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.

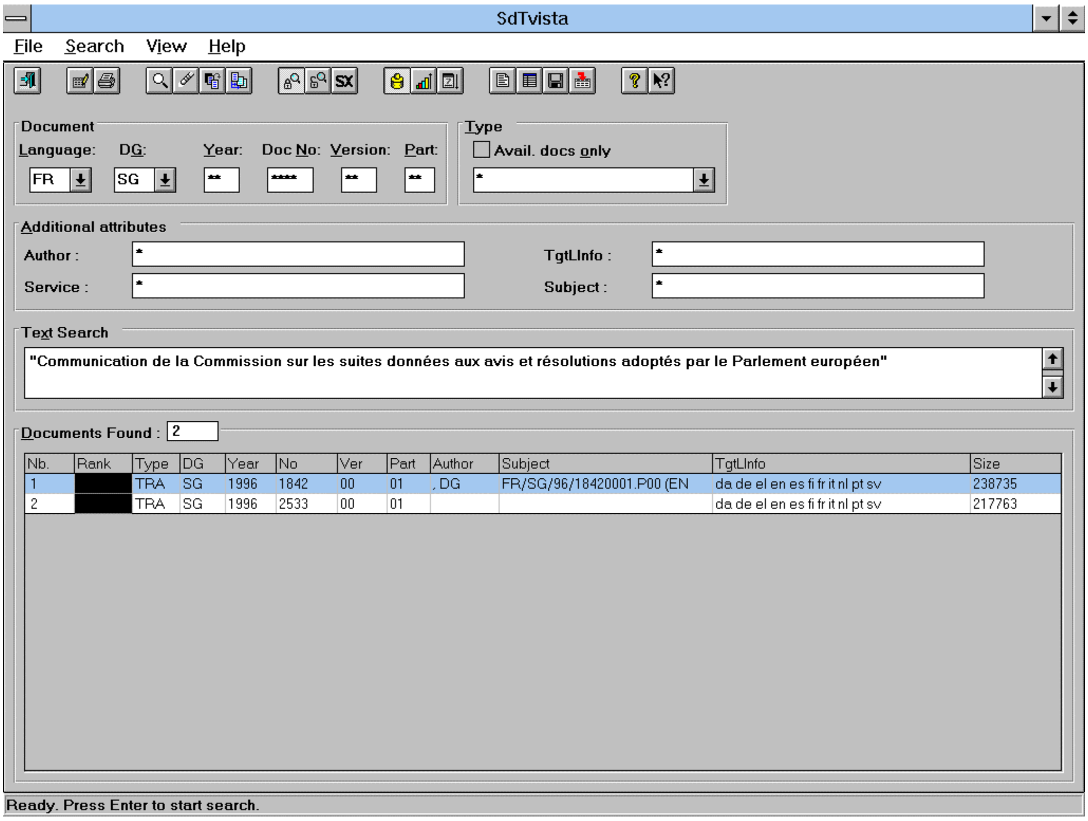
If one abstracts away from organisational aspects (which parts are done by secretaries, which parts are done by translators etc), a difference has to be made between the preparation of a translation, and the translation itself. In the worst case, a translation memory has to be created from zero.

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.):



