is concerned, a recommendation could be based on the following analyses:

- calculate the degree of internal repetition of the source document (including fuzzy sentences): high value favours treatment with Translator's Workbench;
- create CELEX *ad-hoc* TM if possible, find most pertinent existing server TMs, calculate overall coverage of source document: high value favours TM treatment in general.

The next step would then be to integrate such an expert system in the production management application WINSUIVI, so that the queries necessary for the preferred treatment can already be launched and the results can be saved into a working directory before the translation request even arrives on the translator's desk.

6. BIBLIOGRAPHY

Blatt, A. (1996): The EURAMIS Project. In: Lauer, A., Gerzymisch-Arbogast, H., Haller, J., Steiner, E.: *Übersetzungswissenschaft im Umbruch*. Festschrift für Wolfram Wilss zum 70. Geburtstag. Tübingen 1996, pp. 131-134.

Blatt, A., Martins, P. (1997): EURAMIS, The European Advanced Multilingual Information System. *The ELRA Newsletter* 1997-2, pp. 3-5.

Reinke, U. (1997): Integrierte Übersetzungssysteme. Betrachtungen zu Übersetzungsprozeß, Übersetzungsproduktivität, Übersetzungsqualität und Arbeitssituation. *Lebende Sprachen* 1997-3, pp. 97-106.

A Centralized Approach to Managing Multiple Lexical Resources

Susan McCormick

ABSTRACT

The rapid expansion of SAP in markets around the world has brought with it an urgent need within the company for high-quality translation that adheres to SAP-specific terminology standards. Both the trend toward outsourcing and the increased use of automatic translation tools depend critically on quick and reliable access to official company terminology. SAP is therefore implementing a strategy that will make a central terminology database easily accessible not just to all of the people who need it (internal employees, customers, consulting agencies), but to translation and terminology tools as well. This includes the MT systems, Metal and Logos, whose data will be interchangeable with SAP database data.

1 BACKGROUND

As a large and growing international software developer, SAP has translation needs that can no longer be met using traditional translation management techniques. On-screen documentation alone is currently being translated into 26 target languages from source documents generated in either German or English. To address the expanding translation volume that has come with this rapid growth, the company has adopted more flexible strategies, notably the outsourcing of translation work to qualified agencies, and the use of new technology to automate and streamline the translation process.

A critical supporting element in this approach is the generation and management of company-specific terminology. In order for SAP to successfully outsource its translation jobs, it must be able to provide translators working in remote locations with the official terminology required for a given language and domain. And introducing new technology to make the translation process more efficient will work only if the integrity of company terminology can be assured. By increasing its reliance on outsourcing and job automation, therefore, SAP has also highlighted its need for a central store of company terminology that can be quickly and easily accessed by both translators and translation tools.

2 SAPterm and STERM

In-house translators have long used the central SAP terminology database, SAPterm, to help them with their translations. At present, SAPterm contains over 70,000 entries in approximately 130 SAP-specific subject areas. Programmed in R/3, SAPterm is accessible primarily to internal staff and can be updated by users with administrator's authorization only. Subsets of the database can be extracted and made available to external translators, customers, and partners, if required.

As SAP has expanded, SAPterm's usefulness has diminished, both in the information it contains and in its functionality. To address this, a new, second-generation SAP terminology database, STERM, is under development. STERM improves on SAPterm by offering 1) more terminological information, 2) improved coverage of the languages associated with newer SAP markets, 3) transparent links to glossaries, 4) general coding guidelines for all languages, and 5) open access to all internal SAP translators for updating. STERM also requires that terms include minimal grammatical information to allow for interchangeability with other terminological/lexical databases such as MT system lexicons or Trados‘ Multiterm.

While STERM is an improvement over SAPterm, it is considered a stepping stone to a more comprehensive approach that would make SAP terminology available, probably over the Web, to anyone with a ‘need to know.’ Open access of this sort would make it possible for SAP to take greater advantage of the linguistic expertise of its consultants, customers, and external translators by having them create directly the terminology they need for their languages and subject areas.

3 MT at SAP: Metal and Logos

The Multilingual Technology Department at SAP has used MT productively since 1991. There are currently two systems that are in active use, Metal’s German-English and Logos’ English-French. To run successfully, each of these systems must have system lexicons that are current with SAP’s terminology database. Right now, the Metal German-English lexicon contains upwards of 70,000 transfer entries in 49 SAP subject areas; the Logos installation has over 43,000 lexicon entries in 45 SAP subject areas and an additional 600 SAP semantic rules.

In order to keep the MT lexicons up-to-date, translators must constantly check official SAP terminology for changes and then make the appropriate entries and edits in the MT lexicons. This is usually done via the Metal or Logos lexicon interface, both of which support an entry-by-entry user processing mode. The process is often tedious and time-consuming, pointing up the fact that *essentially the same terminology set is being coded at least three times*[2]*, on three different platforms, in three different formats.*

4 Centralizing Terminology: The OTELO Central Lexical Database

While MT has allowed the MLT Department to post measurable productivity gains, the linguistic/administrative overhead associated with maintaining each system lexicon in isolation has appeared unnecessarily high. After several years of working around the problem, SAP decided to opt for a central repository of SAP terminological data that would be compatible with both Metal and Logos, i.e., a new database and format that would allow users to code terminology just once and then exchange it easily into other formats.

To achieve this, SAP became a full partner in the OTELO project, an EU project with the aim of developing a central translator’s environment. The environment would bring together already existing technology by offering unifying standards, formats, and interfaces for NLP products like MT, TM, and TDBs.

Central to the OTELO concept is the OTELO Lexical Database, which contains entries that can be exchanged using OLIF (OTELO Lexicon Interchange Format). With OLIF, SAP will be able to create and manage terminology in a central OTELO database and easily export its data to Metal or Logos; translators will be able to convert

[2] There are other terminology databases in use at SAP for specific applications. For instance, the Asian Language group has developed its own RDB for its members in Asia; SAP also uses Trados’ Multiterm in conjunction with Translator’s Workbench.

SAPterm/STERM entries to OLIF and load them to the OTELO database where they can be used to automatically update Metal or Logos lexicons.

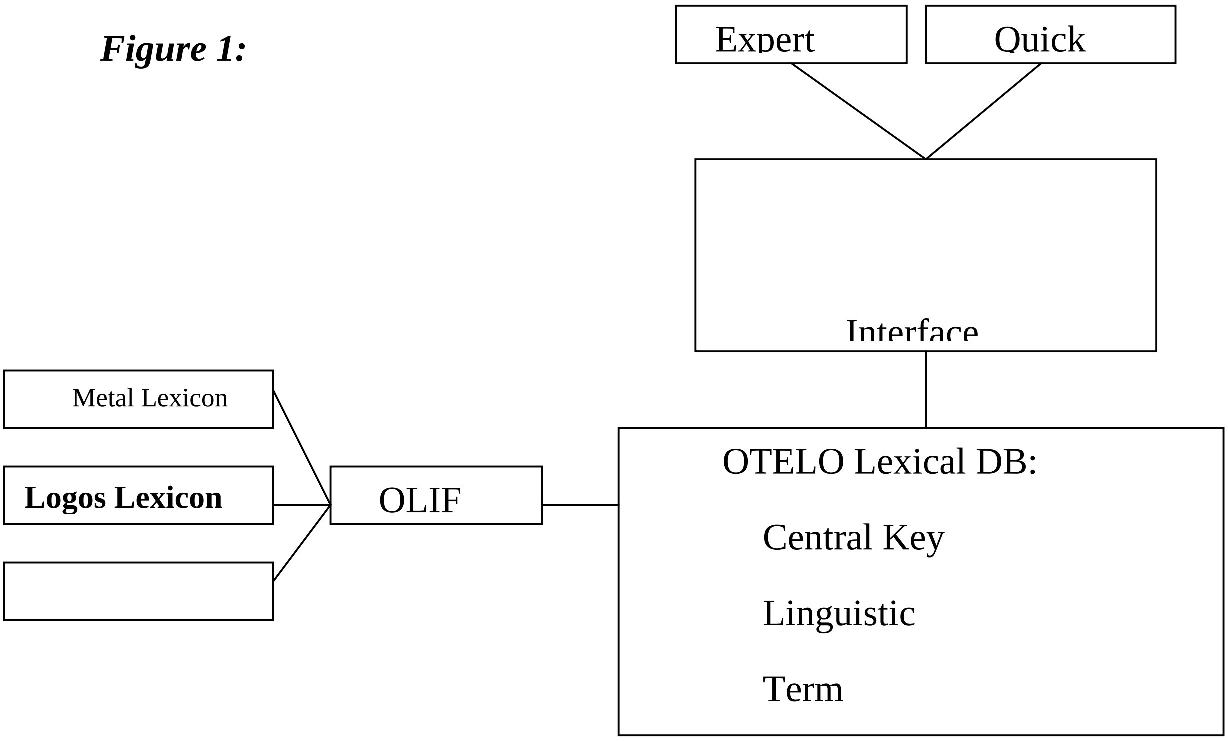
The OTELO central lexical database, designed to accommodate traditional as well as MT lexicography, has four basic partitions associated with a central key:

- Linguistic: semantic, syntactic, morphological information
- Terminology: standard terminological information, such as prose definitions, examples, user comments
- Transfer: information on transfer from language to language
- Cross-Reference: information on word relations, e.g., synonymy, taxonomy, part-whole

Using the expert interface, advanced users may code complete entries, including the specification of selectional restrictions and lexical transformations. For quick, repetitive jobs, a scaled-down quick interface is available.

With the OTELO database in place, translators in the MLT Department should, for the first time, be able to consolidate their terminology in a central location:

Figure 1:



5 Converting to OLIF

To successfully import SAPterm/STERM entries to OTELO, an interchange program must be available that takes into account some fundamental differences between the OTELO and SAP databases in terms of structure, format and defining conventions:

- SAPterm and STERM are concept-oriented; the OTELO lexical database is lemma-oriented.

- Key fields for an OTELO entry are *language, canonical form, subject, area,* and *part of speech;* the *part of speech* field is optional for SAPterm and is, in fact, rarely coded. The feature *gender* is important for MT analysis, but is also sparsely coded in SAPterm. (In STERM, both *part of speech* and *gender* are obligatory, but the current 70,000 SAPterm entries that will be migrated to STERM do not usually contain this information.)
- Conventions for formulating the canonical form are not the same. In OTELO, for example, an *adjective-noun* multiword is entered in the order that is unmarked for the given language, e.g., *adjective noun* for English or German, *noun adjective* for French. In SAPterm, conventions differ from language to language. In German, for example, *adjective-noun* multiwords are entered as *noun, adjective.* In English, they are entered as *adjective noun.* (Conventions for STERM have been drawn up to be compatible with OTELO, but, again, none of the existing SAPterm entries reflect the new conventions.)

Within the OTELO framework, SAP has written an interchange program that converts entries from SAPterm to OLIF. During the conversion process, some simple linguistic algorithms are used to automatically fill in gaps in features like *part of speech* and *gender,* as well as convert canonical form strings to the OTELO standard.

Entries can first be selectively downloaded from SAPterm so that they look a little more like a lemma-orientation than the standard SAP display. The output is a flat file of entries with a small set of lexical features:

Figure 2: Sample SAPterm Entry:

```
<Path>R/2: Vertrieb (RV)


<German>Abbuchen, automatisches
<Creator>FISCHERF
<Creation date>19930817
<Changed By>SATTLER
<Change Date>19930921
<Status>
<Gender>
<Category>
<Unauth.Synonym>Abbuchung, automatische


<English>automatic order filling
<Creator>FISCHERF
<Creation date>19930817
<Changed By>SATTLER
<Change Date>19931015
<Status>
<Gender>
<Category>
```

Figure 2 shows a case in which the German term *automatisches Abbuchen* is identified with the English term *automatic order filling* in the subject area *Vertrieb (Sales and Distribution).* Note that the feature *Category (= part of speech)* has a null value for both the German and English. The feature *Gender* is present for both terms, even though grammatical gender is not relevant to English. It has been left uncoded in the German. Also, as mentioned above, the convention for formulating the canonical form, in this case an *adjective-noun* string, is different in German than in English.

When converted to OLIF, the part-of-speech value (*CAT)* has been assigned, gender *(GD)* has been derived for German based on the morphology of the noun, the feature *gender* has been discarded in English, and the canonical forms have been regularized to adhere to the OTELO conventions:

Figure 3: OLIF Entries

<ENTRY>

<MONO>

<CAN=automatische Abbuchen>

<LG=de>

<CAT=noun>

<SA=RV>

<CE-AUTHOR=FISCHERF>

<CE-DATE=1993-17-08>

<L-AUTHOR=SATTLER>

<L-DATE=1993-21-09>

<TSTAT=new>

<GD=(n)>

<USE=online>

<ETYP=mw>

<NOSYN=automatische Abbuchung>

<SOL=R/2>

</MONO>

<XFR>

<ENTRY>

<MONO>

<CAN=automatic order filling>

<LG=en>

<CAT=noun>

<SA=RV>

<CE-AUTHOR=FISCHERF>

<CE-DATE=1993-17-08>

<L-AUTHOR=SATTLER>

<L-DATE=1993-15-10>

<TSTAT=new>

<ETYP=mw>

<SOL=R/2>

</MONO>

<XFR>

<CAN=automatische Abbuchen>

<LG=de>

<EQ=FULL>

(SA=*subject area;* TSTAT=*technical status;* ETYP=*entry type;* SOL=*industry solution*; EQ=*equivalence;* REV=*reversible*)

The SAP entries in OLIF are represented as monolingual entries (MONO), each with a full-equivalence transfer (XFR) to a target language. In addition to supplying part-of-speech information and regularizing canonical forms, the program has also analyzed the canonical forms for entry type and decided that they are multiwords. All of this derived information will be helpful not only in building the OTELO database, but also for the transition from SAPterm to STERM, since we can easily and automatically fill out information that is missing from SAPterm.

While SAPterm/STERM entries will provide the basis for new OTELO entries, Metal and Logos entries must also be converted to OLIF. For example, the Metal transfer entry in Figure 4 will be represented as well in the OTELO database:

Figure 4: Metal Transfer Entry

"automatische Abbuchen" NST --> "automatic order filling" NST
Pref S.0.0.00 Tag (SAP-RV)
<< Sap SAP Gaston 4-Feb-92 >>

Since the SAPterm entry in Figure 3 and and the Metal entry in Figure 4 refer to nouns with the same canonical forms, in the same languages and subject areas, the entries will be merged into a single entry in the OTELO database. During the merge, linguistic/lexical feature values will be unified where possible and the information that the entry exists both in SAPterm and in Metal will be maintained. Users will thus be able to interface with a single entry instead of managing multiple entries in several different databases.

5 Using the Central Lexical Database

It is clear that a major impetus for creating a common lexical database for SAP terminology is the need to reduce the administrative work involved in keeping several similar databases up-to-date with another. In addition to the advantages already discussed, the centralized approach should further lighten the administrative workload by offering:

- **A unitary treatment of subject-area codes:** At present, the requirements of Metal, Logos, and SAPterm mean that three separate subject-area schemas are maintained for essentially the same subject-area hierarchy. In OTELO, a single schema exists from which the others can be mapped.
- **Simple options for comparing terminology from different sources:** Merging entries in OTELO allows the user to make quick, easy checks for things like discrepant translations, i.e., cases where one system assigns one target translation for a given source word and another system assigns a different one.
- **Facility for making global changes:** Changing or deleting entries based on global criteria is easily done and applied to all relevant databases represented in OTELO.
- **Easy Import/Export of terminological data:** OLIF is an open, SGML-type format (see Thurmair et al. (1998)) to which other common formats can be easily adapted. Its coverage is relatively broad and eclectic, ranging from traditional lexicography and terminology features to the more detailed MT requirements. This alleviates the difficulties often encountered with terminology interchange.

By making it easier to generate SAP entries in different formats, the new approach should also make the startup costs for a new MT system far less onerous. With OLIF and a common lexical database, the development of a new system lexicon should require much less manpower.

In general, SAP sees the move towards integrating its various lexical and terminological information into a single, central source as a viable way of addressing its terminology needs as it expands. If translators have quick access to official terminology, if related translation tools can be brought together to support the central standard, the company will be better able to deliver consistent, clear documentation to its customers.

References

McCormick, S. (1997) "Lexical Resource Integration Requirements for OTELO," WP3.4 Technical Report, EU Project LE-2703.

Ritzke, J. (1997) “Common Lexical Resource Format/CDB Specification:CDB Features and Values,” WPA1.1 Technical Report, EU Project LE-2703.

Steffens, P., editor. (1995) *Machine Translation and the Lexicon*. Springer.

Thurmair, G., J. Ritzke, S. McCormick (1998) "The Open Lexicon Interchange Format OLIF." In *Proceedings TAMA Conference*, Vienna.

Using Automated Translation in a Corporate Setting

Lou Cremers

Company introduction

Océ Technologies is a manufacturer of copiers, printers, plotters, design & engineering equipment and supplies. It has operating companies in 30 countries and is active in 80 countries. Océ employs 17000 people worldwide; 3000 are based at the head office in Venlo, the Netherlands.

Implementation of automated translation

After prototyping an MT system at R&D to demonstrate its feasibility, a commercial MT-system (Logos) was introduced in 1995 for translation of documentation. The main reason was initially to reduce the increasing translation costs. Additional considerations were quality and shorter release cycles of the product documentation.

The introduction of MT required preparation of terminology, adaption of the use of the English language in the manuals, and a modification of the workflow to incorporate MT. It was especially important to establish a consistent and complete terminology database, a basic requirement for MT.

It was a corporate decision that service documentation was to be translated only in four languages handled by the commercial MT system. User documentation is being translated in an increasing number of languages (currently 10 to 15).

Soon after the MT system started 'production' it was combined with a Translation Memory system (XL8) which was replaced early last year by a TM-system more suited to our needs (Trados Workbench). By now, MT has migrated into automated translation: the combination of MT and TM, two complementary technologies.

Problems encountered

A lot of effort had to be put in the integration between MT and TM, as well as in the integration of MT in the documentation workflow. Although MT and TM come as end-user products, quite some tooling had to be developed at all levels, in order for automated translation to work efficiently.

Once the MT/TM combination started translating, several other problems had to be solved: different document formats, sheer size and numbers of documents, lack of functionality in the MT or TM system and ,of course, network and disk space problems. Synchronised management and distribution of terminology for reference and MT purposes appeared to be a problem as well.

The source text quality of the Technical Service Manuals, written in English by Dutch technicians varied from author to author. In order for the MT system to work efficiently, the source text must be grammatical and must follow certain writing guidelines.

The pre-translation of documentation also required changes in the interaction with translation agencies who had to shift their activity from full translation to edit translation. We had to find a way to monitor the efficiency of this new way of working.

Positive factors

On the other hand, the combination of MT and TM could be employed successfully because of factors such as an existing central documentation department. This made it much easier to influence important issues such as the writing process, document structure and formatting and provided the possibility to adapt existing workflow and methods for the benefit of automated translation.

Result

After almost two years of automated translation, Océ now has achieved effective re-use of translations, and made some progress towards improved quality of both source and translated documentation and a distributable terminology database.

Volume problems caused by the number of FrameMaker files to be translated, was reduced by combining the text into a single large .rtf file that is sent out for translation editing. The files are rebuilt upon return.

If time permits, the source text is corrected, but more often problems are still corrected in translation editing.

Translation requests need to be scheduled to allow for terminology work, review of the source text and consideration of the copy editor's workload.

A number of tools have been developed to automate the translation process as much as possible. They vary from C-programs, shell scripts to Word-macro's and serve to improve integration between different programs, perform format conversions, protect text passages etc.

As a consequence of the growing share of software in our products, the release time has shortened. Nevertheless the scheduling of translations has kept pace.

And last but not least: Océ managed to reduce the cost of external translations by about fifty percent on all documentation handled by MT/TM.

Conclusions

Automated translation by MT and TM can be successfully implemented in a corporate setting under certain conditions.

It was important to start with a manageable and scalable pilot project to establish feasibility.

The translation process must be integrated with the documentation authoring and review process. A central documentation department can serve as a coordinating point for translation. It can issue guidelines, keep a terminology database and ensure proper review.

Considerable technological and linguistical expertise is required in the documentation department.

Workflow, Computer Aids and Organisational Issues.
Margaret King

Introduction.

The burden of this article is that since translation services and agencies can vary enormously in the kind of work they do and how they do it, the introduction of electronic documents and the tools that make use of them into the translation process needs to take account of the differences. In particular, the consequent changes in work flow patterns may be very different, ranging from doing little more than offering another possible way of doing things at some point in the translation process to radically changing the way work is divided and tackled.

The ideas put forward here are based on studies carried out for the Translation Services of the European Commission and for the Linguistic Services of the Swiss Federation, as well as on work in progress with the World Intellectual Property Organisation and in the context of a European project, TransRouter. The first three are all concerned with ways of introducing computer aids into existing translation services and with the consequences of doing so, while the last aims at developing a software tool which, on the basis of document characteristics and other constraints, will help the translation manager to decide how best the task of producing a translation should be accomplished - whether, for example, it lends itself to treatment with a translation memory system, whether it can be adequately dealt with by machine translation, whether a particular terminology resource would be helpful, whether it is best dealt with by human translation and so on. The author is however solely responsible for the contents of this article, and nothing said here should be construed as being the official policy of any of the bodies mentioned.

Translation Scenarios.

The article is structured around examination of three different translation scenarios. Once again, while based on real experience, all are over-simplified pictures which, while reflecting more or less faithfully some aspects of their translation work, make no claims to be an accurate representation of any particular organisation.