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

□ Avail. docs only
*

Author : *
Service : *
TgtLInfo : *
Subject : *

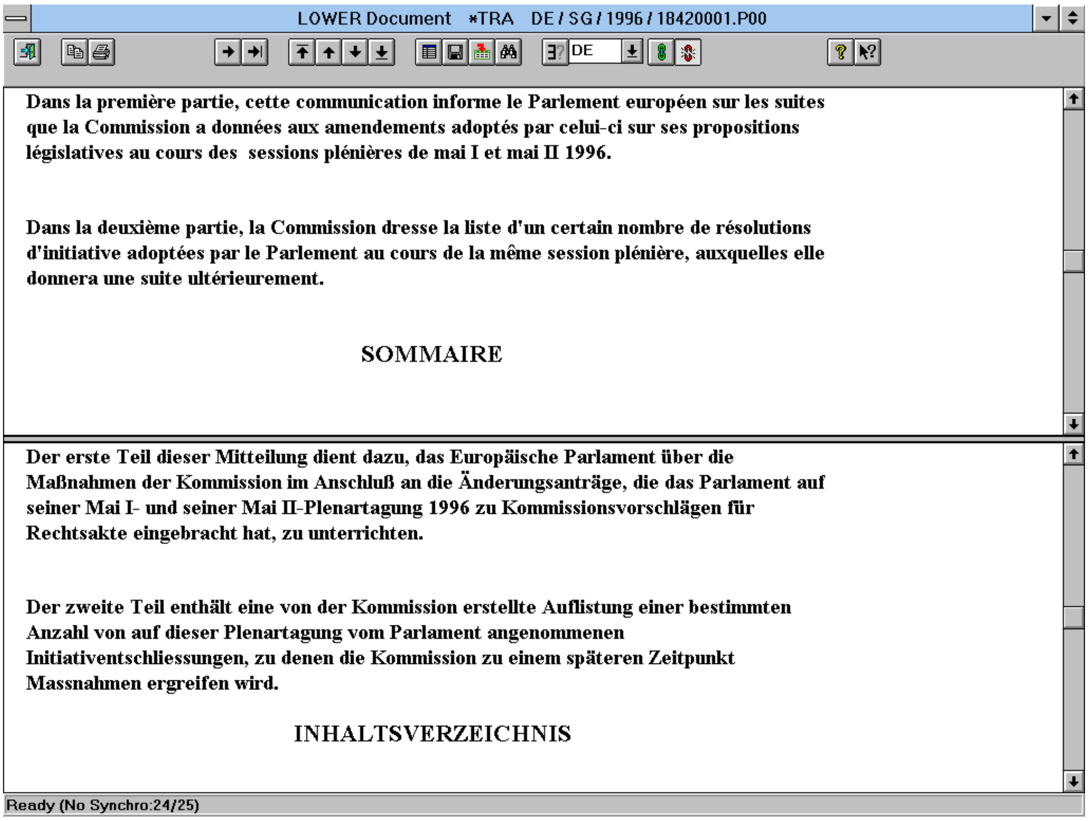
Text Search: "Communication de la Commission sur les suites données aux avis et résolutions adoptés par le Parlement européen"

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

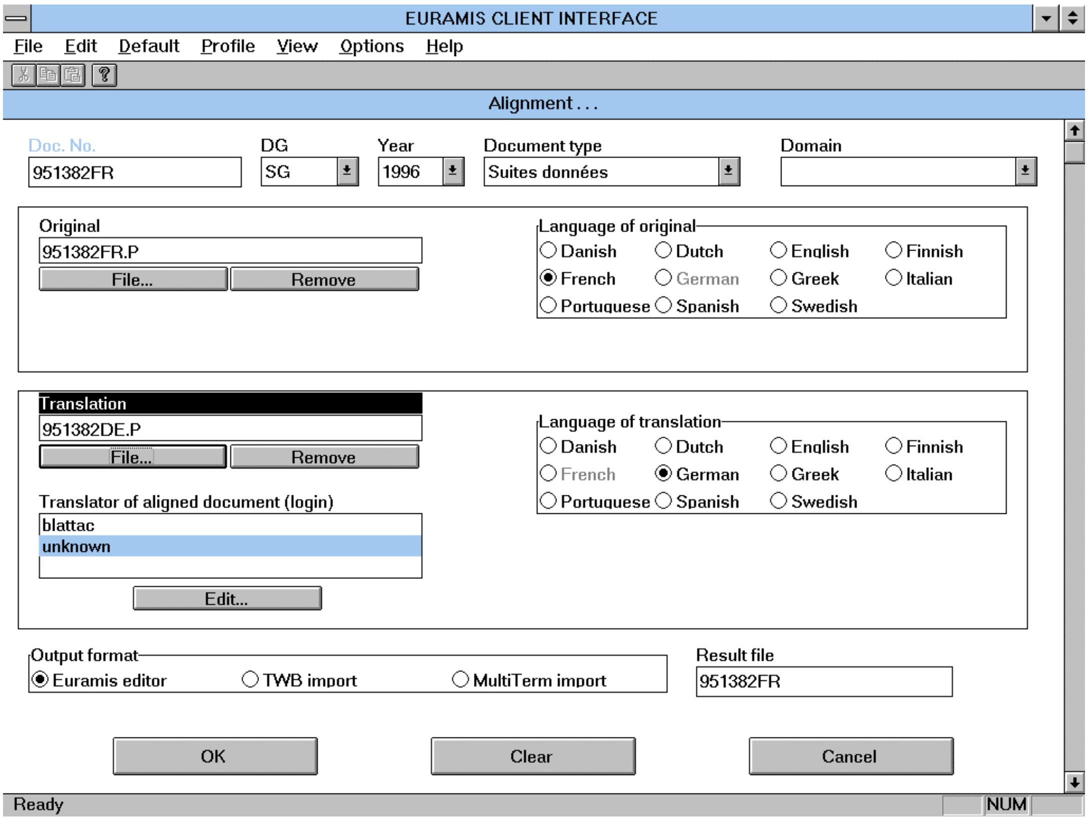
(2b) Regulation (EEC) No 210/69 of the Commission of 31 January 1969 on communications between Member States and the Commission with regard to milk and milk products