In each case, a brief description of the scenario is given, followed by some suggestions about where translation technology might be introduced. Finally, in each case, we ask how the introduction of new tools and new ways of working might change the working lives and habits of those involved.

Scenario one: a small homogeneous unit.

As it is.

First, consider a fairly small group of about ten translators. A major part of their work is legal translation, systematically adding to the parallel text versions of a growing body of legislation. Although this work must be accomplished within reasonable delays, they are not subject to extreme urgency. Consistency is of very great importance: not only must new translations be internally consistent in the way they deal with terminology and phraseology, the body of law must be consistent across time.

All members of the group are used to dealing with electronic documents. They work on PCs, have access to a local terminology base which they share with a number of other translation services and also have access to Eurodicautom. They have limited access to dictionaries on CDRom (not all translators have CDRom readers, not all useful dictionaries are available).They do not however use any translators' aids tools other than those associated with a word processor, of which the spelling checker is the most heavily used. Translations are delivered in electronic form. The idea has been mooted that this being so, the group might produce camera ready copy rather than the electronic version being sent on to a publisher.

All members of the group are staff translators. They work in offices on the same floor, conveniently grouped around a central area which contains much of the reference material they need in printed form. It is therefore very easy for them to consult one another, and chance meetings in the central area reinforce ease of contact.

As it might be.

The scenario sketched above suggests the potential utility of several computerised tools. Here we will concentrate on just three.

A preoccupation with consistency in translation suggests that a **translation memory system** could be very useful, especially if the memory contained the body of law which had already been translated.

Such a memory is likely to be large, perhaps very large, and if searching times are not to become unacceptably long, the question arises of how the memory should be structured to avoid this. In the particular context outlined here, the hierarchical nature of the body of law may provide an answer to that question. In other contexts it may not be so straightforward.

Organisational issues also arise: who is to build the memory, how is it to be updated, who will decide what translations are fed into the memory and at what point? Building the memory involves the more or less fastidious task of aligning two texts and checking the results. How time consuming and annoying construction of a translation memory is depends essentially on two factors: how reliable the alignment has to be, and the number of formatting mismatches between the two texts. Two examples from recent experience will help to show how sensitive aligners are to differences of formatting. In a large document, it turned out that a table present in the English version had been omitted from the French version. Not surprisingly, this gravely perturbed the alignment of all the text that followed. But, since the document was long, spotting the disparity between the two texts would not have been all that easy. In the second case, the layout of the title of one version had been done using hard carriage returns, in the other it was one continuous line. This resulted in different segmentations of the two texts, and, again, in false alignments. These kind of problems can become heavily burdensome when a memory is being constructed from existing texts where, naturally enough, page layout and formatting on each version has been done independently by two different people, with the sole consideration of how the text will look on the page in mind.

In some cases, it may be enough to have most of the alignment correct. With our second example above, accepting the alignments given would only have resulted in one segment being missing from the memory, which would not have perturbed overmuch its functioning. In other cases, though, very reliable alignment is needed, and then each alignment must be checked. Although this task does not necessarily require extensive or profound knowledge of the two languages, it is time consuming and not very interesting work. There is an obvious intimate connection between formatting problems and the amount of time needed to check alignment accuracy.

In the context of our first translation scenario, it is likely that very reliable alignment is required, and that the task of creating a memory will be fastidious and time consuming.

Updating the memory may raise additional organisational issues. Most commercial translation memory systems seem to have been designed from the view-point of an individual translator, who can decide for himself what translations to put into the memory when. He may indeed decide to update the memory sentence by sentence as his work proceeds, thus facilitating the task considerably. Once the context is that of a large organisation, mechanisms need to be put into place for validating the translations to be added to the memory and ensuring that the updating is done. We shall return to this issue below in the context of terminology management.

Most translation memory systems come packaged with a **concordancing facility**, which allows the translator to search for all occurrences of a word or phrase in the memory and view them in context along with their translations. This facility is of obvious interest in a context where consistency across translations and across time is important.

With the same need in mind, the group could well benefit from an **electronic archive** independent of the translation memory. The archive would contain not only the body of law but also a substantial amount of related material, and would support a focused search through parallel language versions. For example, the translator could ask for all instances where term is *not* translated in a specific way, or where two or more terms co-occur. This provides an easy access to reference material as well as a tool for checking consistency. And, of course, such an archive would be useful to a considerable wider community of users.

Once again, however, the construction of such an archive raises issues of how it should be structured, created, and maintained.

Consequences on working patterns.

The use of word processors has already radically changed workflow patterns. There was a time when the translators dictated their translations, which were subsequently typed by secretarial staff. With the introduction of PCs secretarial staff have been much reduced, and the translators type themselves. They do not, however, as yet, have to be overly concerned with formatting of text or of page layout. This will come if they are eventually obliged to produce camera ready copy. At that point, either the translators will have to acquire a new set of skills or secretarial staff will have to be increased again in order to cope with the new kind of work being asked of the service.

It is worth labouring this point a little. Although the chain of producing a draft, having it typed, correcting the draft, having it re-typed, repeating the process if necessary until an acceptable final version is produced, then sending the final version to a publisher where it is probably re-typed is clearly wasteful of time and human resources in an electronic age, it is only recently that translator training has begun to include the acquisition of word processing skills. Maximum efficiency is unlikely to be gained by requiring people with one set of skills to exercise a different set for part of their time.

Other effects on working patterns depend critically on the answers to some of the organisational issues raised above. The questions connected to the creation and maintenance of an electronic archive will be left aside, on the grounds that a single small group of translators is very unlikely to be charged with this task. For the rest, let us imagine two extreme situations.

First, imagine that the creation of the translation memory is contracted out, and that maintenance of it is entrusted to what we might call a translation technician who, while not being a translator, is very familiar with translation technology. When a document for translation is received, the technician makes sure that it is in a suitable electronic form and extracts from the central translation memory a memory tailored to this translation. (Most translation memory systems offer facilities for doing this as part of the package). The translator receives the text to be translated and the tailored memory, which he can then modify if he wishes during the translation process. The translator does not modify the central memory in any way. What changes in the translator's life is that instead of using a naked word processor, he has access to translation memory and concordancing software through his word processor. He needs to learn how to interact with this software, but that is all. When the validated version of the translation is ready it is passed back to the translation technician who updates the central memory.

A rather less attractive scenario is to imagine that the translators themselves are responsible at least for updating the memory if not for creating it, and that they have direct access to the central memory for this purpose. Two versions of how this might be done are possible.

Most translation memory systems automatically update the memory during the translation process. In the first scenario above, the translator was not working with the central memory, so his new translations were only going into the tailored memory created for this translation. One might imagine that updating could be accomplished by allowing the translator to work directly with the central memory and taking advantage of the automatic updating facility. The translator would then have to learn how to edit the translation memory in order to correct translations automatically entered (every translator occasionally changes his mind, or realises he has made a mistake), and even so the risk of introducing incorrect translations into the memory increases, as does that of damaging its integrity by accident.

Most of those risks can be avoided if the translator works with a copy of the central memory or learns how to create an independent tailored memory for himself. Updating the central memory is then done only after all revision s have been incorporated into the translation and the validated version produced. This may mean that the translators now also have to learn to use an alignment tool and spend some of their time validating the results of alignment. Alternatively, the translator has to 're-translate' the document in order to make use of the automatic updating facility.

Scenario two: a slightly larger, more disparate group.

As it is:

This group consists of about twenty people, most of whom are freelance and part time. They deal exclusively with highly technical documents, each of which is on average about half a page long. Sentence structures are limited, but coordinations and relative clauses abound. Documents contain a great deal of terminology, much of it new terminology. Between 1,200 and 1,300 such documents arrive for translation each week. Deadlines are known well in advance, but the freelance translators are encouraged to keep up a steady rate of production by being paid piece rate.

Translators choose their own working method. The majority type their translations onto PCs, although a few prefer to dictate and one or two hand-write their work. All work arrives on paper. Those translations which are produced electronically are delivered electronically. Dictated or handwritten translations are sent to a separate typing pool for typing. The same typing pool is responsible for copy typing the original documents and for up-loading them into a central data base, where the translations also are stored. Experimentation with acquisition of the originals in electronic form has recently started, but paper can be expected to be a major medium for some time to come. The group has recently acquired access to the term bank TERMIUM, and is very enthusiastic about its helpfulness. They also have access to a specialised data base containing material of the same document type as that which they treat.

Not all translations are systematically revised. When revision is done, it is usually done by the permanent staff on paper copies. The corrections are then transmitted to the translator who incorporates them himself into the final text, or corrections are made in the typing pool.

Although the majority of the staff are freelance, for security reasons they work on the premises of their employer. The unit is rather overcrowded, with the result that a wide central corridor now houses both the library and a desk with work stations. Contact between staff who are present is so easy as to be almost unavoidable, but given that many are part time, not all staff meet all other staff regularly.

As it might be.

The most obvious and pressing need for this group is for easy access to existing terminology resources and for tools to assist with the acquisition and management of in house terminology.

Some projects in this direction are already under way. It has long been the case that paper dictionaries can be acquired easily and quickly, and this is now being extended to dictionaries on CDRom. As mentioned, access to TERMIUM is available, although not all translators as yet have a direct access from their own screens. Access to the Internet is being prepared, with special bookmark files provided that will lead directly to potentially useful sites. A word of caution is in order here, though: although much terminology is available on the net, it is of very uneven quality. The provision of special bookmark files can be used to steer the translator towards good quality resources, as well as avoiding a lot of unnecessary surfing.

A **local terminology management system**, accessible through the translator's word processor springs to mind as the translator's aid most likely to dramatically facilitate terminology research and management. However, introduction of such a tool, as always, raises technical and organisational issues.

First, the advantage of a terminology resource coupled with a word processor is that the translator can search directly from the source text and can incorporate the solutions he finds directly into his translation. If the source is not in electronic form, then the first of these advantages disappears. Thus, an essential first step is to transform the written source documents into electronic form through scanning or perhaps dictation software. An extra task is thereby added to the work flow, that of checking the results of scanning or dictating. It is also perhaps worth noting that even if the originals are available in electronic form, many translators will prefer also to have a paper copy available, partly because it is easier to keep an overall view of the document from a paper copy than from scrolling on the screen, and partly because it is far more comfortable to move the head around, looking sometimes at the screen, sometimes down at the desk than to look fixedly at the screen all the time, getting backache and headache while doing so. Ergonomic considerations are of great importance in ensuring acceptability of any electronic tool.

The second advantage is only a real advantage if the local terminology resource is rich enough to provide a sufficiently large number of successful hits. The translators estimate that TERMIUM currently provides a solution in better than 50% of all cases. If the local terminology source cannot at least approach the same success rate, the translators will soon go back to consulting TERMIUM, copying down the solution or printing the fiche and typing the solution in.

The underlying problem here, of course, is that of terminology acquisition. An obvious source of terminology is the translations themselves, but it is unrealistic to install an empty terminology resource and ask the translators themselves to enter terms as they work - especially when the translator is being paid piece rate. It seems likely then that another essential preliminary to installing a local system is to invest in the creation of resources to stock it.

As with the translation memory system discussed in the last section, updating and maintaining the local terminology source will raise issues of validation and of organisation. Most terminology management systems allow the translator to update the term bank as he works, and provide tools to make it easy to do so. In an organisation, encouraging translators to update the term bank obviously helps to solve the problem of acquisition of new terms, but at the same time runs the risk of damaging the integrity of the term bank itself. Multiple entries may be created, sometimes justifiably (terms do change over time or over subject matter), sometimes not. Where more than one solution exists, the poorer one may be chosen through inadvertence or ignorance and perpetuated through its existence in the term bank. All this suggests the wisdom of putting into place a validation mechanism, and centralising the maintenance and updating of the term bank. But centralisation once again contributes to making acquisition more difficult.

Essentially, two options exist here. The first is to persuade the translators to collaborate and to make it very easy and quick for them to do so, perhaps by asking them to do no more than to mark new terms with fluo pen on a finished original or translation and pass the marked copies on to the central updating section. The other alternative is to provide those responsible for updating with tools to help identify new terminology from the results of translation.

Although we have mostly concentrated in this section on terminology needs, let us take advantage of the fact that some of the translators in this group still dictate their translations to suggest that their work might be facilitated by the introduction of **voice dictation software**. These softwares have made a great deal of progress in the last few years, and even though training for individual voice patterns is still required in order to obtain the best results, once the training has been done the results can be surprisingly good. But use of dictation software raises another set of issues, this time connected to physical layout of the working place and to ambient noise. The microphones used are quite sensitive, and even with a robust system a ringing telephone can cause some chaos. Space constraints prevent us from going into these issues here, and in any case, office layout and noise conditions are very specific to particular locations. Nonetheless, given that use of dictation softwares promises to have a profound effect on translation work, it is worth being aware that their introduction also brings with it new problems to be solved.

Consequences on working patterns.

As with our previous case, how working patterns are affected depends to a very large extent on answers to some of the management and organisational questions. Once again, let us imagine two alternative situations.

First, let us imagine that every translator now has access to a rich local terminology source through his word processor, as well as to external resources and to the web. These latter can be accessed without having to close the word processor. Care has been taken to ensure that interfaces are easy to use and pleasant to look at, ideally offering the translator a choice of window arrangement, of screen colours and so on. The local terminology source is regularly and rapidly updated centrally with new terminology culled from all the translators' work.

Original documents either arrive in electronic form or are pre-processed into electronic form before the translator receives them. When a translation is finished, an electronic copy of it and of the original is automatically available from a central data base for use in terminology identification or updating of other linguistic resources.

Preparing a translation now becomes mostly a matter of sitting at a screen. There will still be recalcitrant cases where colleagues will be consulted and library searches made, but they will be much less frequent than before. This has some clear advantages: it is quicker to prepare a translation, consistency across translations is enhanced, duplication of work is minimised. But is also has some disadvantages in that social contact between translators is cut down, and the physical effects of spending long periods working on a computer are known to be disagreeable. Awareness of the disadvantages is a first step towards countering them.

The other scenario involves the translator either having to scan originals in himself, or, more probably, receiving the raw results of scanning and having to check and if necessary correct them before he can start work on the translation. He is also expected to feed any new terminology he defines into the local terminology management system, either whilst he is doing the translation or whenever he finishes a translation. The disadvantages here are that the translator is being asked to spend his time on word processing rather than on translation, and that the integrity of the term bank is in danger. On the other hand, making a direct contribution may mean that the translator becomes more aware of how improving the contents of the term bank facilitates later translation work.

Scenario three: one department of a large translation service.

As it is:

This is a department responsible for translation into a large number of languages. Internally, it is organised into smaller groups of about a dozen translators responsible for translation into one specific language, but parallel translations are often required into several languages at the same time. Requests for translation come from a large number of different sources. They are roughly linked in subject matter, but come in a very wide range of document types, ranging over a number of important dimensions. Documents can be very brief, a page or less, to very long, several hundred pages. They can be highly confidential, somewhat sensitive or not sensitive at all. They can be information or publicity material destined for a wide public or extremely technical. They can be legal texts, where every translation has legal force. Translation requests can be one-off, or documents may come in a series of versions with each version differing more or less substantially from the previous version. The urgency with which translations are required can range from immediately (within the next couple of hours) to several months down the line.

Most, but not all, documents arrive in electronic form. All translations are delivered in electronic form. Individual translators choose their own way of working: some dictate, with the translation then being typed within the language group, others work directly onto a PC. Translators have access to EURODICAUTOM and to the Internet. The department makes use of freelances for a substantial amount of translation (over 20%), and typing too is occasionally contracted out by groups which have difficulty in finding or keeping typing staff. The in-house translators all work in the same building. Freelances work from home or in translation agencies and may be geographically located a considerable distance away. Decisions on when and how to make use of freelances are typically made within the group responsible for translation into a specific language, although freelances are centrally recruited and the list of freelances is centrally maintained.

Substantial support is available to the in-house translators. Help with terminology or with specific language problems is available, as is computer support. Some translation technology tools have already been introduced : in-house translators have access to a machine translation system, as do staff outside the translation service, local terminology management systems have been introduced, there is easy access to a large central terminology bank, translation memory systems are being introduced. Individual translators make use of translation technology tools if they wish to do so, and usage varies greatly.

As it might be :

The most striking thing about this scenario is how different it is from the two preceding ones. The work to be done is much more varied, the size of the operation is much greater and translation management is correspondingly much more complex. And herein lies the moral of our tale. If we think about how translation technology can be of help to this group, it is immediately obvious that different tools will be of use in different contexts. It would be impossible to draw up an exhaustive account here, so let us be content with some examples.

Dictation software may be useful to someone who has to translate a short but very urgent document. In the same context, a **translation memory system** will probably not be useful, just because too much time is needed to prepare the document and a tailored memory.

When a document comes in several versions, **document comparison software** may be very useful to spot the differences between the new version and the preceding version : even if the requester is supposed to have marked the differences he may have forgotten to do so or not done so completely. In contrast, document comparison software is pointless when only one version of a document is ever going to exist.

This group sometimes has to deal with very long documents where time does not allow the document to be translated by a single translator. In this context, **tools which support group work** are potentially very useful. Some of the same tools might be useful to translators in different language groups when translations of the same document into different languages are being produced in parallel. They are simply irrelevant to someone who needs to work quickly on his own.

Extensive use of freelances chosen from a centrally maintained list raises questions of **quality control** in a particularly acute way. It is no longer possible to get to know the freelances and learn who can be trusted to produce high quality work without revision when it proves to be impossible to revise all work. Could tools be developed that would help in signalling potential poor quality ?

Even on the basis of this very small sample, it becomes clear that different documents need to be treated in different ways, and that it is a combination of document characteristics and external constraints such as deadlines or particular language combinations that help to determine what the appropriate treatment for a document is. It also becomes clear that translation technology is not only about translation aids, it is also about translation management tools.

Consequences on working patterns.

Consciously deciding that different documents need to be treated in different ways means that workflow itself is document oriented. To illustrate this, let us again take just a couple from the wealth of possible examples.

First, let us imagine that a series of documents (a translation dossier, let us call it) falls into a category where machine translation has been known to give good results, perhaps because effort has been invested in specialising the machine translation system for this type of document, and where each document in the dossier re-uses material from other documents in the dossier. Let us also imagine that a large central translation memory already exists.

A possible way of treating this dossier would be for someone - preferably not the translator, as already discussed above - first to check for possible formatting problems, spell-check the originals as a precaution, and extract a tailored translation memory from the central memory. The same person could check in a central electronic archive for missing pertinent reference material and extract it from the archive.

The translator would then receive a packet containing documents to be translated, a tailored translation memory and reference material. Choosing the first document to be treated (using as criteria characteristics such as a high degree of internal repetitiveness or that substantial elements of the document are likely to be repeated in subsequent documents), the translator first asks for a machine translation and checks the results. Satisfactory results are used to feed the tailored translation memory, and the rest of the document translated using the memory interactively, that is the translator updates the memory as he moves through the text. This operation is repeated for each document in the dossier. When the translation is complete, it is sent for revision. Revisions are incorporated into the text, and the final text used to update the main translation memory. With his example, it is not strictly necessary that all documents arrive for translation at the same time, it is sufficient to recognise that a document is probably the first of a series.

With this first example, translation is still a black box to the outside world : the individual translator's way of working has changed considerably, but for the requester he still sends a document for translation and gets a translation back, without being aware of how the translation has been produced.

If the document to be dealt with is a very long document, multi-authored with different parts of the document becoming available at different times, one might imagine a workflow that would break open the black box. Here, as soon as it is known that the document will be produced, a team of translators is assigned as a sort of task force to the development and translation of the document. They can be consulted by the authors as work progresses (for terminology or language questions for example - some of the authors may not be writing in their own language) and can start work on producing the translation of different parts of the document as they come available. The author/translator team uses e-mail and groupware to support collaborative work. The translation of different parts of the document is tackled as each translator sees fit.

The main point here, and one which deserves to be laboured much more than space permits, is that workflow does not just a concern the translation activity alone, but the whole context in which that activity is done.

Conclusion.

The main conclusion to be drawn from all this is quite simple. Some translation technology tools and some resources are likely to be of benefit to almost everybody. It is hard to imagine, for example, a translator who could not benefit from **a uniform interface through which a wide range of different dictionaries could be consulted**, instead of having to familiarise himself with a new interface every time he buys a new CDRom, or a translator who does not rejoice at having a rich terminology source readily available and easy to consult. Other translation technology tools are best suited to specific situations and to different contexts of work. A critical factor in deciding on the potential utility of the tool is the type or types of documents to be dealt with and the external constraints on translation production. Maximum benefit from introducing translation technology can be gained by careful preliminary analysis of what is really needed and of the consequences of introducing it.

A secondary, but nonetheless very important, conclusion is that introducing translation technology into the workflow introduces also new tasks, which are in many cases not best tackled by translators. In several places in this article, we have imagined the existence of translation technicians as support staff : it may be that a new profession needs to be born.

References.

EAGLES Evaluation Group (1995). Final report. Available from CST, Njalsgade 80, DK 2300 Copenhagen; or electronically at
http ://wwwissco.unige.ch/projects/ewg96/ewg96.html

King, M. (1995). The European Commission Translation Service : A Case Study. Available from the author or electronically as an appendix of the report cited above.

Workflow Automation of Translation Projects

Paul Kaeser

People in many different sectors of the economy are talking about workflow systems, although the subject has only been discussed occasionally in the field of technical documentation. Concrete applications and solutions have been rare thus far. The following case study illustrates the possibilities and the, as yet unexploited, potential.