(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

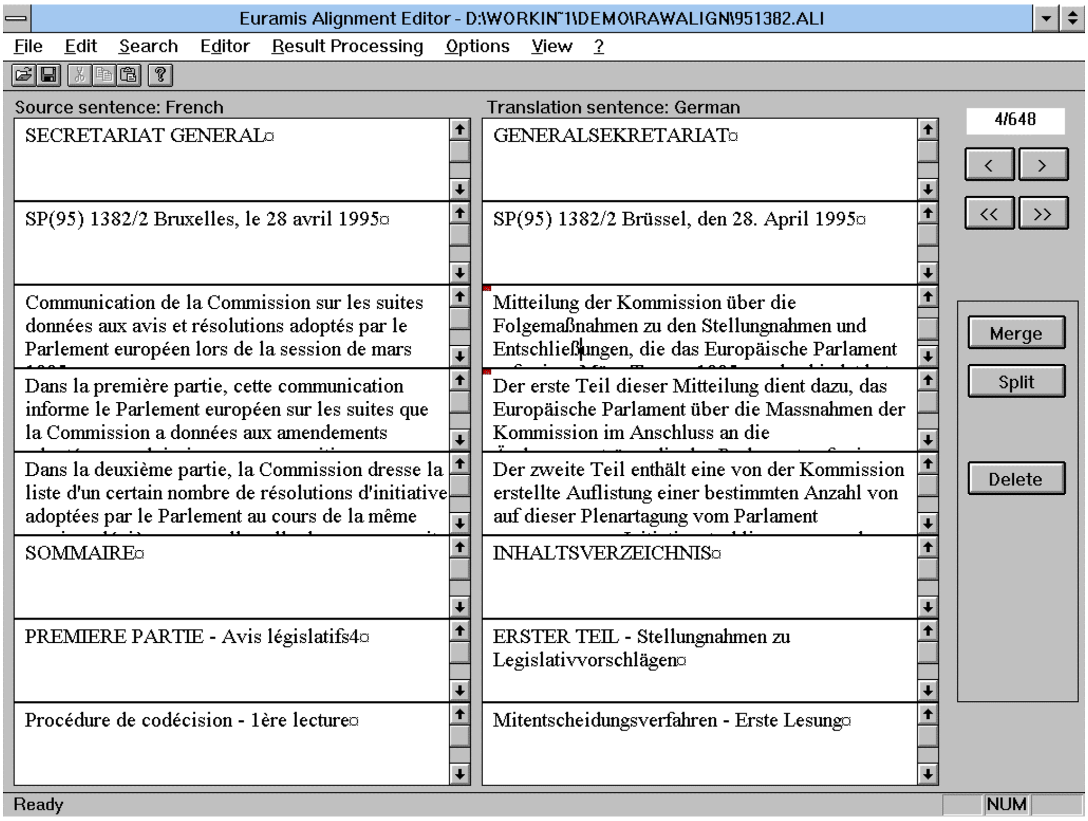
Result file: 951382FR

Ready

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:



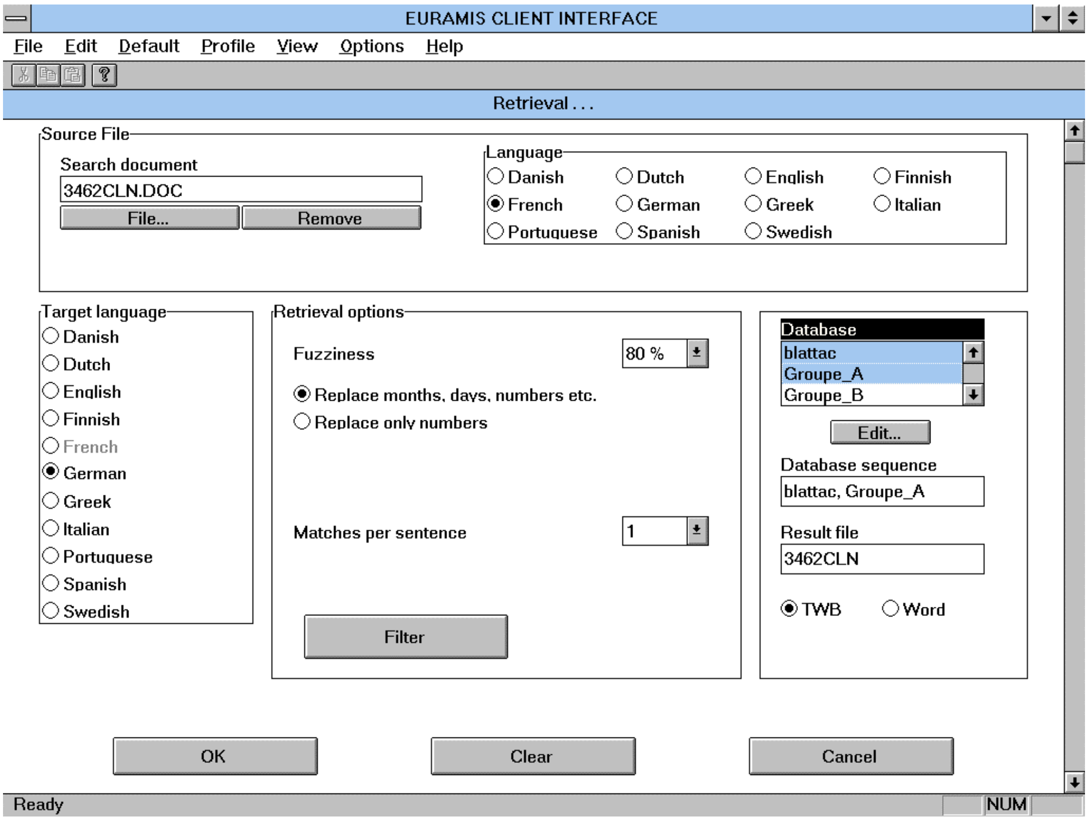
Source sentence: French	Translation sentence: German
SECRETARIAT GENERAL¤	GENERALSEKRETARIAT¤
SP(95) 1382/2 Bruxelles, le 28 avril 1995¤	SP(95) 1382/2 Brüssel, den 28. April 1995¤
Communication de la Commission sur les suites données aux avis et résolutions adoptés par le Parlement européen lors de la 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:



EURAMIS CLIENT INTERFACE

File Edit Default Profile View Options Help

Retrieval . . .

Source File

Search document: 3462CLN.DOC

File... Remove

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

Filter

Database: blattac, Groupe_A, Groupe_B

Edit...