The preparation of technical documentation, particularly when publishing in several languages or on a large scale, requires the full cooperation of several specialists. The interfaces between these specialists are still handled manually for the most part, thus causing losses at the interfaces which diminish overall performance.

One of the central themes for the STAR Group's software development departments is to achieve the transparency of processes in the field of documentation through standardization and automation of the interfaces.

STAR is well-known to many as the producer of the TRANSIT translation tool, but in addition to this, STAR is working on a series of standard and specialized software products, all of which aim to optimize the entire process of technical documentation up to and including multilingual and multimedia publication. SGML has an important role to play here.

Exploiting the efficiency of the specialized standard tools while simultaneously controlling the subprocesses through a host control system and total integration were additional challenges.

In most cases, preparing and publishing technical documentation involves intensive team work. For this reason, technical documentation is fundamentally ideally suited to workflow solutions and these solutions generally become a requirement as a consequence of the range of disciplines which need to be handled (writing and editing, graphics, translation into a series of different languages, multilingual DTP etc). In practice, however, they are also required because of the large volume of the documentation, tight deadlines and the variety of versions required.

Commercially available applications, specifically in the form of authoring or translation tools, are usually only aimed at enhancing the productivity of a single job or process and until now, too little attention has been paid to the harmonious and conflict-free cooperation between working groups, or in modern jargon, to groupware solutions.

In the past few years, some specialized applications have made it possible to improve productivity markedly, but it is for this precise reason that the costs generated at the interfaces have risen sharply as a percentage of the whole, which is why they have become the focus of much interest. Added to this is the fact that, in many cases, team members now often work in several different locations rather than just in one place and may even be spread over several different countries.

The automation of the translation process

The possibilities which already exist will become clear from the optimization of the translation and foreign language publication processes using a purpose-built translation workflow system. The potential to be extracted in this situation is less to do with the translation itself and much more to do with handling and interfaces in the translation environment. The translation itself will already have been substantially optimized using TRANSIT, a commercially available standard translation tool. The translation environment demands a certain amount of administrative and data processing work and these activities can, in certain cases, exceed the actual effort demanded by the translation itself.

Although the translation process and the related handling processes are of primary interest at this stage, this subprocess must be incorporated into the documentation process in a meaningful and conflict-free manner. Upstream and downstream processes must be stable, ie they must handle the appropriate input and output interfaces reliably, if a subprocess is to be automated.

The translation process starts with the author

The truism which states that the method used to prepare documentation affects the efficiency of subsequent processes has by now become sufficiently well-known within the documentation community (although whether this knowledge acted on in every case is another question).

Apart from the linguistic and terminological requirements on documentation to make it suitable for translation, elements such as standardized document structures and correctness in formatting also play a part in the automation of the translation process. For example, tables and indented paragraphs created with the tab key are frequently found, even today.

Put another way, documents should be structured in such a way that as few functions are used as possible, better still no functions at all. Texts outside automatic pagination are graphically controlled using paragraph formats. In this way, a translated text can subsequently be formatted automatically by the DTP system with practically no intervention. At the same time, this procedure also has the side-effect of significantly increasing the ease with which the source text can be maintained.

Translation Memory Systems

STAR began development of the TRANSIT Translation System in the mid eighties. The concept of an integrated translation environment, and in particular the principle of Translation Memory, was barely taken seriously at all in the industry. Today there are virtually no medium or large-scale translation projects planned without the use of this technology.

A first step towards the integration of TRANSIT into the documentation process was the development of interfaces to all popular DTP systems (Interleaf, Framemaker, MS-Word, Quark Xpress, Pagemaker etc.) and to data formats (SGML, HTML etc).

These developments have since become indispensable parts of the daily routine in the industry.

The next step towards integration is extensive automation of processes, so that no more time is wasted on pre-editing and post-editing work. This has been achieved by making TRANSIT fully capable of integration and by adding server capabilities to it. This means that TRANSIT can take commands from other (control) applications, process them and send the results back to the control application.

The specialized work, the translation-relevant actions, are handled by the TRANSIT standard application.

The concept of the Translation Workflow System developed by explained below.

Mode of Operation

A daemon on the server system (MS-Windows NT 4.0 in this concrete case) continuously monitors specified directories. As soon as an authoring system saves data to one of these monitored directories, a job file also supplied by the authoring system is read. The information a human “contractor” would also require to complete a job is stored in the job file. The job is processed by the system fully automatically.

The workflow system now automatically passes on appropriate commands to the TRANSIT Translation Memory System which is also installed on the server. These are, in particularly, the following:

- Conversion, segmentation and automatic pretranslation using data identified by the workflow system as the most suitable reference material. Creation of control data for the translator’s client computer.

- Once the data has been prepared for the translator in this way, it is compressed into an archive file. The workflow system now searches in the database for the translator best suited for this job. Criteria include, for example, type and volume of documentation, language combination and the translator’s current availability. The archive file is now sent to the translator via FTP (with the options of network, modem ISDN or Internet).

- TRANSIT Light (the translator version of TRANSIT) is installed at the client (translator) side along with a workflow system client. The client reads the data from the server, installs the translation job automatically and informs the translator on screen.

- The translator now only has to click on the job-number and TRANSIT loads itself automatically with the parameters required for this particular job and the information to be translated. The translator now “only” has to translate and check his or her translation.

The return route is similar, once again fully automated, as soon as the translator approves the translation for dispatch.

The workflow system on the server now sends the same data to one or more revisers (depending on the specified process, eg language reviser, technical reviser). Certain automatic checks (eg consistency, code page, SGML/HTML structure, translation completeness etc) can be incorporated at this point before the translation is returned to the authoring system or customer. These checks enable the authoring system to paginate the foreign language data directly.

All handling, both at the server and client end, is fully automatic. The system tracks each job until it is returned, sets and monitors deadlines, sends reminders for missed deadlines etc. All actions and translation volumes are continually logged recording precisely who did or approved what and when, an important element in QA systems.

The information can be exported in Excel format at the end of the month or at other times, sorted by translator and/or customer. This means that invoicing can also be automatic, for instance. Messages are sent to the system administrator in the event of any problems, allowing for appropriate intervention, or discussion with the translator concerned.

All translations completed using this system are automatically added to the reference pool (ie Translation Memory management is automatic) so that, once completed, translations can be automatically reused by workflow server at any time.

The ability to integrate as an important criterion for tool evaluation

The suitability of the tools to be integrated is critical in ensuring that a system of this type can function correctly.

The capability of integration ought to become the principal criterion in the evaluation of such tools for professional environments.

When would a workflow environment make sense? In principle, it makes sense to automate any time the stability of a complete process or authorship can be influenced. This is in the interest of overall productivity and is not, therefore, merely an end in itself. In practice, there are two typical scenarios which demonstrate the system's potential particularly clearly:

- A large volume of small documents which need to be regularly translated into or out of several languages to a deadline (eg service bulletins)

- Larger scale documentation which is prepared and maintained in modular fashion (eg systems documentation). This makes "Simship" (the simultaneous publication of all language versions) a relatively stress-free possibility for the translator, even when publications are on a particularly large-scale. This case is typical for CD-ROM or website publication.

The objectives are identical in both these cases:

- Reduction of costs by avoiding administrative and handling overheads

- Shorter throughput times

- Prevention of errors by cutting out the human factor in routine procedures

- Reliable monitoring of the status and quality of work being carried out

What are the prospects for a wider distribution of the integrated publication chain? A strong trend towards publishing in several formats simultaneously can already be detected. At least five different media are current today: paper, microfilm, online help, CD and the Web.

The PARS MT Family: Practical Usage

Michael S. Blekhman (Paper not presented)

PARS is a family of machine translation systems developed by Lingvistica ’93 Co. for the following language pairs:

- Russian to and from English, German, and Ukrainian;
- Ukrainian to and from English and German.

The systems run on IBM PCs and have the following characteristics:

- they run under Windows 3.1 and above, Windows 95, Windows NT, in stand-alone and network modes; DOS versions are available for PARS/ER, PARS/U, and PARS/RU;
- PARSes are designed to make draft translations of scientific, technical, business, and socio-political texts in the subject areas covered by their dictionaries; in particular, PARS/ER bidirectional dictionaries relate to such subject areas as business, mathematics, physics, chemistry, aviation, engineering, automobile building, oil/gas, etc., altogether over 900,000 in each part, English-Russian and Russian-English; **a very large general dictionary of about 300,000 words and idioms** will be made available to the customers later this year;
- the translation modes are «from file to file», «from Clipboard to Clipboard», and «from WinWord to WinWord»; in the latter mode, PARS is started directly from WinWord 6,0, WinWord 7.0, or WinWord 97, and the target file is placed in a separate Word window under the source one, preserving the source text formatting; in the «Clipboard to Clipboard» mode, PARS can translate HTML files and Windows screen Helps, as well as files generated by all Windows-based text processors such as Word Perfect, Write, etc;
- polysemantic words and phrases are marked with asterisks in the target file so that the user could select a more appropriate translation from the panel of optional translations;
- each system has a flexible dictionary editing program;
- a special dictionary compilation technology is used to develop new dictionaries; one of the major sources of terminology is a set of Polyglossum dictionaries supplied by our partners, ETS Publishers, Moscow, Russia.

Practical usage

I made a kind of classification to outline the circle of PARS users. Unfortunately, any kind of serious statistics is impossible due to awful computer piracy in ex-Union. The only thing I can be «proud of» is that PARS as well as Polyglossum are very popular with the pirates: Igor Fagradiants, director of ETS Publishers, claims that about 300,000 piratic compact disks with our systems have been made in Russia since 1996 up to the present time.

Individual users

A very numerous subgroup is made up by *students who need their diplomas and other kinds of papers to be translated from Russian into Ukrainian.* We hope to meet their requirements with the COPERNIC CD-ROM and convince at least some of them to abstain from using piratic disks. COPERNIC is a project launched last year jointly by the Ukrainian Ministry of Education and Lingvistica ’93 Co. The disk comprises the basic versions of each of the 5 PARSes (without specialist dictionaries), it is supplied with a user’s guide, and costs $12 for school, college, and university students, and $28 for the rest of customers.

Some people want to communicate with people living abroad. PARS/U, is bought, in particular, by Americans and Canadians wishing to communicate with their friends and relatives residing in Ukraine. One of them told me: «They speak Ukrainian, while I speak English. The only way to communicate is to use a computer program». I wonder if one of the international pen pal organizations might be interested in using PARSes for communication purposes. It would certainly require serious modifications to the systems in order to take into account peculiarities of this style, but the idea itself seems rather promising to me.

Professional free-lance translators make up another subgroup, though less numerous. Their language pairs are mainly English, German, French, Italian to and from Russian. Some of them like MT systems, some prefer MAT software (electronic dictionaries such as Polyglossum), while others buy both. My opinion is, however, that the majority of this group are still our **potential** clients. The fact is that the foreign languages departments of Ukrainian universities train people who are good at languages but have no idea of the computer as translator’s everyday tool. Introducing elements of language engineering at such departments would contribute a lot to expanding the circle of our conscientious clients!

There is a group of *individual users who require Russian to English translation of scientific texts.* Here is an example. A scientist asked me to translate his medical paper for submittance to a serious British journal. When I looked through the text, there was only one thing which I understood - I could not do without PARS because the paper was abundant in «awful» medical terms. I faced a dilemma: either to translate the text manually looking every second or third word up in the Polyglossum Russian-English medical dictionary, or to let PARS make a draft translation and post-edit it. I chose the latter variant, and the paper was accepted.

Corporate users

MT and MAT systems seem to be very popular with corporate users. Generally speaking, all kinds of *organizations, both state-owned and private, use PARS/RU for translating official documentation, including that of financial, scientific, and technical nature, between Russian and Ukrainian.*

Many *Ukrainian banks use PARS/RU for translating financial documentation, such as official instructions, between Russian and Ukrainian.* Here is another example. In autumn, I installed PARS/RU in one of the banks in the town of Saki, the Crimea. They use it to translate megabytes of instructions they receive electronically from the Ukrainian National Bank. Those texts are written in Ukrainian, the country's state language, and the problem is that many people in the Southern and Eastern parts of Ukraine doesn't even **understand** Ukrainian, to say nothing of speaking it.

A tendency that gains popularity is making *MT systems part of integrated products*, such as PRAVO, a system very well-known in Ukraine. It is supplied on CD-ROM and comprises the full set of Ukrainian laws and decrees, with a retrieval system and our Ukrainian-Russian translation module. Later this year, Ukrainian to English and German modules will be added in May.

I am especially proud that PARS/ER is used for translating Russian medical abstracts into English for the *Medical Practice* journal published in Kharkov. I do it myself, first running the texts through PARS and then post-editing the raw translations. *Using MT systems for translating abstracts in scientific journals* may become a tendency.

Large plants and design bureaus that export their products are among the users of the PARS/ER system. The Antonov Aviation Design Bureau in Kiev as well as the Yangel Spacecraft Bureau in Dnepropetrovsk are among them. We supplied PARS/Avia to them, which includes the core Russian-English-Russian system and a number of terminological dictionaries on aviation, space, communications, etc. Their reaction is very important for me: they say that PARS is better for translating technical documentation, while Stylus by ProMT is preferable for business correspondence. Well, we'll try to be up to the mark in all the aspects!

A new tendency is using PARS *to translate Russian textbooks and lectures into English for foreign students coming to sudy at our universities* (see below).

MT can and should also be used for purely academic purposes. An example is using PARSes at Kharkov State Polytechnical University in the course of machine translation at the Department of Intelligent Information Systems. Presently, we are going to set up a department of language engineering at Kharkov Slavonic University. I plan to implement all our systems there.

Access to Internet and E-mail will contribute to a higher role of MT. However, this will require not only technical (which is comparatively simple) but also linguistic solutions because colloqial texts, which are very often to be found on web sites, to say nothing of E-mail messages, are very hard to translate automatically. I am sure that Internet and MT will stimulate each other greatly. And this application is very promising. The fact is that Internet resources are in fact unaccessible to Russian and Ukrainian speaking scientists because of the language barrier, and so are the Russian and Ukrainian publications for the English-speaking community. You should take into account that the state system of scientific information, which was the pride of the former Soviet Union, does not exist in Ukraine for a number of reasons, so Internet will be a very good, though not the only source of information if the decision will be taken to build up such a system in this country.

In 1996-1997 Olga Bezhanova, my elder daughter, used PARS to produce draft translations of large scientific and technical text corpora. She described her experience elsewhere (MT News International, Proceedings of the MT Summit VI). In this paper, I would like to summarize the results she obtained.

Post-editing PARS-made scientific translations

This work was ordered by The Russian Foundation for Fundamental Research (RFFR). Olga was supposed to translate the «RFFR Annual Bulletin» that comprised about 400 pages presented in the WinWord 6.0 format (about 1.2 MB), namely the titles and bibliographic data for several thousand research projects in the following areas:

- mathematics and information science;
- physics and astronomy;
- chemistry;
- biology and medicine;
- geosciences;
- liberal arts;
- databases and books issued in Russia.

Each document in the Directory comprised approximately 5,500 pieces of information, each consisting of the author’s surname, project title, identification number, name of the institution (University, research institute, etc.), and the city/area of residence.

Thus the task generally consisted in translating not complete texts but the titles of research projects, **each title having two to fourty words**.

The customers required a similar English text presrving the source text styles and formatting. The customers also stipulated that **the surnames were to be transliterated according to the rules of the English language**, while the titles of institutions were to be translated.

The work was supposed to be done within approximately a month. Taking into account the days-off, the translation was made during 34 days, 5-7 hours a day, consisting in **post-editing the texts translated by PARS**. Numerous misprints in the sorce text (the better half of which composed Latin letters instead of Cyrillic ones in Russian words) slowed down the whole process.

A great number of scientific terms in the source text relating to numerous subject areas required using quite a number of various dictionaries. Olga says that **translating texts of such volume by one person for such a short period of time without using machine translation software would be impossible**.

Before the translation session, the Polyglossum system of dictionaries was activated on CD-ROM, which made it possible to access any of the dictionaries without exiting from WinWord.

Words not found by PARS were looked up in Polyglossum dictionaries.

When Olga was making the translation, PARS lacked several terminological dictionaries that were entered into the system later. They would have increased MT quality greatly. The names of these dictionaries are given for the corresponding subject areas.

Here are the results for each subject area.

Mathematics and information science

This chapter of the Directory comprised 800 titles of research projects in the above subject area.

In order to translate this chapter, the following dictionaries were set up in PARS (in a descending order of priorities): computer dictionary (25,000 terms in each part, Russian-English and English-Russian), technical (76,000 terms), general (35,000 words and phrases).

The mathematics dictionary (85,000 terms) was made later.

It is to be noted that the dictionaries used did not cover the subject area completely, so Olga had to refer to the Polyglossum dictionaries when post-editing the documents on mathematics. As to those on information science, they were translated by PARS fairly well.

The main difficulty when working with this chapter consisted in translating phrases comprising surnames of «foreign» mathematicians, as, for example, *Langevin equation*. Because many similar phrases were absent both in PARS and in Polyglossum, Olga had to look them up in The Great Soviet Encyclopaedia, which presents names of well-known scientists in their native languages.

Physics, astronomy

The following dictionaries were set up in PARS for translating this chapter (consisting of 1290 titles): technical, radioelectronics (50,000 terms), microelectronics (20,000), general.

Due to the absence of a special dictionary on physics and astronomy in PARS (the dictionary on physics, 80,000 terms, was developed later), post-editing this chapter was more difficult. The Polyglossum dictionaries, comprising about 1,500,000 terms of various subject areas, were of great help.

The main problems arose in rendering the names of the planets and their satellites, which Olga managed to find in the English-Russian astronomy paper dictionary.

Chemistry

For translating this chapter (659 titles), the technical and general dictionaries were set up in PARS. The chemical dictionary (50,000 terms) was not yet present in the system.

This chapter was the most difficult to translate since it comprised quite a lot of specific chemical terms, such as *фталоцианин (phthalocyanine), редокс (oxidation-reduction), рацемат (racemoid), аценафтен (acenaphthene), гваяцил (guaiacyl),* etc.

Olga had to look up the words not found either by PARS or by Polyglossum in the Russian-English Dictionary of Chemical Reactions and in the English-Russian Dictionary of Petroleum Chemistry and Processing because, generally, the difficulty consisted in the spelling of the unknown chemical terms.

For example, it was clear that the English translation of the term «стирил» could not differ seriously from the Russian variant, but the translator was not sure whether it was «styril» or «stiril». She found the word «styryl» in one of the paper dictionaries, which put an end to the troubles.

It was very difficult to translate complex terms consisting of several components, for example, «винилхалькогенополигалогенбензол». PARS failed to translate such words, that is why it took Olga 6 days to post-edit this comparatively short chapter.

Coming across a word consisting of several components, the translator usually broke it in sense-bearing parts and translated them in turns. Thus, the term

винилхалькогенополигалогенбензол

was broken into *винил, халькоген, полигалоген*, and *бензол*. The resulting «simple» words were translated and united in one. It's only natural that such work was very labor-intensive and occupied much time.

Biology, medicine

This chapter (908 titles) was the second most difficult to post-edit. The following PARS dictionaries were used for translation: medicine (20,000 terms), aviation medicine (24,000 terms), technical, general.

The chemical dictionary as well as that on biotechnology (10,000 terms) was entered into PARS later.

The main difficulty consisted in translating the names and genders of animals and insects. Again, Olga could not do without the Russian-English paper dictionary by A.I. Smirnitski, in which she found such terms as «иглокожие», «ракообразные», - «echinodermata», «crustacea», etc. The dictionary comprises quite a number of biological terms.

Geosciences

This chapter comprised 752 projects in such subject areas as geology, paleontology, archaeology, ecology, etc. PARS translated this chapter very well, owing to the presence of geological and ecological dictionaries; this raised the translation quality several times as compared with translating such chapters as «Chemistry» and «Biology, Medicine».

The chapter was translated in two stages. It was split into two portions of nearly equal sizes that were translated by PARS using the following dictionaries:

The first portion: geological (11,000 terms), technical, general.

Embarking on the translation of this chapter, Olga didn't yet know that it comprised many documents on ecology, that is why the ecological dictionary was not chosen for translating the first portion. When post-editing it, she saw that the better half of the words not found in the system dictionaries related to ecology, so she also set up the PARS ecological dictionary (18,000 terms) for translating the second portion.

The second portion: technical, geological, ecological, general.

By the way the dictionary on oil and gas (70,000 terms) was developed later. More than that, the geological dictionary was extended to comprise 27,000 terms.

When post-editing this chapter, Olga actively used The Random House Unabridged Dictionary to clear up the spelling of such words as *Тетис (Tethys), пегматит (pegmatite)*, and many others. In particular, she made use of the table of geological periods given in this dictionary. It is to be noted that this was the only source where she managed to find translations of a large number of geological terms. It only took her 4 days to translate this chapter, which would have been impossible without using the Random House dictionary.

Humanities and social sciences

This chapter (214 titles) was the easiest to translate. Two PARS dictionaries were used: general and economy (55,000 terms).

Despite the fact that this chapter occasionally comprised separate terms of biology, geology and ecology, the number of words not found by PARS was very small.

«Databases of the 95-96ies»

This chapter was translated by PARS very well using the general and computer dictionaries. Post-editing consisted in making minor corrections to the machine translation.

Generally speaking, the draft translations provided by PARS were of different quality, depending on the source text subject area and, accordingly, on the presence of terminological dictionaries.

A few translations required no post-editing at all or «cosmetic» post-editing. For example:

151. Офицеров В.И. Исследование структурной организации конформационных антигенных детерминант на примере белков оболочки вируса гепатита А.

Machine translation:

151. Ofitserov V.I. Research of the structural organization of conformational antigenic determination on the example of the proteins of the capsule of hepatitis virus A.

On the other hand, in some cases the translation offered by PARS had to be corrected completely to obtain the correct text. The fact is that if the system dictionary doesn't have a set expression, PARS translates it word for word, sometimes making it hard to understand. For example, the phrase «принимающий решения» was translated «receiving decisions», on analogy with «receiving letters». The phrase «образ жизни» was rendered as «image of life» instead of «way of life», etc.

One of the merits of the PARS system is the translation variants option. Here is an example.

580. Носиков В.В. Поиск и изучение антигенных детерминант, связанных с аутоиммунной деструкцией островковых бета-клеток при инсулинозависимом

сахарном диабете, с сипользованием библиотек бактериофагов, экспрессирующих широкий спектр разнообразных пептидных эпитопов.

Machine translation:

580. Nosikov V.V. Search and the studies* of antigenic determinants bound with the autoimmunity disruption of islet beta-cages* at инсулинозависимом sugar diabetes, using the libraries of bacteriophages expressing the broad* spectrum of diversified* peptide epitopes.

A double click on the asterisk will display the list of translation options for this word/phrase. Having chosen one of the variants and pressed the button, the post-editor inserts it into the text instead of the initial one.

In the above text, the following translation variants were offered: studies (research, analysis), cages (cells), broad (wide, capacious, extensive, large-scale), diversified (miscellaneous, diverse). The most important substitution is cerainly ‘cells’ instead of ‘cages’.

Transliterations are presented as translation variants for proper names, which is very useful when post-editing machine-made translations of bibliographic documents. For example, in the above text, *Nosikov V.V.* was substituted for *Носиков В.В.*

6.1. *Post-editing PARS-made technical translations*

The order from Kharkov Aviation University consisted in translating from Russian into English a large corpus of technical texts on aircraft building for Iranian students who were coming to sudy at the University. 650 K of source Russian texts were presented as WinWord and DOS files. They were translated by PARS and then post-edited. Both PARS for Windows and PARS for DOS were used, with a number of terminological dictionaries.

The translator had very little time for doing the work. An original Russian text of 20-30 pages used to be given every day at noon, the request being to «please translate it not later than tomorrow morning!».

The situation being so tense, sacrifices as to the stylistic purity of the end-translations had to be made in order to submit them as soon as possible. The translator’s was to make the texts grammatically correct and understandable, omitting a number of stylistical details, such as repetition of several «of»-clauses, misuse of articles in those cases where this did not affect understanding, etc.

This allowed Olga to come up with translations that were grammatically and lexically correct though stylistically far from ideal in a number of cases. Here are two examples of machine translations left as they were, without any post-editing:

«Requirements to the execution of the outlines of modern airplanes and assurance of inter-changeability of their aggregations».

«This advantage is especially noticeable on the larger level of loading».

Again, technical terms were the main difficulty, but the problem was solved by using Polyglossum were PARS failed to translate a term: the time needed to find in

Polyglossum a word not translated by PARS and paste it into the text is, as a mathematician would say, «negligibly little». Besides, all the «new» terms were immediately entered into the corresponding PARS dictionaries by Lingvistica '93 dictionary officers, which made PARS «cleverer» with each text translated.

At the same time, there were also some sentences generated by PARS that had to be changed completely, as, for example, the following one:

«After switching-on pumping station, if via 5 with pressure is not is heaved above 8 Pa actuates signaling table ...ABORT».

If all or most of the sentences had been translated so poorly, post-editing would have been much more difficult, which would have made MT quite or almost useless. However, it only took the translator about **2-2.5 hours to post-edit 30 K of texts using PARS, and, what is very important, the work itself was not so boring and tiresome as manual translation**. In other words, **editing machine-made translations was 3-4 times easier than translating the same texts manually**.

Using PARS (plus Polyglossum, if necessary), the translator can prepare 20-30 pages a day and not feel exhausted.

MULTIDOC – Controlling Language in multilingual Documentation

Jörg Schütz

1. Introduction

MULTIDOC is a European project of the Fourth Framework Programme within the Language Engineering Sector. It is founded on the specific needs and requirements of product documentation expressed by several representatives of the European automotive industry, among them are Bertone, BMW, Jaguar, Renault, Rolls-Royce Motor Cars, Rover, Volvo and others. The focus of the project is particularly on the multilingual aspects of product documentation. Therefore, the general goal is to define and specify methods, tools and workflows supporting stronger demands on quality, consistency and clarity in the technical information, and shorter lead times and reduced costs in the whole production value cycle of documentation including the translation into multiple languages.

The results of the project are applicable to any other component or system manufacturing business; thus, they are not restricted to the automotive industry. The project is divided into two phases: an inception and elaboration phase, the so-called MULTIDOC Concerted Action (LE3-4230), and a construction or development phase, the so-called MULTIDOC Project (LE4-8323). The first phase has been finished by the end of last year, and the second phase has started in January 1998.

2. Basic Requirements and Vision

The aim of the MULTIDOC Concerted Action was to identify the problem areas and to specify solutions for the European automotive industry when it comes to multilingual product documentation and also set a roadmap for the future. In software engineering, this phase is usually called the inception phase of an iterative software development process. During inception we establish the business rationale for the project and decide on the scope of the project. This is also the phase where we get the commitment from the project sponsor(s) to go further; in our case this was the successful evaluation of our MULTIDOC Project proposal.

The Concerted Action also included parts of the elaboration phase of a software development project. In elaboration, we collect more detailed requirements, do high-level analysis and design to establish a baseline architecture, and create the plan for construction which is the actual software production phase consisting of many iterations. In our domain, the most crucial bottlenecks comprise the following business areas that needed further elaboration:

- More and more languages in which product documentation has to be published; there is a drastically increased focus on Asian and East-European markets.
- Increasing costs for translations.
- Lead times in the document production process and in the translation process.
- Poor or no possibility to measure and control the translation process.
- Inconsistent use of information structure and information content.

All project partners agree that besides the quality of the product the services associated with the product and the accompanying documentation of the product must be seen as an integral part of the product. To satisfy the demand for high-quality technical documentation, the documentation has not only to be comprehensible and up to date, it has to be produced and delivered (including the accessibility to new or up-dated information) with modern technologies. The following scenario shall exemplify the intended direction:

> *Mr M is the proud owner of a new, environmentally clean car which was assembled according to his wish list from a huge variety of car components of the automotive manufacturer. The first contact with his new car was well before the actual delivery in a virtual reality animation, where Mr M was able to check his colour selection, the harmonisation of the chosen colour with the selected materials of the car interior, as well as first virtual driving tests.*
>
> *Mr M is also very satisfied with the delivered car documentation; he got the personalised documentation right after the deal was contracted. This documentation is not only personalised but also customised to his specific car: this includes the appropriate colouring of all graphics in the documentation and the text itself, where we do not find any generalised references such as "... applies to specific countries.", "... according to the model variant." and so forth.*
>
> *With the CD-ROM edition of the documentation Mr M can directly search for information on his PC at home, and he is also provided with multimedia enhanced information about his car. This documentation is also available in his car via the on-board computer. This computer maintains each service work and possible repair work in its storage, and therefore permits customised service and repair measures and fault tracing at dealer's workshops. Not to mention, that these data are also available via the world-wide computer network of the car manufacturer.*
>
> *Mr M gets new information about his car and about new products and services of the car manufacturer via his Internet access at home; for this he has subscribed to the free information service which in addition is parametrisable according to his information demands.*
>
> *Now, on a business trip to the south of Spain Mr M has to stop with a defect late at night. The workshop that is alarmed due to the 24-hours assistance service is somehow lost: this kind of defect is not listed in the technical service documentation and therefore not a standard service and repair routine. However, the immediate access to the hotline information service via the computer network identifies the same defect two days ago in Oregon/USA. The problem solution that the Spanish mechanic gets on his workshop screen is in English. So he activates the translation-on-demand button and receives a Spanish translation within a few seconds. This translation is not perfect, but the necessary repair steps and the terminology of the needed tools and parts is correct due to a multilingual terminology knowledge base which is maintained by the car manufacturer; therefore the mechanic does not care about the 'ser/estar' errors of the delivered Spanish text.*
>
> *Not every service information is available in all languages that are supported by the car manufacturer; the translation of service information is a matter of information need, but every available information is accessible through "pull" technology (for example, the hotline information service). Information of common interest is distributed in all languages through "push" technology; this ensures a fast and efficient update of all product documentation. The defect of Mr M's car is now available in English and Spanish (after a correction of the computer translation); after an in-depth analysis of the defect this information might become available in all supported languages.*

This scenario, which of course is partly a vision, shows the necessity of integrating services, documentation and networked information technology (IT) solutions with the support of modern, multilingual language technology (LT). The identified documentation bottlenecks and parts of this scenario then form the basis of the MULTIDOC vision of an Abstract Documentation Factory (ADF).

3 MULTIDOC Virtual Application

3.1 Translation Engineering

Within the Concerted Action a so-called virtual application was defined. It constitutes a compromise between the present situation of product documentation in the different automotive companies and the MULTIDOC vision of an ADF that is based on the concept of Translation Engineering. We talk about a virtual application because this application is based on the generalisation of the different processes and workflows maintained in the automotive product documentation environments of the MULTIDOC partners. For this we analysed the existing documentation processing chain and we identified the stages for initial improvements taking into account the historical evolution of product documentation. The sequential stages of today's documentation value cycle is profiled in Figure 1 below.

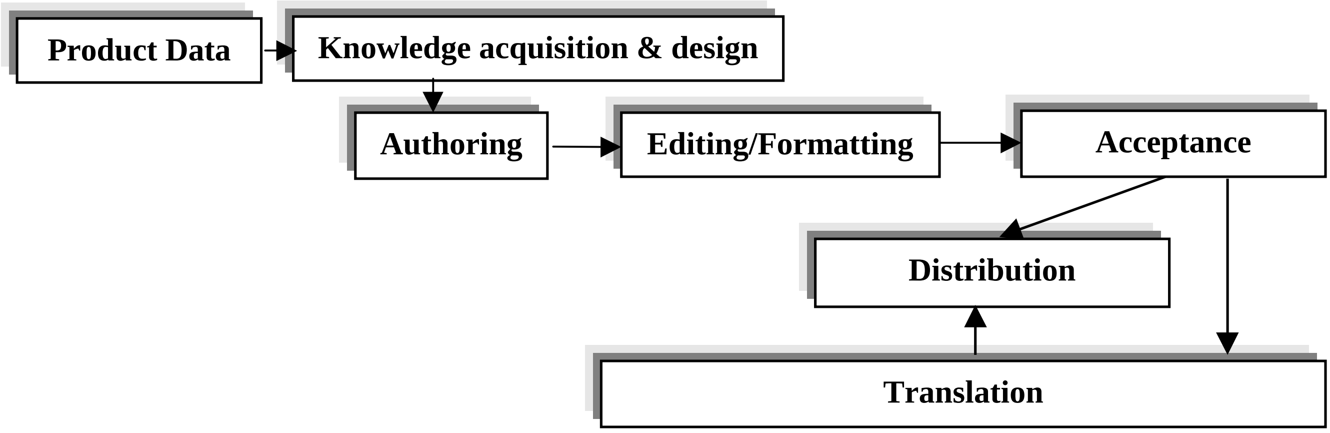


Figure 1: Today's Documentation Workflow

This analysis led to the definition of a strategy for an efficient and effective employment of language technology to bridge the gap between the present situation in product documentation and the MULTIDOC vision. This strategy is based on the present situation and has to be maintained with various restrictions for the different automotive companies but with the common interest to work toward the ADF vision that is shared by all companies, however, with different ways to reach the vision.

The virtual application is the result of the elaboration phase, and it allows for a smooth and cost effective transition of the business, because we have first and foremost concentrated on the existing process stages, where several control capabilities for the source language, such as spell, grammar and style checking functionality in the authoring stage, as well as the control of terminology consistency in the product data stage, the knowledge acquisition stage and the authoring stage, support the technical writer and other knowledge workers in identifying and defining information objects in an SGML authoring environment.

In broad terms, an information object is either a meaningful, non decomposable SGML marked-up text unit or a composition of such text units. The virtual application shall already include steps toward Translation Engineering that is the operational foundation of the MULTIDOC vision. Translation Engineering (TE) as a business strategy is concerned with:

- Fostering the use of information objects preferably linked with product data to ensure the consistent use of information structure and information content.
- Optimising the translation production chain through the employment of different multilingual language technologies, including multilingual generation.
- Linking of source language information and target language information to facilitate better maintenance, quality assurance and quality control.

TE as a methodology for multilingual documentation will help to drastically reduce cost, to shorten lead times and to further improve the quality of technical documentation. In addition, the core business and the component or system manufacturing industry benefits in terms of:

- accelerating the building of enterprise-wide and industry sector wide knowledge systems (repositories and knowledge bases) based on Web technology (intranets and extranets) including multimedia (text, graphics, video and animation, virtual reality) and multimodal (language and speech navigation) capabilities,
- improving the semantic content of information objects (accuracy and quality),
- speeding up of the translation processes (today, in limited cases the translation process could be substituted by multilingual generation and symbolic authoring, which in particular has to be seen in combination with controlled languages),
- reorganising of the overall production process (lean multilingual documentation).

A snapshot of this information distribution scenario is shown in Figure 2 below.

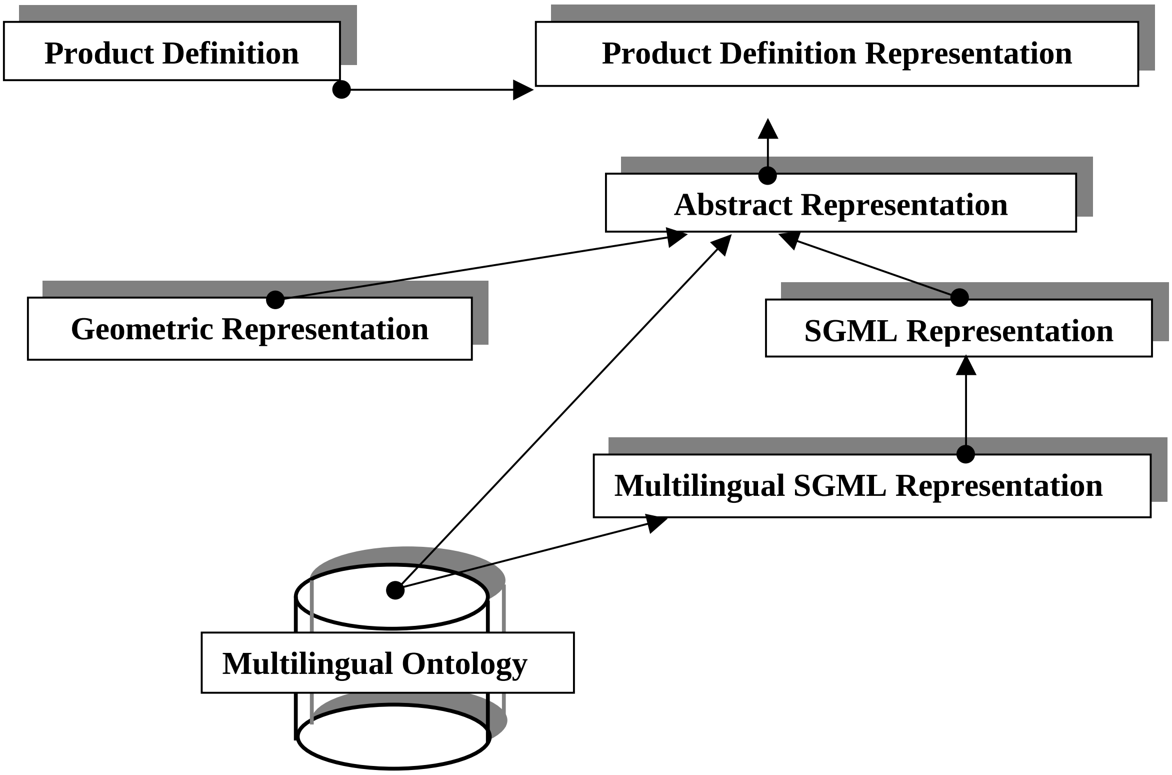


Figure 2: Information Distribution

TE as such will revolutionise the current way of thinking in technical documentation because the whole documentation process is oriented toward multilingualism. This new business scenario includes a push/pull policy for technical information delivery and retrieval in an automotive dealer’s workshop in combination with a translation-on-demand policy (see the scenario above). TE is responsive to the new business demands, and it will harmonise and unify the most crucial documentation requirements in areas such as the consistent use of technical information in structure and content, the efficient and effective reuse of information objects based on standardised information structures (increased retrieval hit-rates), and the terminological and multilingual orientation of the whole information production process.

3.2 Abstract Documentation Factory

The vision of the ADF comprises the complete re-organisation of the documentation processes. This means that LT in general, and specifically multilingual LT including translation technology, will make the move from a supporting technology to an enabling technology. The most important areas for this development are:

- Graphics and other multimedia incarnations, such as video, animation and virtual reality applications, may enrich or even replace text in certain information objects and facilitate new approaches to information production such as symbolic authoring.
- Translation-on-demand policy to allow for an efficient and effective control of the actual translation needs, because not all information objects need to be stored in every language that is supported by the business (translation management).
- Compilation of documents from multilingual information objects, either already stored in a foreign language, translated on demand, or generated from an abstract representation; this allows for the simultaneous delivery (publishing) of multilingual documentation.

The focus of the ADF is on the following three main components:

- Multilingual terminological ontology as a means for representing domain knowledge (the subject of technical documentation) linked with natural language semantics. Figure 3 below shows the different views on the ontology organisation at the upper level.
- Object Modelling Technique (OMT) as a theoretical foundation for analysis and design, and as an implementation platform based on distributed object environments such as, for example, CORBA.
- Agent technology as the overall umbrella for construction, and as an alternative implementation platform, especially for networked applications.

In the ADF, knowledge producer and knowledge consumer will operate in virtual environments brokered by software agents. A software agent acts autonomously on behalf of a person to fulfil the person’s goal or task. Agents are also key enabler for push technology which is used in information update tasks and information retrieval tasks.

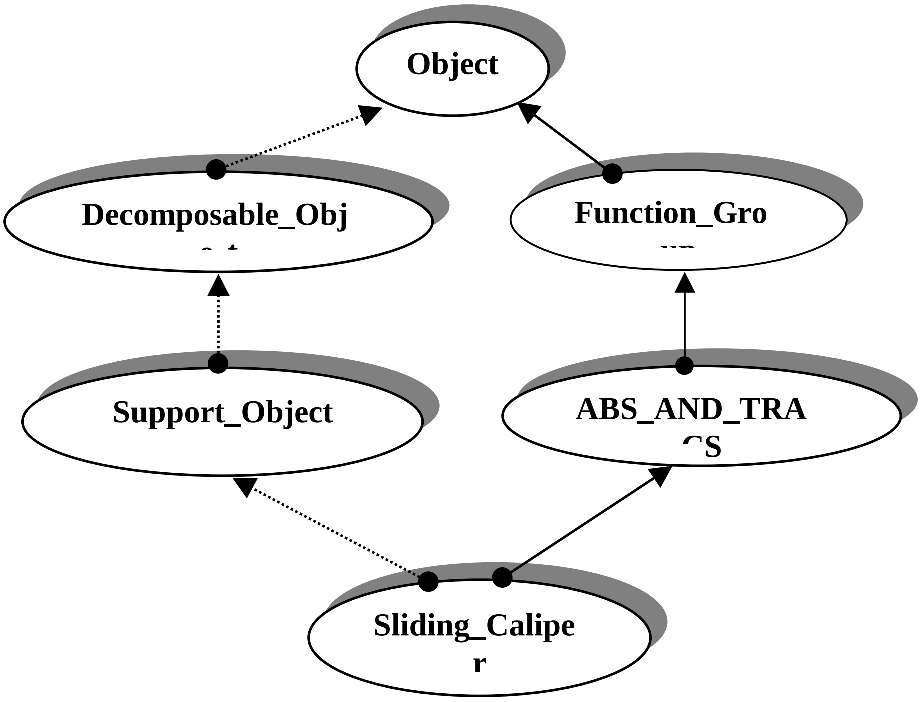


Figure 3: Different Views on the Ontology

3.3 Validation

All development strategies have been validated with a cost/benefit appraisal based on a hypothetical business calculation of a virtual automotive enterprise. We have taken this way to further maintain the generalisation direction which we already followed in the other phases of the Concerted Action. However, our profitability assessment is based on actual calculations made by the MULTIDOC partners for their specific enterprise situation. The validation stage also included a risk analysis consisting of:

- Critical analysis of our approach.
- Analysis of changes to the human resources, the technical infrastructure and the organisational environment.
- Identification of any restrictions, constraints, risks and problems hitherto not taken into account.

In the following, we will demonstrate that the effective control of terminology helps to reduce costs at a very early stage of the documentation workflow. This is motivated by the costs that are needed to detect and repair a terminology error.

Let us assume that a unit cost of one is assigned to the effort required to detect and repair an error during the authoring stage, then the cost to detect and repair an error during the data gathering, harmonisation (synchronisation between product data and product documentation) and documentation design stages (which are similar to the requirements stages in software engineering) is between five to ten times less. Furthermore, the cost to detect and repair an error during the maintenance stage is twenty times more. The reasons for this large difference is that many of these errors are not detected until well after they have been made. This delay in error discovery means that the cost to repair includes the cost to correct the offending error and to correct subsequent investments in the error. These investments include rework (perhaps redesign) of documentation, rewrite of related documentation, and the cost to rework or replace documentation in the field. Figure 44 below shows the cost pyramid of the different stages of error detection and correction.

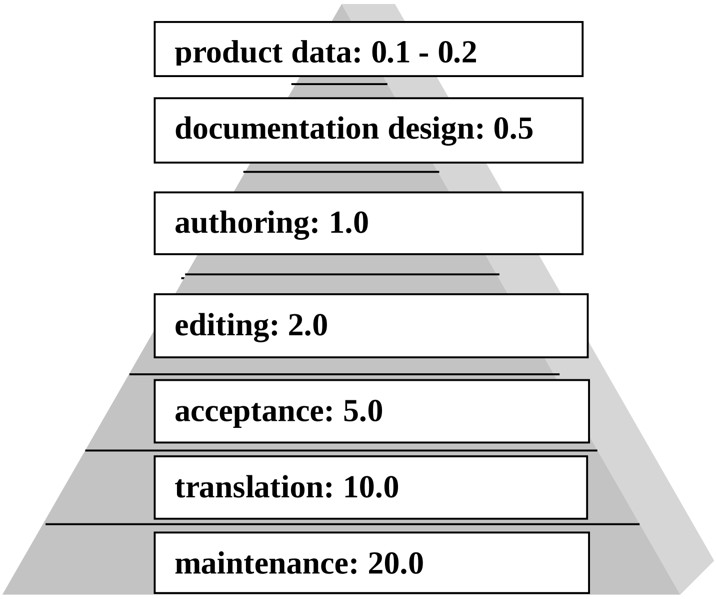


Figure 4: Cost Pyramid for Detecting and Correcting Terminology Errors

This shows that errors made at early stages in the documentation workflow are extremely expensive to repair. If such error occurred infrequently, then the contribution to the overall documentation cost would not be significant. However, terminology errors are indeed a large class of errors typically found in complex technical documentation. These errors could be between 30 % and 70 % of the errors discovered in technical documentation.

It seems reasonable to assume that a 20 % or more reduction in terminology errors can be accomplished at various levels of organisational maturity. Because of the multiplying effect, any such reduction can have a dramatic overall effect to our project's bottom line.

The user groups who are in the focus of our work are, on the one hand, the knowledge producers of documentation departments and translation departments (technical writers and translators, designers and engineers, and so forth) and the knowledge consumers in the automotive workshops (mechanics and technicians).

4 Conclusions and Perspectives

Both approaches, "bridging the gap" and ADF, are centred around the MULTIDOC terminological ontology as the primary information source. The parallel development allows for an optimal use of resources, and permits a straight forward implementation of the ADF based on already existing LT modules and components. It should be noted that the ontology centric approach has several benefits which have an direct impact on the most important business demands in multilingual documentation. Among them the most significant ones are that it

- supports harmonisation and standardisation between product data environments and product documentation environments;

- ensures better control of information object production, translation and deployment because of clearly defined responsibilities and quality assurance measures, including a transparent workflow control;
- fosters a better integration of supplier information and subsidiary information (core business as well as associated businesses);
- abandons the need of an end-control within the information object production value cycle because of the distributed responsibilities with an integrated control functionality;
- fits with existing and emerging networked computing environments, including advanced agent technology.

More details on the project can be obtained from our web pages at URL:

http://www.iai.uni-sb.de/MULTIDOC.

5. References

Bevan, Nigel (1996): User-centred Design. Deliverable D3.1.2 v1.2 of IE 2016 project INUSE.

Bradshaw, Jeffrey M. ed. (1997): *Software Agents*. AAAI Press/MIT Press, Cambridge, Massachusetts, USA.

Church, Ken and Hovy, Eduard (1993): Good applications for crummy machine translation. *Machine Translation* 8,

Cockburn, Alistair (1997a): *Structuring Use Cases with Goals*. Web document at URL http://members.aol.com/cockburn/papers/usecases.htm.

Cockburn, Alistair (1997b): *Using "V-W" Staging to Clarify Spiral Development*. Web document at URL http://members.aol.com/cockburn/papers/ vwstage.htm

Englander, Robert (1997): *Developing Java Beans*. O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Flanagan, David (1996): *Java in a Nutshell*. O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Flanagan, David (1997): *Java in a Nutshell*. Second Edition covering Java 1.1, O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Foyler, Martin (1997): *UML Distilled*. Addision Wesley Publishing Company, Reading, MA, USA.

Herdman, David (1997): No mirrors and magic here - You can write common code for Unix and NT. *Sunworld On-line*, April issue and May issue.

Howells, Ian (1997): Document Management Beyond 2000: More of the Same, a Strategic Tool or Spider in the Web?. In: *Document World*, Issue 1, IMC and AIIM publication, Powerhouse Solutions Ltd., Surrey, England.

Huntington, Samuel P. (1996): *The Clash of Civilizations*. Simon and Schuster, New York, USA.

Maguire, Martin (1997): RESPECT User Requirements Framework Handbook. Deliverable D5.1 of Information Engineering (IE) 2010 project RESPECT. HUSAT Research

Institute, Loughborough, UK.

Object Management Group (1995): The Common Object Request Broker: Architecture and Specification Version 2.0

Orfali, Robert et al. (1996): *The Essential Distributed Objects Survival Guide*. John Wiley & Sons, Inc., New York, NY, USA.

Orfali, Robert and Harkey, Dan (1997): *Client/Server Programming with Java and CORBA.* John Wiley & Sons, Inc., New York, NY, USA.

Stark, Heather, et al. (1996): *Ovum Evaluates: Document Management*. Ovum Ltd, London, England.

Reese, George (1997): *Database Programming with JDBC and Java*. O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Schütz, Jörg (1996a): Combining Language Technology and Web Technology to Streamline an Automotive Hotline Support Service. In *Proceedings of AMTA-96*, Montreal, Canada.

Schütz, Jörg (1996b):. Network-based Machine Translation Service. In*: Proceedings of the EAMT Machine Translation Workshop*, Vienna, Austria.

Schütz, Jörg (1997): Utilizing Evaluation in Networked Machine Translation. In *Proceedings of TMI 97*, Santa Fe, New Mexico, USA.

Sowa, John F. (1984): *Conceptual Structures: Information Processing in Mind and Machine*. Addision Wesley Publishing Company, Reading, MA, USA.

Sowa, John F. (to appear): *Knowledge Representation: Logical, Philosophical and Computational Foundation.* To be published by PWS Publishing Company.

Tucker, Hugh A. (1996): Harmonization of SGML and STEP. ISO (TC 184/SC4/WG3/T14 Product Documentation) STEP/SGML Tutorial.

UML (1997): Unified Modelling Language (UML) Documentation, Version 1.0. Rational Software Corporation, Santa Clara, California, USA. Available on the Web at URL http://www.rational.com.

Unicode (1996): *The Unicode Standard: World-wide Character Encoding*. Addison- Wesley Publishing Company, Reading, MA, USA.

Supporting Controlled Language Authoring

Pim van der Eijk and Jacqueline van Wees

Introduction[3]

Since the early 1990s, Cap Gemini Language Technology (currently part of Cap Gemini's Advanced Technology Services group, Utrecht, the Netherlands) has been developing software to support large scale document creation and translation of technical documentation using controlled sublanguages, and deploying this software in customer projects. From its inception, activities have concentrated on Controlled Languages satisfying severe lexical and syntactic restrictions, such as on-line help texts, software manuals, and aerospace maintenance manuals.

The formalism for analysis grammars has a built-in mechanism for word-level, morpho-syntactic and terminological error correction. In addition to this, it is possible to specify more general correction transformation annotations to rules. Grammars can be compiled into correction modules that can be integrated in commercial DTP products for interactive use by technical writers.

Controlled Languages and applications

Controlled sublanguages are derived variants of sublanguages, constructed to impose precise coverage bounds and application-specific additional constraints such as improved understandability, ambiguity reduction and increased ease of (machine) translation. User acceptance and clear business benefits are important factors determining feasibility and success of Controlled Language implementations.

The business case for investment in (computer support for) Controlled Language is application and customer dependent. In some cases, it can be based on a time to market reduction for localized foreign language versions of products, which can be achieved by shortening editorial review cycles and reducing Machine Translation post-editing costs. In other cases, improved quality of technical documentation can reduce the Mean Time To Repair metric for complex, expensive systems, and thus reduce cost or improve customer satisfaction. Fortunately, case studies demonstrating these benefits exist and increase market interest in Controlled Language technology and services.

There are two important acceptance factors regarding the introduction of a Controlled Language in a user community.

- A first criterion is the degree to which users, both authors and the target audience of the documents, find sample representative sublanguage documents, rewritten in the Controlled Language, to be acceptable paraphrases of the original documents.

Our experience confirms the experience at other sites that rewritten documents often match or exceed the originals in clarity and ease of understanding.

[3] An earlier version of this document appeared in the first Controlled Language Application Workshop, Leuven, 1996.

- A second criterion for a “natural” sublanguage is the ease with which technical writers can create new sublanguage documents in the Controlled Language, and perceive the Controlled Language to be intuitively “close” to the sublanguage on which it is based.

In practice, the second restriction is considerably harder. Grammar restrictions often can only be expressed in a linguistic jargon that is not always easy to explain to technical writers, who normally are domain experts with no or limited linguistic background. This can be alleviated to some extent by using dedicated authors, who are trained and coached well in the use of the system, and useful feedback from the system.

Activities and phases in Controlled Language application

As we define the concept, a Controlled Language is a variant of an existing sublanguage, in which expressions in the sublanguage are related, via a paraphrase relation, to expressions in the Controlled Language that satisfy specific additional constraints. Documents paraphrased, or created from scratch, in the Controlled Language should be able to perform the communicative functions of the document at least as well as corresponding documents in the non-Controlled Language, throughout the various stages in the document lifecycle.

The design of a Controlled Language therefore involves the following activities:

1. Sublanguage analysis;

2. Specification of constraints on the Controlled Language;

3. Specification of a paraphrase relation from expressions in the sublanguage to expressions in the Controlled Language.

In practice, the three classes of activities will be separated temporally into separate (phases of) projects, ranging from initial analysis, as part of an initial feasibility study, to implementation. Sublanguage analysis requires the availability of a representative corpus for the sublanguage. Issues to be looked into during analysis are, for example, word volume, translation workload, lexical growth, parts of speech distribution, terminological ratio, homography and polysemy ratio, lexical coverage projection and linguistic complexity ratio of major phrase structures.

The second element in the specification of the Controlled Language is the specification of the Controlled Language. The specification of the Controlled Language can be formalized as a grammar in a grammar formalism and an associated lexicon that can be compiled into a recognizer or parser of the Controlled Language. In applications involving translation, development of this grammar will normally be synchronized with development of the translation system. Existing industry specifications such as the aerospace industry’s Simplified English standard can be viewed as starting points in the development of these grammars.

The third element of a Controlled Language is the association of expressions in the uncontrolled sublanguage and expressions in its controlled subset. To some extent, it will be possible to formalize this association as lexical or syntactic transformations from the sublanguage into the Controlled Language. There can be zero (no paraphrase in the Controlled Language), a single (rewritable to a single, possibly identical, Controlled Language expression), or many (an ambiguous sublanguage expression) Controlled

Language expressions per sublanguage expression. Part of the association, e.g. the part described as informal stylistic instructions in a style guide, will not be formalizable at all. In some cases, a particular error type can be detected, but not corrected automatically. In these cases, it is sometimes possible to generate informative messages that could help technical writers rephrase the sentence.

To support the authoring process, it is therefore necessary to combine a variety of functions in a single system, viz. recognition and parsing of a Controlled Language, transformation of general sublanguage expressions into Controlled Language, and error correction. Cap Gemini's lingware formalism was designed to incorporate these various types of functionality in a single formalism.

It should be stressed that only some sublanguages allow for a Controlled Language approach because of insufficient lexical or grammatical convergence, or because of inherent ambiguity.

Authoring Controlled Languages

Technical writers often find it hard to create new documents in a Controlled Language (or to rewrite existing documents), especially if a large number of previously acceptable sublanguage constructions can no longer be used. To prevent frustration, they should know how to paraphrase these constructions in the Controlled Language. Apart from training, it is important to provide authors with supporting software to support the authoring process. These supporting function can be divided in checking tools, which generate informative diagnostic messages for authors, and correction tools. The objective is to be able to correct as many errors as possible, and as automatically as possible.

In our system, a correction module accepts a language defined as four successively larger sets.

1. The system recognizes and assigns lexical and structural descriptions to the subset of sublanguage expressions that conform to language control constraints.

2. This set is expanded to include as large a part of the sublanguage as can be transformed, automatically or interactively, to the Controlled Language.

3. A third expansion is inclusion of variant expressions that contain morpho-syntactic errors.

4. Finally, expressions containing orthographic errors are corrected.

An integration of the system with Microsoft Word has been developed and is currently being marketed within the international customer base of Cap Gemini. Similar integrations with other desktop publishing products will be developed, depending on customer demand and market feedback.

Controlled Language analysis and correction lingware

The Cap Gemini lingware formalism was designed to facilitate development of interactive grammar checking applications. Using proprietary LR compiler software, the grammars can be compiled into correction modules, performance of which is fast enough for interactive use on commodity office PCs. In the sample application discussed below, the correction engine is accessed at runtime, as a shared library, from Microsoft Word.

The lexical database is stored separately, and has its own separate maintenance utilities. To obviate the need for computationally expensive run-time morphological analysis, the run-time system uses an exhaustive full-form lexicon.

The valid constructs of the Controlled Language are described using extended context free grammar rules, annotated with dependency relations among attributes. The grammar can be augmented with correction rules, which are similar to normal grammar rules but are enhanced with instructions for local reordering and deletion, insertion of lexical items, and diagnostic messages. In the lexicon, words are organized into synonym sets, individual members of which can be marked as (non-)preferred. Per rule, word forms are organized in syntactic equivalence classes based on attribute dependencies, which are used to carry out morpho-syntactic (e.g. agreement) and terminological (use of unapproved word forms) corrections.

As an illustrative example, consider the following English input sentence, which contains non-preferred terms, a morpho-syntactic and an orthographic error:

> *Check that leading edges conforms to values in teh table.*

It is converted automatically to the following ‘correct’ Simplified English sentence:

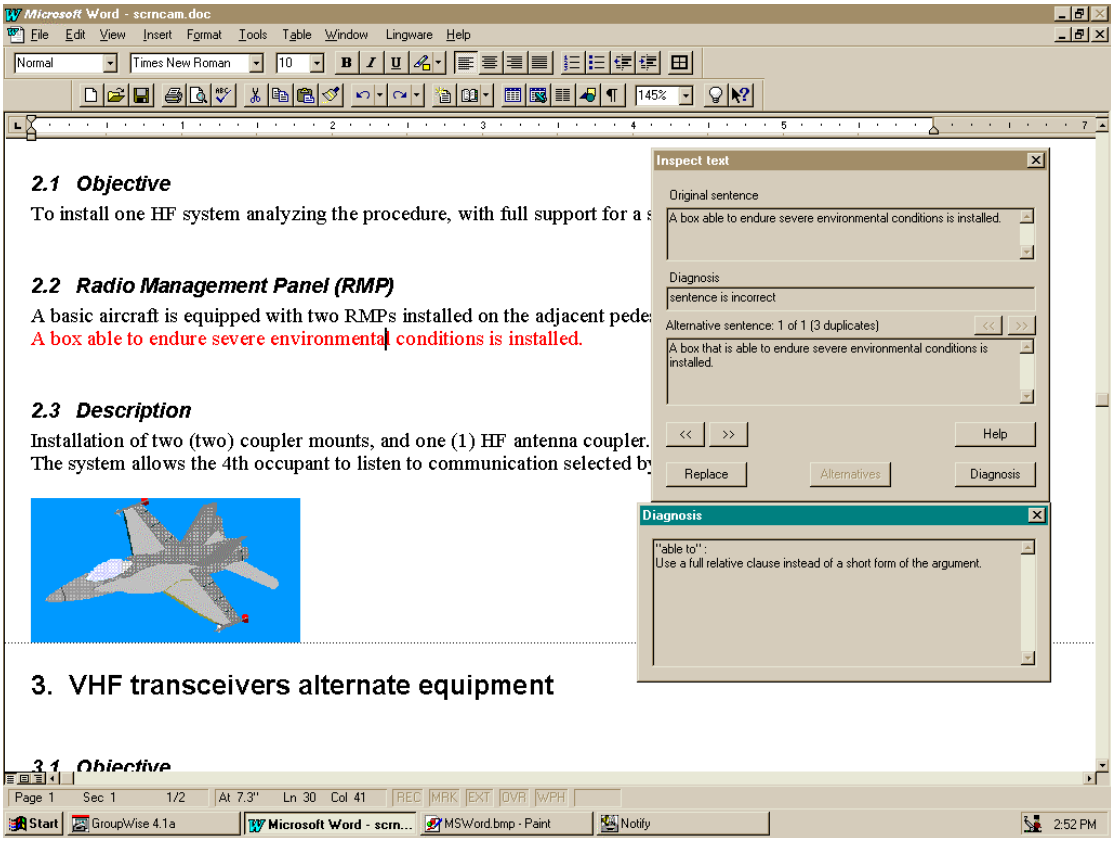
> *Make sure that the leading edges agree with the values in the table.*

First of all, and least interestingly, the misspelled article *teh* is corrected to *the* via a fuzzy string matching mechanism. In analysis, the non-approved word form “conforms” is connected to a synonym set that has “AGREE” as approved word. In the grammatical context, this lemma is associated with the inflected forms “agree” and “agrees”, the first of which is selected because of agreement dependency with the subject noun phrase. Similarly, the preposition “to” is associated with the generic complement PP preposition. The word form “with” is selected because of agreement in the attribute *pform* with the verb.

The unapproved word “checks” is associated with three approved constructions, viz. “MAKE SURE”, “MEASURE”, and “EXAMINE”. The latter two take NP complements and the former a sentential complement, as appropriate in the case at hand. The Noun Phrase rewrite rule contains an insertion instruction that supplies the missing article preceding the plural noun. This sentence can therefore be corrected in a completely automatic fashion. Use of “check” with a complement NP would be ambiguous between “MEASURE” and “EXAMINE”. In interactive use, the correction engine would consult with the user to obtain the necessary disambiguation information.

Integrating language correction in an authoring environment

Controlled Language correction, as a supporting function in a document creation process, is naturally viewed as an extension to standard document editing functions. Modern desktop publishing products support this view by offering integration toolkits that can be used to add specialized functionality to the core DTP functionality. As an example of such an extension, we developed a prototype integration of a Controlled English correction system in Microsoft Word. The following example shows the application of the editor to a sample aerospace document.


Microsoft Word - scrncam.doc
File Edit View Insert Format Tools Table Window Lingware Help
Normal
Times New Roman
10
145%

2.1 Objective

To install one HF system analyzing the procedure, with full support for a s

2.2 Radio Management Panel (RMP)

A basic aircraft is equipped with two RMPs installed on the adjacent pede

A box able to endure severe environmental conditions is installed.

2.3 Description

Installation of two (two) coupler mounts, and one (1) HF antenna coupler.
The system allows the 4th occupant to listen to communication selected by

3. VHF transceivers alternate equipment

3.1 Objective

Inspect text

Original sentence: A box able to endure severe environmental conditions is installed.

Diagnosis: sentence is incorrect

Alternative sentence: 1 of 1 (3 duplicates)

A box that is able to endure severe environmental conditions is installed.

Diagnosis

"able to" :
Use a full relative clause instead of a short form of the argument.

The translator’s workbench and beyond: off-line add-ons to on-line tools

Pierre Lewalle

Abstract

The advent of new information technologies at an ever increasing pace opens up new avenues for promising developments in the field of language (and translation) processing. Simultaneously, the drastic drop in prices of the technology turns prospects into affordable wonder-solutions, which tend to live up to expectations of actual users only to a limited extent, and probably even less so to those of cost-sensitive managers. Progress in natural language processing research has been sustained in recent decades, but it would be fallacious to make believe that solutions are available off the shelf to do anything with language that needs to be done in today’s communication environment, in no time and at no cost. While the irrational, often summary, dismissal by traditionalist circles of technological support which may help translators in their daily work is outdated, it is necessary to depict a clear picture of the current situation and the characteristics of the present day working conditions : good quality translation will continue to rely on the skills of trained, widely-read and skilled human translators. The difference with past practice lies probably in the fact that the resources indispensable for the work will be more easily available and at a lower cost, that the mass of information which can be used will increase dramatically in scope, that repetitive tasks will be facilitated by the use of robots which never get tired or bored and can keep on sustaining search processes long after a human brain would have ceased to function properly. What will not change, on the other hand, is that human translators can translate only so many pages a day with a consistent reliability, and that gains in performance in this respect will be necessarily limited. As a corollary, expected cost savings will become reality only if a number of preconditions are met : efficient use of interactive tools will be dependent on proper training of the users; constraints inherent to the automated systems will have to be understood and accepted; reasonable expectation targets will have to be taught after discounting promotional mirages; long, difficult preparations will have to be undertaken to pave the way for the implementation of the new technologies. Prospects for savings will eventually depend on the quality of the archive materials to be used and the corresponding level of investment required to recycle them for electronic processing, the compliance with recognized standards for newly produced materials and the resulting re-usability of such materials for a variety of purposes. Only on that basis will it be possible to embark on a large-scale automatic processing of text materials for efficiently supporting translation work, but also for document indexing, information management at large and global communication development for the benefit of the world community.

Multilingual Tools at the Xerox Research Centre

Jean-Pierre Chanod

Introduction

The Xerox Research Centre Europe (http://www.xrce.xerox.com for more information) pursues a vision of document technology where language, physical location and medium - electronic, paper or other - impose no barrier to effective use.

Our primary activity is research. Our second activity is a Program of Advanced Technology Development, to create new document services based on our own research and that of the wider Xerox community. We also participate actively in exchange programs with European partners.

Language issues cover important aspects in the production and use of documents. As such, language is a central theme of our research activities. More particularly, our Centre focuses on multilingual aspects of Natural Language Processing (NLP). Our current developments cover more than ten European languages and some non-European languages such as Arabic. Some of these developments are conducted through direct collaboration with academic institutions all over Europe.

The present article is an introduction to our basic linguistic components and to some of their multilingual applications.

1. LINGUISTIC COMPONENTS

The MLTT (Multilingual Theory and Technology) team creates basic tools for linguistic analysis, e.g. morphological analysers, taggers, parsing and generation platforms. These tools are used to develop descriptions of various languages and the relation between them. They are later integrated into higher level applications, such as terminology extraction, information retrieval or translation aid. The Xerox Linguistic Development Architecture (XeLDA) developed by the Advanced Technology Systems group incorporates the MLTT language technology.

Finite-state technology is the fundamental technology on which Xerox language R&D is based. It encompasses both work on the basic calculus and on linguistic tools, in particular in the domain of morphology and syntax.

Finite-state calculus

The basic calculus is built on a central library that implements the fundamental operations on finite-state networks. It is based on long-term Xerox research, originated at PARC in the early 1980s. The most recent development in the finite-state calculus is the introduction of the replace operator. The replacement operation is defined in a very general way, allowing replacement to be constrained by input and output contexts, as in two-level rules but without the restriction of only single-symbol replacements. Replacements can be combined with other kinds of operations, such as composition and union, to form complex expressions.

The finite-state calculus is widely used in our linguistic development, to create tokenisers, morphological analysers, noun phrase extractors, shallow parsers and other language-specific linguistic components.

Morphology

The MLTT work on morphology is based on the fundamental insight that word formation and morphological or orthographic alternation can be solved with the help of finite automata:

1. the allowed combinations of morphemes can be encoded as a finite-state network;
2. the rules that determine the form of each morpheme can be implemented as finite-state transducers;
3. the lexicon network and the rule transducers can be composed into a single automaton, a lexical transducer, that contains all the morphological information about the language including derivation, inflection, and compounding.

Lexical transducers have many advantages. They are bi-directional (the same network for both analysis and generation), fast (thousands of words per second), and compact.

We have created comprehensive morphological analysers for many languages including English, German, Dutch, French, Italian, Spanish, and Portuguese. More recent developments include Czech, Hungarian, Polish, Russian, Scandinavian languages and Arabic.

Part-of-speech tagging

The general purpose of a part-of-speech tagger is to associate each word in a text with its morphosyntactic category (represented by a tag), as in the following example:

This+PRON is+VAUX_3SG a+DET sentence+NOUN_SG .+SENT

The process of tagging consists in three steps:

1. tokenisation: break a text into tokens
2. lexical lookup: provide all potential tags for each token
3. disambiguation: assign to each token a single tag

Each step is performed by an application program which uses language specific data:

- The tokenisation step uses a finite-state transducer to insert token boundaries around simple words (or multi-word expressions), punctuation, numbers, etc.
- Lexical lookup requires a morphological analyser to associate each token with one or more readings. Unknown words are handled by a guesser which provides potential part-of-speech categories based on affix patterns.
- Disambiguation is done with statistical methods (Hidden Markov Model), although we also experiment with fully rule-based methods.

Noun Phrase Extraction

For the purpose of terminology extraction from technical documents we designed a tool which applies finite-state techniques to mark potential terms, especially noun phrases corresponding to given regular patterns. The noun-phrase extraction tool consists of several modules: language independent programs (tokeniser, part-of-speech disambiguator, and noun phrase mark-up) and language dependant data (finite-state transducers and transition probabilities). This modular architecture allows rapid extension to different languages. Currently, implementations for 8 languages (Dutch, English, French, German, Hungarian, Italian, Portuguese, Spanish) exist; more languages (e.g. Czech, Polish, Russian) are in preparation.

Noun phrase (NP) mark-up applies finite-state automata describing noun phrase patterns. These patterns rely on the simple (non-ambiguous) tagger output format, i.e. they consist of regular expressions on sequences of tokens and tags.

A very simple noun phrase description for a given language (e.g. French) may consist in a (possibly empty) sequence of adjectives followed by a noun and another sequence of adjectives. The automata which describe noun phrases are compiled into the final NP-mark-up. The compilation script uses the directed replace operation for the longest match and inserts brackets around maximal NPs (according to the NP patterns). The final NP-mark-up transducers are non-ambiguous, i.e. for every input they provide a single output containing non-recursive bracketing for NPs.

The following examples from the current realisations for French, Dutch and Spanish illustrate the application of the complete chain of tokenising, part-of-speech disambiguation and noun phrase mark-up:

> *Lorsqu'on tourne le commutateur de démarrage sur la position auxiliaire, l'aiguille retourne alors à zéro.*
>
> Lorsqu'/CONN on/PRON tourne/VERBP3SG le/DETSG
> **NP{commutateur/NOUNSG de/PREPDE démarrage/NOUNSG}**
> sur/PREP la/DETSG
> **NP{position/NOUNSG auxiliaire/ADJSG}**
> ,/CM l'/DETSG
> **NP{aiguille/NOUNSG}**
> retourne/VERBP3SG alors/ADV à/PREPA
> **NP{zéro/NOUNSG}**
>
> *De reparatie- en afstelprocedures zijn bedoeld ter ondersteuning voor zowel de volledig gediplomeerde monteur als de monteur met, minder ervaring.*
>
> De/ART **NP{reparatie-/CMPDPART en/CON afstelprocedures/NOUN}** zijn/VAFIN bedoeld/VVPP ter/PREP **NP{ondersteuning/NOUN}** voor/PREP zowel/CON de/ART **NP{volledig/ADJA gediplomeerde/ADJA monteur/NOUN}** als/PREP de/ART **NP{monteur/NOUN}** met/PREP minder/INDDET **NP{ervaring/NOUN}**
>
> *Para asegurar el funcionamiento óptimo de los vehículos, así como la seguridad personal del técnico, es imprescindible seguir los métodos apropiados de trabajo y los procedimientos correctos de reparación.*
>
> Para/PREP asegurar/VINF el/DETSG **NP{funcionamiento/NOUNSG óptimo/ADJSG de/PREP los/DETPL vehículos/NOUNPL}**,/COMA así~como/CONJ la/DETSG **NP{seguridad/NOUNSG personal/ADJSG del/PREPDET técnico/NOUNSG}** ,/COMA es/AUX imprescindible/ADJSG seguir/VINF los/DETPL **NP{métodos/NOUNPL apropiados/VPASTPARTPL de/PREP trabajo/NOUNSG}** y/CONJ los/DETPL

NP{procedimientos/NOUNPL correctos/ADJPL de/PREP reparación/NOUNSG}

Naturally, in a terminology management application, noun phrase extraction leads only to the selection of candidate terms. This automatic selection remains to be validated by human terminologists.

Additionally, by combining monolingual NP extraction as described above with alignment techniques based on statistical methods, one may extend the application to bilingual terminology extraction. Candidate terms are first extracted independently for language A and B. Aligned terms are then spotted by evaluating how often a given bilingual pair of terms (T_a, T_b) appears within aligned sentences. Again, in terminology management, bilingual extraction as well as alignment needs to be further validated by human specialists.

Incremental finite-state parsing

Finite State Parsing is an extension of finite state technology to the level of phrases and sentences.

Our work concentrates on shallow parsing of unrestricted texts. We compute syntactic structures, without fully analysing linguistic phenomena that require deep semantic or pragmatic knowledge. For instance, PP-attachment, co-ordinated or elliptic structures are not always fully analysed. The annotation scheme remains underspecified with respect to yet unresolved issues. On the other hand, such phenomena do not cause parse failures, even on complex sentences.

Syntactic information is added at the sentence level in an incremental way, depending on the contextual information available at a given stage. The implementation relies on a sequence of networks built with the replace operator. The current system has been implemented for French and is being expanded to new languages.

The parsing process is incremental in the sense that the linguistic description attached to a given transducer in the sequence relies on the preceding sequence of transducers, covers only some occurrences of a given linguistic phenomenon and can be revised at a later stage. The parser output can be used for further processing such as extraction of dependency relations over unrestricted corpora. In tests on French corpora (technical manuals, newspaper), precision is around 90-97% for subjects (84-88% for objects) and recall around 86-92% for subjects (80-90% for objects).

2. APPLICATIONS

2.1. LOCOLEX: a Machine Aided Comprehension Dictionary

LOCOLEX is an on-line bilingual comprehension dictionary, which aids the understanding of electronic documents written in a foreign language. It displays only the appropriate part of a dictionary entry when a user clicks on a word in a given context. The system disambiguates parts of speech and recognises multiword expressions such as compounds (e.g. *heart attack*), phrasal verbs (e.g. *to nit pick*), idiomatic expressions (e.g. *to take the bull by the horns*) and proverbs (e.g. *birds of a feather flock together*). In such cases LOCOLEX displays the translation of the whole phrase and not the translation of the word the user has clicked on.

For instance, someone may use a French/English dictionary to understand the following text written in French:

Lorsqu'on évoque devant les ***cadres*** *la séparation négociée, les rumeurs fantaisistes vont apparemment toujours bon* ***train***.

When the user clicks on the word *cadres*, LOCOLEX identifies its POS and base form. It then displays the corresponding entry, here the noun *cadre*, with its different sense indicators and associated translations. In this particular context, the verb reading of *cadres* is ignored by LOCOLEX. Actually, in order to make the entry easier to use, only essential elements are displayed:

cadre I: nm
1: *[constr,art] (of a picture, a window) frame
2: *(scenery) setting
3: *(milieu) surroundings
4: *(structure, context) framework
5: *(employee) executive
6: *(of a bike, motorcycle) frame

The word *train* in the same example above is part of a verbal multiword expression *aller bon train*. In our example, the expression is inflected and two adverbs have been stuck in between the head verb and its complement. Still LOCOLEX retrieves only the equivalent expression in English *to be flying around* and not the entire entry for *train*.

train I: nm
5 : * [rumeurs] aller bon train : to be flying round

LOCOLEX uses an SGML-tagged bilingual dictionary (the Oxford-Hachette French English Dictionary). To adapt this dictionary to LOCOLEX required the following:

- Revision of an SGML-tagged Dictionary to build a disambiguated active dictionary (DAD);
- Rewriting multi-word expressions as regular expressions using a special grammar;
- Building a finite state machine which compactly associates index numbers with dictionary entries.

The lookup process itself may be represented as follows:

- split the sentence string into words (tokenisation);
- normalise each word to a standard form by changing cases and considering spelling variants;
- identify all possible morpho-syntatic usages (base form and morpho-syntactic tags) for each word in the sentence;
- disambiguate the POS;
- find relevant entries (including possible homographs or compounds) in the dictionary for the lexical form(s) chosen by the POS disambiguator;
- use the result of the morphological analysis and disambiguation to eliminate irrelevant sections;
- process the regular expressions to see if they match the word’s actual context in order to identify special or idiomatic usages;

- display to the user only the most appropriate translation based on the part of speech and surrounding context.

Besides being an effective tool for understanding, LOCOLEX could also be useful in the framework of language learning. LOCOLEX also points out that existing on-line dictionaries, even when organised like a database rather than a set of type-setting instructions, are not necessarily suitable for NLP-applications. By adding grammar rules to the dictionary in order to describe the possible variations of multiword expressions we add a dynamic feature to this dictionary. SGML functions no longer point to text but to programs.

2.2. Multilingual Information Retrieval

Many of the linguistic tools being developed at our Centre are being used in applied research into multilingual information retrieval. Multilingual information retrieval allows the interrogation of texts written in a target language B by users asking questions in source language A.

In order to perform this retrieval, the following linguistic processing steps are performed on the documents and the query:

- Automatically recognise language of the text.
- Perform the morphological analysis of the text using Xerox finite state analysers.
- Part of speech tag the words in the text using the preceding morphological analysis and the probability of finding part-of-speech tag paths in the text.
- Lemmatise, i.e. normalise or reduce to dictionary entry form, the words in the text using the part of speech tags.

This morphological analysis, tagging, and subsequent lemmatisation of analysed words has proved to be a useful improvement for information retrieval as any information-retrieval specific stemming. To process a given query, an intermediate form of the query must be generated which he normalised language of the query to the indexed text of the documents. This intermediate form can be constructed by replacing each word with target language words through an on-line bilingual dictionary. The intermediate query, which is in the same language as the target documents, is passed along to a traditional information retrieval system, such as SMART[4]. This simple word-based method is the first approach we have been testing. Initial runs indicate that incorporating multi-word expression matching can significantly improve results. The multi-word expressions most interesting for information retrieval are terminological expressions, which most often appear as noun phrases in English.

2.3. Callimaque: a collaborative project for virtual libraries

Digital libraries represent a new way of accessing information distributed all over the world, via the use of a computer connected to the Internet network. Whereas a physical library deals primarily with physical data, a digital library deals with electronic documents such as texts, pictures, sounds and video.

We expect more from a digital library than only the possibility of browsing its documents. A digital library front-end should provide users with a set of tools for querying and retrieving information, as well as annotating pages of a document, defining hyper-links between pages or helping to understand multilingual documents.

[4] This software is available for research purposes at ftp://ftp.cs.cornell.edu/pub/smart.

Callimaque is one of our projects dealing with such new functionalities for digital libraries. More precisely, Callimaque is a collaborative project between the Xerox Research Centre and research/academic institutions of the Grenoble area (IMAG, INRIA, CICG). The goal is to build a virtual library that reconstructs the early history of information technology in France. The project is based on a similar project, the Class project, which was started by the University of Cornell several years ago under the leadership of Stuart Lynn to preserve brittling old books. The Class project runs over conventional networks and all scanned material is in English.

The Callimaque project includes the following steps:

- Scanning and indexing around 1000 technical reports and 2000 theses written at the University of Grenoble, using Xerox XDOD, a system integrated with a scanner, a PC, a high-speed printer, software for dequeueing, indexing, storing, etc. Numerised documents can be reworked page by page and even restructured at the user's convenience. 30 Gbytes of memory are needed to store the images. Abstracts are OCRed to permit textual search.
- Documents are recorded on a relational database on a UNIX server. A number of identifiers (title, author, reference number, abstract, etc.) are associated with each document to facilitate the search
- Multilingual terminology derived from multilingual abstracts allows the system to process non-French queries.
- With a view to making these documents widely accessible, Xerox has developed software which authorises access to this database by any client using the http protocol used by the World Wide Web. The base is thus accessible via any PC, Macintosh, UNIX station or even from a simple ASCII terminal (The web address is http://callimaque.grenet.fr).
- Print on demand facilities connected to the network allow the users to make copies of the scanned material. This connection will subsequently develop towards a high output ATM network.

2.4. Xerox Translation and Authoring Systems (XTRAS)

2.4.1. XTRAS Terminology Suite

2.4.1.1. TermFinder: Multilingual Terminology Extraction

TermFinder enables the user to semi-automatically create multilingual terminology, hence ensuring a huge productivity increase over manual terminology creation. TermFinder is based on the linguistic components described above, especially NP extraction tools and alignment. TermFinder supports Dutch, English, French, German, Italian, Spanish, and Portuguese. Any of these languages can be source or target.

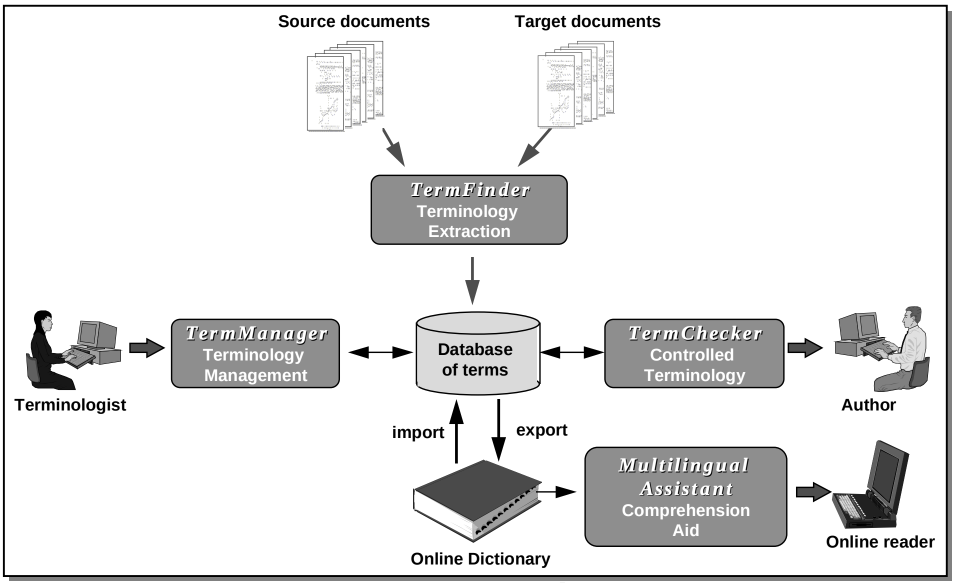
In addition, Danish, Swedish, Finnish, Norwegian, Czech, Hungarian, Russian, Romanian, Polish, Arabic, Japanese, Korean are under development.

Built on top of Open Database Connectivity (ODBC), the database independent layer from Microsoft, TermFinder is independent from a specific database. TermFinder supports SGML, HTML, XML, iso-8859-1 and Rich Text Format documents.

2.4.1.2. TermManager : Terminology Database in Context

TermManager is the complement to TermFinder. It enables one to quickly manage the terminology that was created with TermFinder. One can modify it, add terms, remove others, and add specific information. The Term In Context view enables users to see all occurrences of a term in the context of the original sentences.

TermManager uses several views to display the terminology: Form View, to view all the information related to a term, Table View, to see information related to several terms, Dictionary view: to see terms that are related. One can define filters to see only a subset of the database. One can customise fonts, colours. One can create one's own fields to store user defined information.



XTRAS Terminology Suite

2.4.1.3. TermChecker : Controlled Terminology Tool

The terminology that has been built using TermFinder can then be used by TermChecker to provide authors with interactive feedback, to help them increase the terminology consistency. This tool can be used both by the author for the source terminology and by the translator for the target terminology.

TermChecker is fully integrated with word processors. It provides the same look and feel than the standard spell checker function.

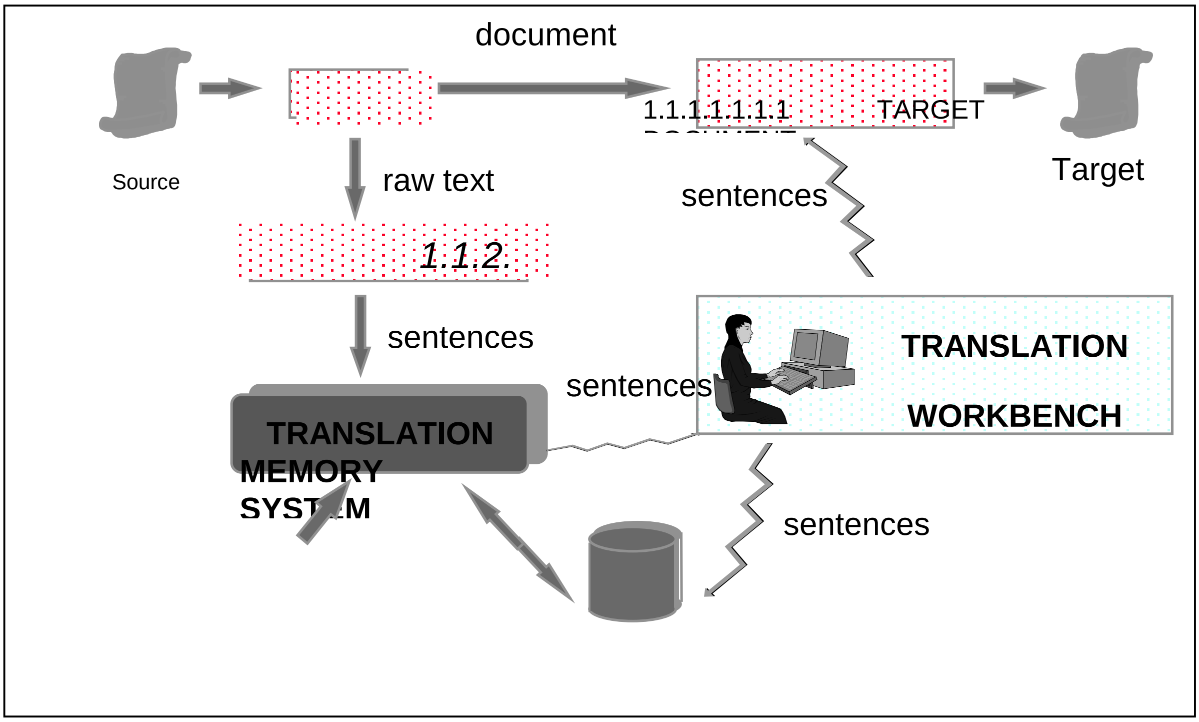
2.4.1.4. Multilingual Assistant : Comprehension Aid Tool

This Multilingual Assistant provides translation of words in context, using a general or specialised dictionary. It can differentiate between similar expressions that should be translated differently ("apply to" vs. "apply *something* to"). The Multilingual Assistant is based on the results of the Locolex project described above.

2.4.2. XTRAS Translation memory

Translation memory helps the translation process by recognising previously translated texts: the system "keeps" sentences that have been previously translated, with their corresponding translation. When a new document has to be translated, or an updated version of an existing document, the translation memory can rapidly find identical or similar sentences and retrieve them for the translator to view. This will save any unnecessary duplication of work for the translator whilst increasing consistency and quality of translations. By cutting down on the repetitious and routine work, Translation Memory frees up the translator to focus on new texts and thereby reduce the overall time and cost of translation.

How does it work?



XTRAS Translation Memory Overview

The Filter: A filter receives the source document to be translated which it parses, extracting information about the structure, such as titles, styles, paragraph marks etc. The process simultaneously extracts the text itself, plus some additional formatting, such as character style, (bold, italic, underlined…) in order to store as much data as possible to reduce the efforts of the human translator. This format information is stored independently from the format of the input document and so can relate to parts of text as well as the whole text. Additional data can be added such as page numbers, document identification etc. etc.The filter can read the most well known document formats (RTF SGML HTML MIF Interleaf) and in this way is word processor independent. The filter reads character codes in English, French, German, Italian, Spanish, Portuguese and Dutch for the source documents. An indefinite number of target languages can be supported when written in Unicode characters.

Segmentation: The input text is split up into units of translation which are to be stored in the translation memory database, normally consisting of whole sentences and their formatting. This formatting is copied to the output sentences without any modifications. However, other pieces of text may be considered as translation units, such as titles, lists, figures, captions etc. and stored accordingly. A list of abbreviations is maintained to enable proper recognition by the user, for example to avoid interpreting every occurrence of a period as the end of a sentence. This list can be extended and modified

Translation Memory System: it performs several functions:

- Manages the translation memory database (storage, administration, import/export)
- Processes the source sentences by retrieving them from the translation memory and/or by retrieving similar sentences
- Retrieves the translation which has been stored for matching sentences (perfect matching) and, in the case of non-identical sentences (fuzzy matching or no match), generating a close translation.

Storage and Administration: Documents to be translated are grouped together to form projects and assigned a manager who will define the characteristics of that project, by domain, customer, source language and target language for example. The manager can add/remove texts to/from the project, delete them, file them and merge two translation memories if required. The database for storage is computationally efficient and can maintain a large amount of information using a minimum of resources. The database holds pairs of sentences, (source and target) containing the following history: the source of the sentence, the source and target languages of the sentence, the number of times the sentence occurs, when the sentence was written and by whom and the last time the sentence was accessed and by whom. The sentences will also carry their original format

As the storage facility can show these details, the project manager will have no trouble in editing and cleaning up texts.

- Import: Various data sources (text files) can be fed into the translation memory, including other translation memory systems for example Trados, IBM TM/2, bilingual dictionaries (by extracting translations).
- Export: The data from the translation memory can be moved to a file of text which contains aligned sentences and to documents using other translation memory systems.

Search and Retrieval: Input for translation memory consists of sentences with some formatting information. Searches for these sentences can take place in more than one translation memory and can be defined and prioritised by the user, to obtain the best matches first. Any differences between the input sentence and the matching sentence are taken into account by the system and include:

- formatting differences; some characters do not have the same style
- case differences
- punctuation differences
- words are substituted; changes in proper nouns, acronyms, numbers
- linguistic differences; one word has the same base form but not the same surface form - number, tense, gender
- insertion or deletion of one or more words; secondary words (adverbs and adjectives) are different but the main words (verbs) are the same
- changes in the order of phrases
- changes in the order of words

Generation of Translation: If there is a difference between the match and the searched sentence, the aim is to find the closest possible target sentence and so minimise the work of the translator. Translation memory can generate such a modified match if the difference is small, for example relating to punctuation, case or number.

The Translator's Workbench: The workbench is the store for sentences and their matches. It allows the translator to translate sentences that have not been found and to verify matches (perfect and fuzzy) that have been found in the translation memory. The workbench can take information from several translators and merge information from several documents. It provides a graphical interface which displays as much information as possible to help the translator work quickly and efficiently.

3. Selected references

Aït-Mokhtar, Salah and Chanod, Jean-Pierre (1997a): "Incremental finite-state parsing", in *Proceedings of Applied Natural Language Processing 1997*, Washington, DC.

Aït-Mokhtar, Salah and Chanod, Jean-Pierre (1997b): "Subject and Object Dependency Extraction Using Finite-State Transducers", *ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

Bauer, D., Segond, F. and Zaenen, A. (1995): "LOCOLEX: the translation rolls off your tongue." in *Proceedings of the ACH-ALLC conference*, Santa Barbara, pp. 6-8.

Chanod, Jean-Pierre, Tapanainen, Pasi (1995): "Tagging French -- comparing a statistical and a constraint-based method" in *Seventh Conference of the European Chapter of the ACL*. Dublin.

Grefenstette, Gregory (1994): *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston.

Grefenstette, Gregory, Heid, Ulrich and Fontenelle, Thierry (1996): "The DECIDE project: Multilingual Collocation Extraction." *Seventh Euralex International Congress*, University of Gothenburg, Sweden, Aug 13-18, 1996.

Hladka, Barbara and Hajic, Jan (1997): "Probabilistic and Rule-based Tagger of an Inflective Language" In *Proceedings of Applied Natural Language Processing 1997* Washington, DC.

Kaplan, Ronald M. and Kay, Martin (1994): "Regular Models of Phonological Rule Systems". *Computational Linguistics*, 20:3 331-378.

Karttunen, Lauri (1994): "Constructing Lexical Transducers". In *Proceedings of the 15th International Conference on Computational Linguistics*, Coling, Kyoto, Japan.

Karttunen, Lauri (1995): "The Replace Operator". *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (ACL-95) 16-23.

Koskenniemi, Kimmo (1983): "A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics". University of Helsinki.

Kupiec, Julian and Wilkens, Mike (1994): *The dds tagger guide version 1.1*. Technical report, Xerox Palo Alto Research Center.

Maxwell, III, John T. and Kaplan, Ronald M. (1991): “A method for disjunctive constraint satisfaction.” In Tomita, Masaru (ed.), *Current Issues in Parsing Technology*. Kluwer Academic Publishers, Dordrecht, pp.173-190.

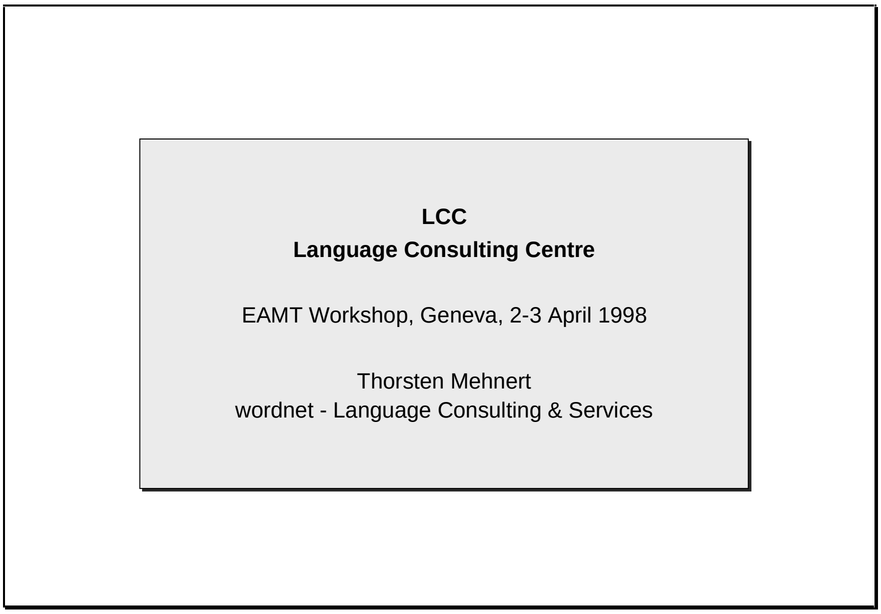
Nerbonne, John, Karttunen, Lauri, Paskaleva, Elena, Proszeky, Gabor and Roosmaa, Tiit (1997): “Reading more into Foreign Languages”. In *Proceedings of Applied Natural Language Processing* 1997 Washington, DC.

Schiller, Ann (1996): “Multilingual Finite-State Noun Phrase Extraction.” In: ECAI '96 workshop on "Extended finite state models of language", Budapest.

Segond, F. and Tapanainen, P. (1995): *Using a finite-state based formalism to identify and generate multiword expressions*. Technical Report MLTT-019, Xerox Research Centre, Grenoble, 1995.

LCC – Language Consulting Centre

Thorsten Mehnert

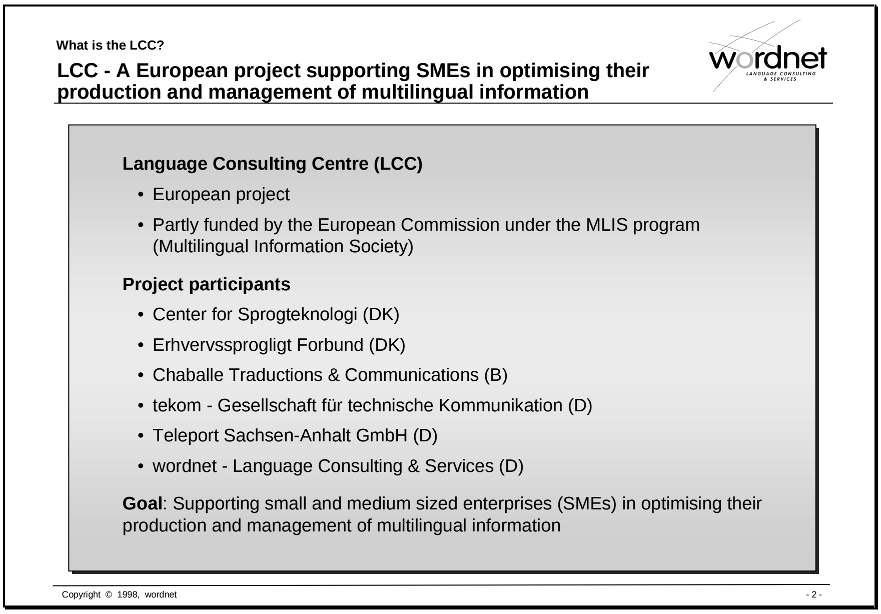


LCC

Language Consulting Centre

EAMT Workshop, Geneva, 2-3 April 1998

Thorsten Mehnert
wordnet - Language Consulting & Services



What is the LCC?

LCC - A European project supporting SMEs in optimising their production and management of multilingual information

Language Consulting Centre (LCC)

- European project
- Partly funded by the European Commission under the MLIS program (Multilingual Information Society)

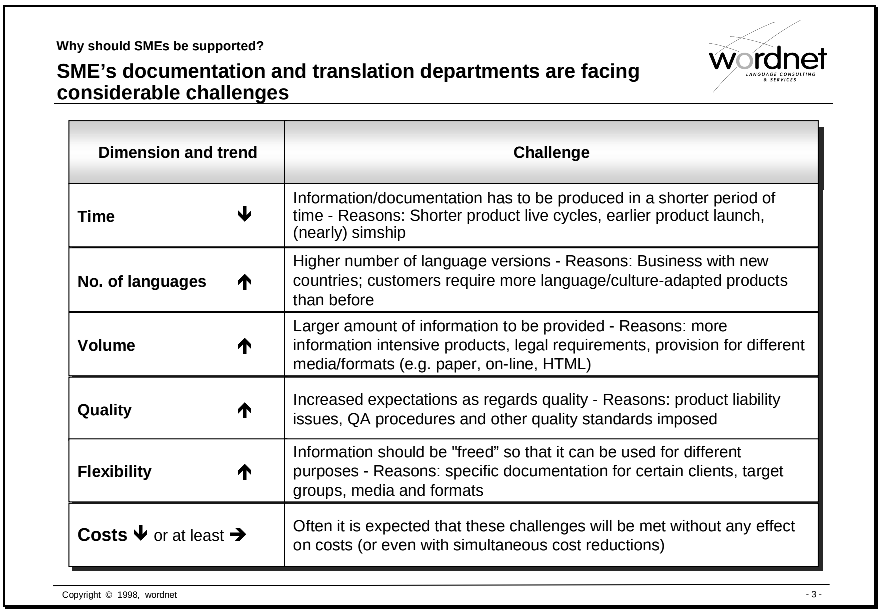
Project participants

- Center for Sprogteknologi (DK)
- Erhvervssprogligt Forbund (DK)
- Chaballe Traductions & Communications (B)
- tekom - Gesellschaft für technische Kommunikation (D)
- Teleport Sachsen-Anhalt GmbH (D)
- wordnet - Language Consulting & Services (D)

Goal: Supporting small and medium sized enterprises (SMEs) in optimising their production and management of multilingual information

Copyright © 1998, wordnet

- 2 -

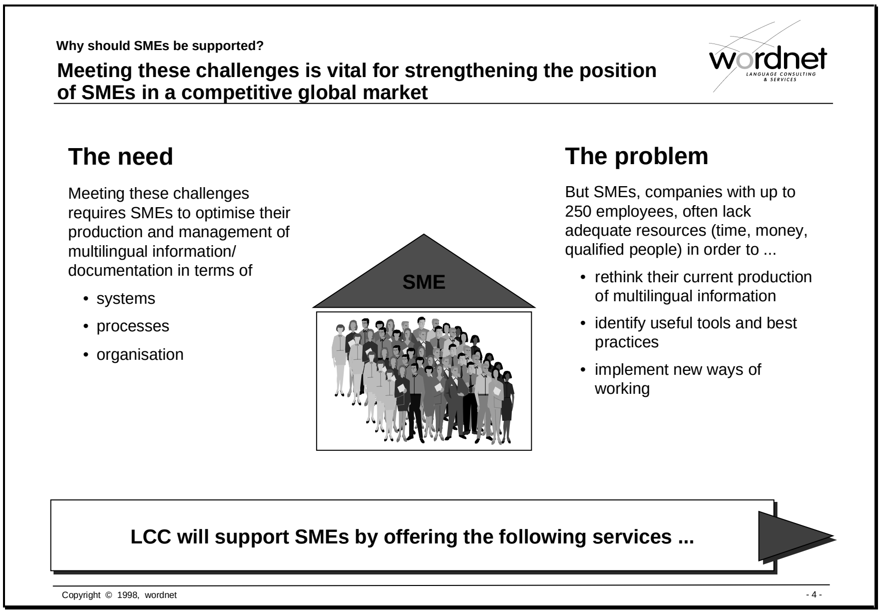


Why should SMEs be supported?

SME's documentation and translation departments are facing considerable challenges

wordnet LANGUAGE CONSULTING & SERVICES

Dimension and trend	Challenge
Time ↓	Information/documentation has to be produced in a shorter period of time - Reasons: Shorter product live cycles, earlier product launch, (nearly) simship
No. of languages ↑	Higher number of language versions - Reasons: Business with new countries; customers require more language/culture-adapted products than before
Volume ↑	Larger amount of information to be provided - Reasons: more information intensive products, legal requirements, provision for different media/formats (e.g. paper, on-line, HTML)
Quality ↑	Increased expectations as regards quality - Reasons: product liability issues, QA procedures and other quality standards imposed
Flexibility ↑	Information should be "freed" so that it can be used for different purposes - Reasons: specific documentation for certain clients, target groups, media and formats
Costs ↓ or at least →	Often it is expected that these challenges will be met without any effect on costs (or even with simultaneous cost reductions)



Why should SMEs be supported?

Meeting these challenges is vital for strengthening the position of SMEs in a competitive global market

wordnet LANGUAGE CONSULTING & SERVICES

The need

Meeting these challenges requires SMEs to optimise their production and management of multilingual information/ documentation in terms of

- systems
- processes
- organisation

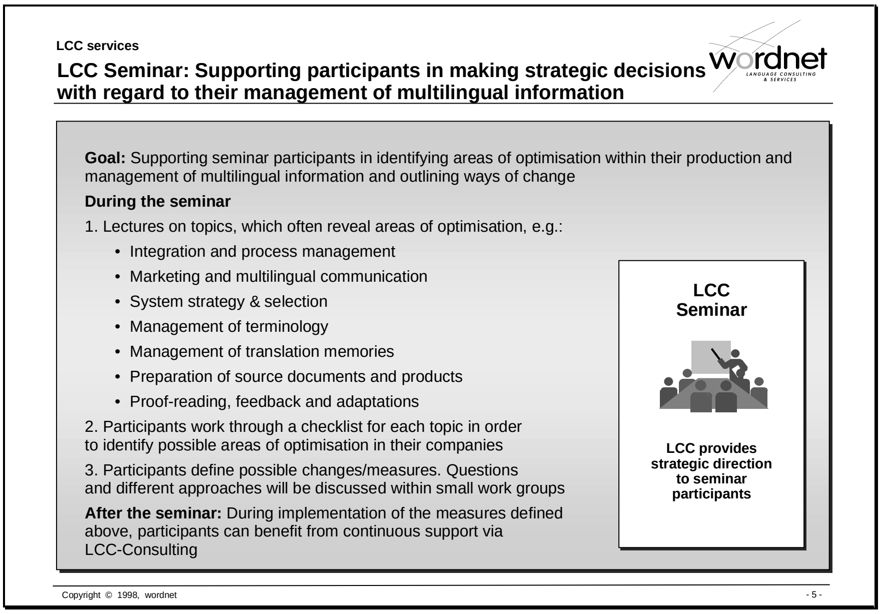
SME

The problem

But SMEs, companies with up to 250 employees, often lack adequate resources (time, money, qualified people) in order to ...

- rethink their current production of multilingual information
- identify useful tools and best practices
- implement new ways of working

LCC will support SMEs by offering the following services ...



LCC services

LCC Seminar: Supporting participants in making strategic decisions with regard to their management of multilingual information

Goal: Supporting seminar participants in identifying areas of optimisation within their production and management of multilingual information and outlining ways of change

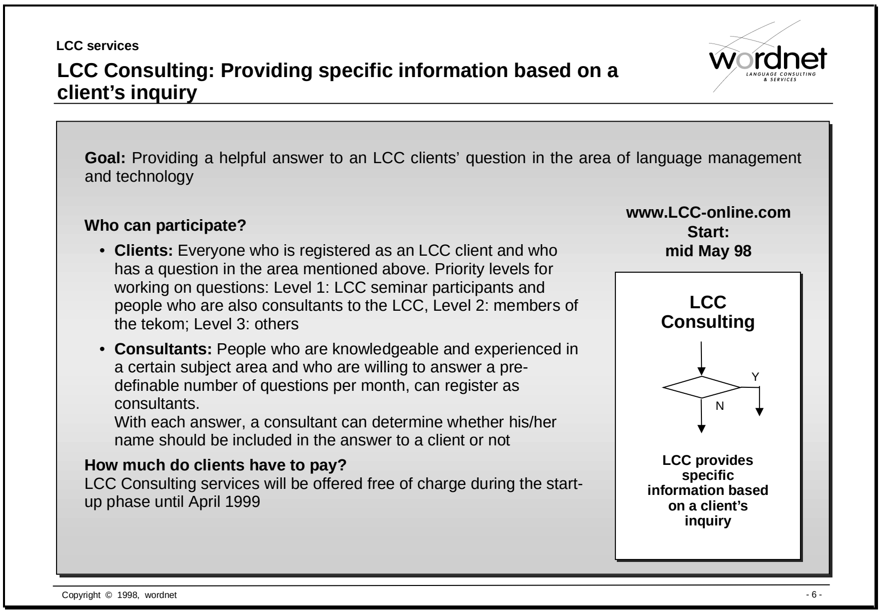
During the seminar

1. Lectures on topics, which often reveal areas of optimisation, e.g.:
   - Integration and process management
   - Marketing and multilingual communication
   - System strategy & selection
   - Management of terminology
   - Management of translation memories
   - Preparation of source documents and products
   - Proof-reading, feedback and adaptations
2. Participants work through a checklist for each topic in order to identify possible areas of optimisation in their companies
3. Participants define possible changes/measures. Questions and different approaches will be discussed within small work groups

After the seminar: During implementation of the measures defined above, participants can benefit from continuous support via LCC-Consulting

LCC Seminar

LCC provides strategic direction to seminar participants



LCC services

LCC Consulting: Providing specific information based on a client's inquiry

Goal: Providing a helpful answer to an LCC clients' question in the area of language management and technology

www.LCC-online.com
Start:
mid May 98

Who can participate?

- **Clients:** Everyone who is registered as an LCC client and who has a question in the area mentioned above. Priority levels for working on questions: Level 1: LCC seminar participants and people who are also consultants to the LCC, Level 2: members of the tekom; Level 3: others
- **Consultants:** People who are knowledgeable and experienced in a certain subject area and who are willing to answer a pre-definable number of questions per month, can register as consultants.
  With each answer, a consultant can determine whether his/her name should be included in the answer to a client or not

How much do clients have to pay?
LCC Consulting services will be offered free of charge during the start-up phase until April 1999

LCC Consulting

LCC provides specific information based on a client's inquiry

LCC services

LCC Consulting: Providing specific information based on a client’s inquiry (cont.)

Some examples for possible questions

- „I am looking for a system which can do the following ... Does such a system exist on the market and where can I find it?“
- „In the seminar I decided on several measure in order to better integrate the documentation activities of our Spanish subsidiary. Here is what happened ... What didn‘t work was that ... What would you suggest as a next step?“
- „ISO 9000 is about to be introduced in our company. I would like to prepare myself for it and I have the following questions ...“
- „I want to get in contact with people who have experience with the system support for tool X. Can you get me in contact with someone?“

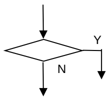
LCC Consulting workflow

- An inquiry will be routed to the LCC project manager who supports the first language of the client
- The project manager will either be able to give an answer him/herself or he/she will reformulate/route the question to another consultant selected from the LCC database

Questions can be written in the following languages
Danish, English, German and French

www.LCC-online.com
Start:
mid May 98

LCC
Consulting



LCC provides specific information based on a client’s inquiry

LCC services

LCC Knowledge Base: Providing general information on language management and technology topics

Goal: Providing LCC clients and guests with the possibility of finding an answer to their question by navigating through the LCC Knowledge Base

Navigation means

- Table of contents, search functions, frequently asked questions
- In addition, clients can receive periodic update information on new or modified Knowledge Base items

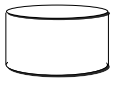
Content

The KB will include information which is of general interest to LCC clients, e.g. answers to the following questions:

- What types of systems exist for the production and management of multilingual information?
- Which steps are recommended to select a system?
- Which translator workbenches (TW) are on the market?
- How to evaluate different TWs with regard to our needs?
- How to exchange translation databases with others who have a different or no translation memory system?
- How to author documents for efficient translation?
- ...

www.LCC-online.com
Start:
end of April 98

LCC
Knowledge Base



LCC provides answers to frequently asked questions, navigation support and information on updates

Summary of the concluding discussion

Dimitri Theologitis and Bente Maegaard

During the Workshop, several issues came up that could not be followed fully during the question and answer part of the presentations. In the concluding discussion, the participants were asked to raise any question that they felt should be dealt with in more depth. Of the subjects proposed some were discussed in more depth and the conclusions are summarised below.

Profiles of human resources

A large translation task involving linguistic technology ideally needs persons with several specialisations. The following were identified:

A **Project Manager** controls the project throughout its lifetime and performs the more or less standard management tasks, including human resources management.

For the design of the project, a **Language Engineer** assumes the role of consultant. He or she analyses the information flows and decides on the type of technology to use, the integration of the various types of tools, the workflow of the project, the necessary human resources and the timings involved. This should be a person with experience in computer translation projects, ideally someone with qualifications in computer technology and with a profound understanding of what translation involves.

A relatively new idea was the inclusion of a **Linguistic Assistant** in the translation process. There are enough tasks, such as text alignment, preparation and management of documents and linguistic resources, batch terminological research, reference document search and post-processing, that need not necessarily be performed by a highly paid translator. These tasks can be dealt with by linguistically sensitive secretaries, for whom the resulting variety would add to their job satisfaction.

Finally, a **Translation Technician** would take care of the more computer-oriented tasks such as the setting up and managing of networks of multilingually enabled computers and the managing of multilingual databases. This person would perform also the function of a general technical assistant whenever there are a sufficient number of translators to make it feasible.

Obviously, the **Translators** themselves take centre stage. Ideally they are involved only after all necessary preparation has been done for them by the support staff. Like the star conductor, they only appear for the final performance once the orchestra has been trained by their assistants.

In practice however, it is rare for projects to be of such size to make feasible the deployment of such a variety of specialised personnel. More likely, the translators themselves need to take on all these roles and become multi-talented in areas that have little directly to do with translation. These new functions are getting added to the job descriptions of translators. And they still need to be good translators…

The Language Assistant could, however, represent a profession with a good future, if one is to judge by the increasing use of translation technology. Much as dentists and doctors employ specially-trained Medical Assistants and lawyers employ specially-trained Legal Assistants, one may envisage that in the same way Language Assistants could very well find satisafctory employment in large translation agencies.

Education in translation technology

Most translators acquire computer technology skills only after they leave university. However, the great increase in the use of computer aids for translation creates a need for training in these tools already as part of the standard translator education. This should be compulsory not only for translators, but also for technical authors, managers of translation projects and, last but not least, customers with large translation requirements.

Universities need to adapt their curricula to this changing environment[5]. It is not yet, however, evident exactly what constitutes the basic, core training for the future professionals. Clearly, courses should include information about terminological databases, translation memories, machine translation – especially since it has moved to the desktop – and basic computer skills with an emphasis on multilingual text processing. Skills on managing technological change would also be very welcome.

Time on a course, however, is limited. To include new topics, the whole balance of translation studies will have to change. For the moment though, the awareness of universities needs to be stimulated – by articles, seminars and concrete feedback from the translation world. The EAMT also has a role to play in this area.

“Merely a management issue…”

To write off the difficulties inherent in the introduction of translation technology as purely a management issue is to understate the challenge. Most organisations tend to underestimate the effort necessary in re-engineering the processes of multilingual document production, and then, when technology is finally introduced, there is a tendency to ape the traditional working methods. This obviously prevents the exploitation of the full potential of the tools.

The eternal question is: should technology dictate the process or should technology follow a given process? The question here is not as philosophical as it might seem. Ten years ago the revolution that translation memories brought to multilingual document production could not even be imagined. Now, companies design their workflow around the technological capabilities of the tools they have chosen.

[5] An attempt at a solution is the LETRAC project, funded in part by the European Commission’s Language Engineering programme, in which nine European universities and organisations try to define a new curriculum for translation studies by means of questionnaires and input from high-level translation professionals.

The fact that machines are rigid has often a beneficial effect on the working methods of humans. Individual translator choices, with otherwise erratic effects on document quality and productivity, tend to be replaced by a streamlined translation workflow, with gains for the final product in terms of quality and faster delivery time.

These changes, however, need to be integrated into the working methods of the whole organisation. In the place of translation as a "necessary evil" multilingualism comes as a business opportunity. The attitude of senior managers needs to change accordingly, but once they are convinced it is usually easy to integrate the necessary changes within the whole organisation.

Translation metrics

In the short time available in the discussion session the subject could not be adequately treated. However, a major concern was the issue of what measures and criteria should be employed in order to decide which technology would be best for the treatment of particular types of documents and texts, including information about the available linguistic resources[6]. Translation metrics in this context was understood to include costs, investments, productivity, but mainly metrics of text and linguistic resources as decision aids.

Productivity is not just about number of pages per day but even more the quotient of the total input and the value of the end product. Investment in technology versus gains in productivity would seem to form a curve with an asymptote, meaning that after a while benefits start to flatten out.

The cost of translation can be measured fairly easily. However, no realistic estimates can be readily established as to the cost of *not* translating! This holds especially true for large political organisations such as the United Nations or the European Commission.

Future EAMT Workshops

Finally, suggestions were made for subjects and for the organisation of future EAMT Workshops. These were:

- Is the use of machine translation declining? What about the balance between use of translation memories and machine translation?

- Should the EAMT promote any particular type of technology?

- What are or should be the visions, the dreams of translators? The question is put in a pro-active way which should allow the focusing of technological developments. What should be coming next?

- Motivation and involvement in the use of translation technology.

[6] Another European project, TRANSROUTER, deals with exactly this issue. Based on data about the text, repetitivity, syntactic complexity, the availability of terminology, translation memories and the performance of the available machine translation system, an expert system would suggest the optimum tools to use in each particular case.

- Languages of the countries of Central and Eastern Europe. The expansion of the European Union and the opening of new markets will create new multilingual needs. What is the position of the European Union?

- The impact of the internet, particularly in respect to the provision and pricing structure of machine translation services.

- Development of exchange standards for terminology, translation memories and their integration in machine translation. Integration of tools and multilingual resources.

- Metrics as decision aids for translation technology. Expert systems (see above).

- Integration of speech recognition in translator's workbenches.

- Multilingual information management.

- Intellectual property rights. Liabilities.

- Suggestion for future workshops: the inclusion of one hour of short commercial presentations. However, it was felt that the present scheme with fewer but more in-depth papers and time for discussion is good.

Next Workshop

The next EAMT Workshop is scheduled for Spring 1999 in Prague. It will focus on the languages of the Central and Eastern Europe, challenges and opportunities.

Company information

Océ Technologies is a manufacturer of copiers, printers, plotters, design & engineering equipment and supplies. It has operating companies in 30 countries and is active in 80 countries. Océ employs 17000 people worldwide; 3000 are based at the head office in Venlo, the Netherlands**.**

SAP AG was founded in 1972 and today is the world leader in integrated business solutions with the two products R/2 and R/3 instaled in more than 80 countries. It is one of the four biggest software companies in the world with more than 10,000 employees and represented in over 40 countries. The Translation Department has a staff of 75 English translators (+15 freelancers) and 30 translators (+35 freelancers and translation bureaus) for the various other languages, and with well over 70 technical writers being involved in writing documentation. SAP reached the point some years ago when the demand for translation of its documentation could not be met by human translators alone. Thus, the Multilingual Technology Department was set up to suppport further aspects of the translation workflow, to investigate new translation tools (e.g. translation memories), to introduce new language pairs in machine translation, and to provide internal technical support for the company.

The **Translation Service of the European Commission** is situated in Brussels and Luxembourg. The SdT houses today about 1 200 translators, 100 linguistic support staff, 100 management staff as well as 500 secretaries and assistants. The world's largest translation service, it produces about one million pages per year, in a combination of in-house, free-lance and machine production. Currently undergoing major technological modernisation. The main translation aids in use, being installed, or overhauled, include: the EURODICAUTOM terminological database; the EURAMIS Linguistic Resources Database and search engines combined with Translator's Workbench and other linguistic applications; the SYSTRAN machine translation system; and a document server with full-text search and retrieval possibilities.

The **Center for Sprogteknologi**, CST, is a research centre under the Danish Ministry of Research and Information Technology. The Centre was established in 1991 with the purpose of promoting research and development in computational linguistics and language technology. CST has some 20 employees with expertise in machine translation, general and computational linguistics, computational lexicography, computer science and Danish and a number of other languages. The Centre participates in European and national research programmes, and performs commercial development and consultancy under contracts with Danish as well as foreign companies.

The **Xerox Research Centre Europe** (XRCE) was formed in 1992 and was originally named the Rank Xerox Research Centre (RXRC). The Centre comprises three organisations located at two sites, Grenoble in France and Cambridge, England. Both sites have research laboratories. In addition, at Grenoble there is a development group (Advanced Technology and Systems). Research in the Grenoble Laboratory focuses on enabling technologies to support innovative document technologies and services both within and between geographically distributed and culturally diverse organisations. The two main areas of interest are Multi Lingual Theory and Technology and Co-ordination Technologies. Advanced Technology and Services draws on each Laboratory as well as Xerox and the external technical community to integrate novel systems and solutions. Since its inauguration, the Centre has grown rapidly and now comprises approximately 90 people, 60 in Grenoble and 30 in Cambridge. Teams at both Grenoble and Cambridge have been very successful in exploring commercialisation opportunities for the Centre's technologies. The MLTT team has led the way through the transfer of its technologies to a new business (Xtras) created within the Document Services Group. It has also been successful in transferring technology for inclusion in products and services from InXsight, XBS and Xerox UK, while within ATS a number of successful joint ventures are underway.

Addresses of speakers, session chairs and organizers

Achim Blatt, Translation Service, Development of Multilingual Computer Aids, CCE JMO B2/21, L-2920 Luxembourg (Tel: +352 4301 33632; Fax: +352 4301 34069; Email: achim.blatt@sdt.cec.be)

Achim Blatt holds a degree in translating from Saarbrücken University, post-graduate studies in translation theory, English philology and information theory, doctoral thesis on the computational treatment of English noun phrases. He was employed at Saarbrücken University from 1980 to 1989; since then, he has been working for the European Commission. His main areas of expertise are machine translation (work on SUSY, Eurotra, Systran) and other NLP products (translation memories, sentence alignment etc.). He currently manages the EURAMIS project (a client-server application which gives access to number of services in the domain of natural language processing).

Michael S.Blekhman, Director, Lingvistica ’93 Co., Vice President, POLYGLOSSUM, Inc., 94a Prospekt Gagarina, apt. 111, Kharkov 310140 Ukraine (Email: blekhman@lotus.kpi.kharkov.ua; or: super@pc101.ai.kharkov.ua)

Dr. Blekhman has been involved in the MT industry for many years, setting up his own company in 1993 to market and develop the PARS system.

Colin Brace, Language Industry Monitor, Eerste Helmerstraat 183, NL-1054 DT Amsterdam, The Netherlands (Tel: +31 20 6850462; Fax: +31 20 6854300; Email: cbrace@lim.nl)

Colin Brace is a writer and consultant specializing in language technology. Since 1991, he has published *Language Industry Monitor*, a newsletter covering developments in the linguistic engineering world. Since November 1997 he has also been joint editor (with Muriel Vasconcellos) of *MT News International*, the newsletter of the International association for Machine Translation. He has written regularly for other computer- and translation-oriented publications, and has given presentations at a variety of conferences. Mr. Brace has also participated in several EU-funded projects.

Jean-Pierre Chanod, Project leader, Multi-Lingual Theory and Technology (MLTT) group; Manager Language Resources Group**,** Grenoble Laboratory, Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France (Tel: +33 4 76 615076; Fax: +33 4 76 615039) Email: Chanod@grenoble.rxrc.xerox.com

Lou Cremers, Océ Technologies, ITC-Translation Services, St. Urbanusweg 43, 5900AM Venlo, The Netherlands (E-mail: lcr@oce.nl; Tel: +31 77 3593444)

Viggo Hansen, Managing Director, Hofman-Bang A/S, Hans Bekkelunds Allé 7, DK-2900 Hellerup, Denmark (Tel: +45 394 88000; Fax: +45 394 88080; Email: vha@hofman-bang.dk)

Since 1995, Viggo Hansen has been managing director of Hofman-Bang A/S, one of the leading patent and trademark attorney companies in Scandinavia. In May 1993 he became manger of a newly-established company Lingtech A/S, a translation company owned by the two predecessor companies of Hofman-Bang, where he implemented the machine translation system PaTrans – a system developed by the Center for Sprogteknologi (Copenhagen) for translating patents.

John Hutchins, University of East Anglia, Norwich NR4 7TJ, U.K. (Tel: +44-1603-592429 or +44-1603-453941; Email: J.Hutchins@uea.ac.uk**,** or: WJHutchins@compuserve.com**)**

John Hutchins is president of the European Association for Machine Translation. He is the author of a number of articles surveying activities in the field, he has given presentations to many MT conferences since the late 1970s, and has written two books on the subject: a history published in 1986, and a textbook (jointly with Harold Somers) in 1992. From its foundation in 1992 until last year he was chief editor of *MT News International* (the newsletter of the International association for Machine Translation). At present he is involved in the compilation of a directory of MT systems.

Paul Kaeser, STAR AG, Wiesholz 35 , CH-8262 Ramsen, Switzerland. Email:pka@star-ag.ch; WWW: http://www.star-ag.ch (Tel. +41 52 742 92 19; Fax +41 52 742 92 92)

Maghi King, ISSCO, University of Geneva, 54 route des Acacias CH-1227 GENEVA (Switzerland) (E-mail: Margaret.King@issco.unige.ch; WWW: http://issco-www.unige.ch/; | Tel: +41/22/705 71 14; Fax: +41/22/300 10 86)

Pierre Lewalle, Computer-Assisted Translation and Terminology Unit, Room 3029, World Health Organization (Tel (41-22) 791 2317; Fax (41-22) 791 3995; Email: lewallep@who.ch)

Susan McCormick, Multilingual Technology Dept., SAP AG, Walldorf, Germany (Email: s.mccormick@sap-ag.de)

Bente Maegaard, Director, Center for Sprogteknologi, Njalsgade 80, DK-2300 Copenhagen S, Denmark (Tel: +45 35 32 90 74, Fax: +45 35 32 9089; Email: bente@cst.ku.dk)

Bente Maegaard holds a M.Sc. in Mathematics and French from the University of Copenhagen, 1970. She was employed at the University of Copenhagen, Department of Applied and Mathematical Linguistics, 1971-90, being a research professor 1984-89, and visiting professor at the University of Geneva 1981. She is currently director of the Center for Sprogteknologi (Centre for Language Technology) since its creation 1991. Her main areas of expertise are machine translation, evaluation methodology, dictionaries, corpora. She has held and holds positions as officer in scientific and other associations, member of editorial boards, reviewer for journals and conferences.

Doris Marty-Albisser, Corporate Language Services, Aeschenvorstadt 48, CH-4002 Basel (Tel: +41-61-288-9841; Fax: +41-61-288-3855; Email: doris.marty@cls.ch)

Doris Marty-Albisser is director of the Corporate Language Services, providing translation services for the major Swiss banks. She holds a degree in translation and an executive MBA. She has been in the translation and language technology business for the last 12 years.

Thorsten Mehnert, Senior Consultant, c/o wordnet - Language Consulting & Services, Schlehdornweg 19, D-35041 Dagobertshausen, Germany. (Email: tm.wordnet@scm.de, Tel: +49-6421-93006, Fax: +49-6421-93018)

Thorsten Mehnert studied computer science and business administration in Hamburg, Baton Rouge and Karlsruhe and holds a diploma (Dipl.-Inform.) from the University of Karlsruhe. After graduation he joined *Gruber, Titze und Partner* (later acquired by the US consultancy *Gemini Consulting*) where he worked as a management consultant on several reengineering projects for clients in telecommunications, the trade and public administration. Since 1996 he concentrates on information and process management in the translation and documentation industry.

Olivier Pasteur, Computer-Assisted Translation and Terminology Unit, Room 3029, World Health Organization (Tel (41-22) 791 2317; Fax (41-22) 791 3995; Email: pasteuro@who.ch)

Jörg Schütz, Institute for Applied Information Sciences (IAI), Martin-Luther-Strasse 14, D-6611 Saarbrücken, Germany. Fax: +49 (681) 389-5140. Email: joerg@iai.uni-sb.de

Dr. Jörg Schütz studied Computer Science, Mathematics and Medicine at the University of the Saarland from which he also received his doctoral degree (MT/CL and AI). Since 1985 he is working for the Institute of Applied Information Sciences (IAI) in Saarbrücken where he is responsible for the institute's R&D as director (the directorship he shares with Prof. Haller). He has been the project leader or supervisor of several research and development projects in the field of Natural Language Processing. He also teaches at the University of the Saarland and acts as a consultant for industrial companies. His current scientific interest is to combine Web technology and language technolgy (Networked MT), and in particular the performance control of such NLP systems.

Dimitri Theologitis, Translation Service, Development of Multilingual Computer Aids, CCE JMO B2/21, L-2920 Luxembourg (Tel: +352 4301 33632; Fax: +352 4301 34069; Email: dimitrios.theologitis@sdt.cec.be)

Born in Athens, civil engineer, specialised in integrated transportation systems and computers. Opted for a major change of career in 1984 when he joined the Translation Service of the European Commission. Responsible for the "Rationalisation of Working Methods" from 1990. In 1994 became head of unit "Development of Multilingual Computer Aids", a multilingual team active in the technological modernisation of the Translation Service.

Pim van der Eijk, Cap Gemini Advanced Technology Services. Email: pvdeijk@inetgate.capgemini.nl

Jacqueline van Wees, Cap Gemini Advanced Technology Services. Email: jwees@inetgate.capgemini.nl

What is the EAMT?

The European Association for Machine Translation (EAMT) is an organization serving the growing community of people interested in MT and translation tools, including users, developers and researchers of this increasingly viable technology.

The EAMT is one of three regional associations of the International Association for Machine Translation (IAMT), which counts a large number of members worldwide. The EAMT is the only organization of its kind in Europe.

MT News International
MT News International is published by EAMT in conjunction with the AMTA and the AAMT three times a year. It includes news of upcoming events, such as workshops and conferences, reports of previous events, company and product news, and updates of research developments.

Five reasons you should join the EAMT

1. Conferences and Workshops
Every two years, the IAMT organizes the MT Summit, a unique conference dedicated to the world of translation technology that alternates between Europe, North America, and Japan. In 1995, the MT Summit was held in Luxembourg; in 1997, it took place in San Diego, California; in 1999, it will be in Singapore; and in 2001, it will be held in Santiago de Compostelo, Spain. In years when the Summit is not being held in Europe, the EAMT organizes an annual workshop on some facet of MT, sometimes in conjunction with another event. EAMT members can attend EAMT- and IAMT-sponsored events at reduced rates.

2. Expertise and networking
The EAMT is an excellent way to learn more about the practical aspects of MT. MT is still an imperfect technology and its members have a lot of collective wisdom to share about the use of MT in working environments. Through the EAMT, you may be able to reach people who have "been there before."

3. MT New International
EAMT members receive MT News International three times a year. Whether you are researcher, developer, or (potential) user, MTNI is an excellent source of information on the world of MT.

4. Other publications
The IAMT regularly produces publications which are available to members at reduced rates, such as conference proceedings, a directory of MT systems and a "yellow-pages" of the members of the MT Community.

5. Bibliographic service
The EAMT offers a bibliographic service to its members. For a nominal amount, members can order photocopies of articles on MT from a wide range of publications in the EAMT archives.

For more information about EAMT see: http://www.lim.nl/eamt
EAMT Secretariat: TIM, 54 route des Acacias, CH-1227 Carouge (Geneva), Switzerland. Email: eamt@cst.ku.dk