Database sequence: blattac, Groupe_A

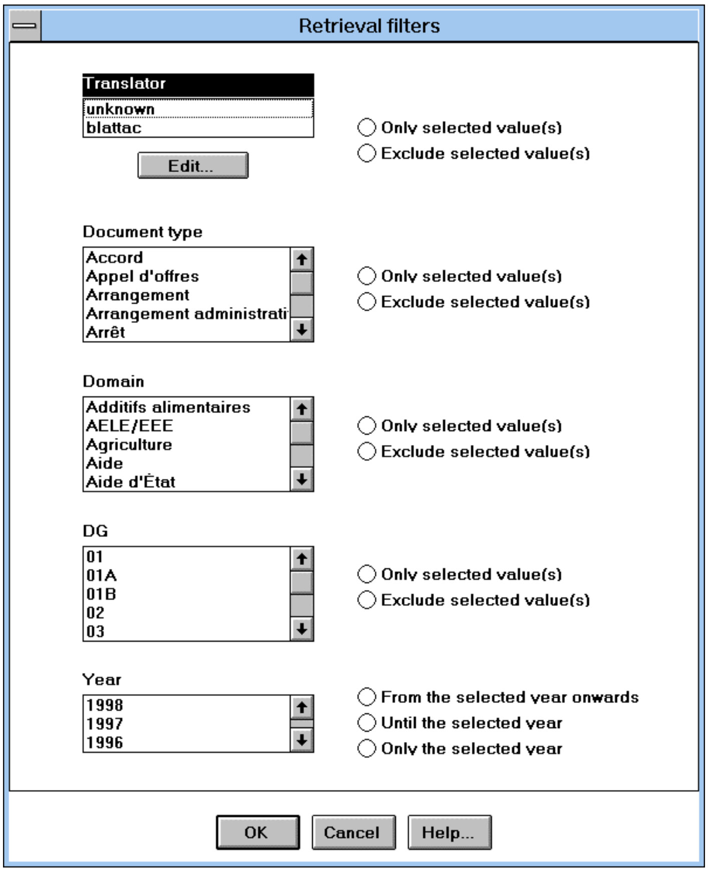
Result file: 3462CLN

● TWB ○ Word

OK Clear Cancel

Ready NUM

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).



Retrieval filters

Translator: unknown, blattac

Edit...

○ 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

OK Cancel Help...

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:

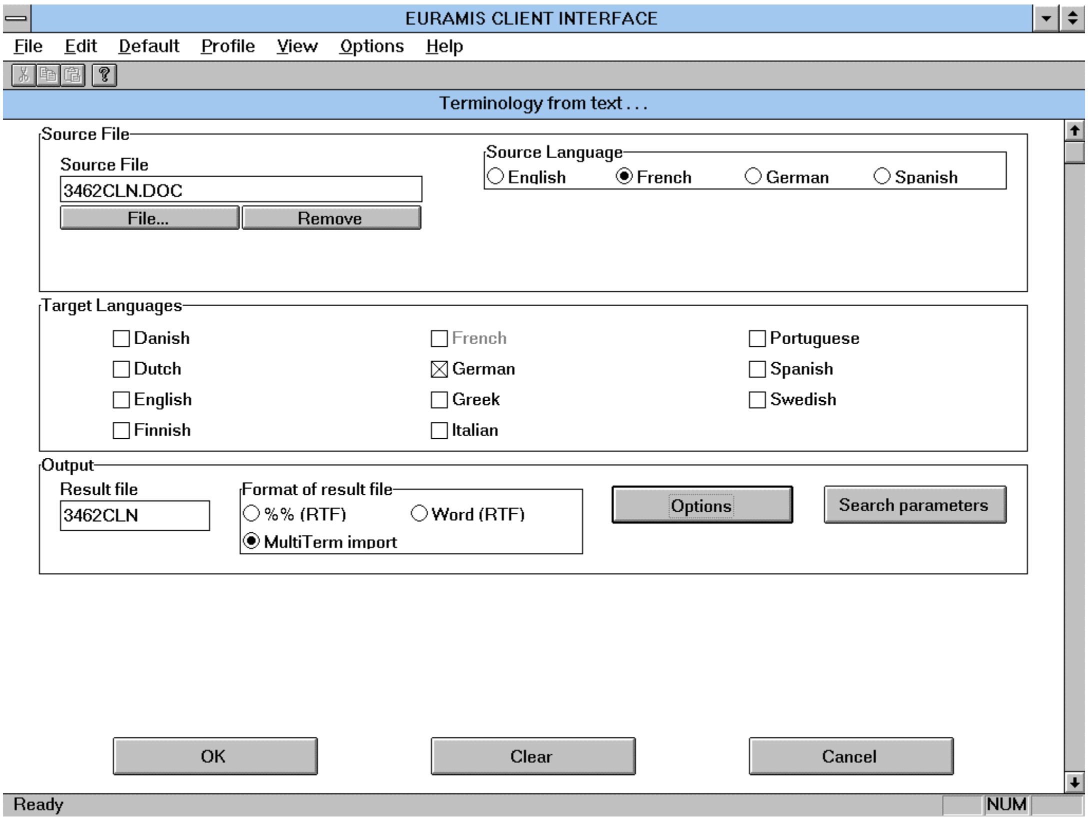
(4b) **La Communauté européenne, représentée par la Commission** de las Comunidades Europeas**, ici représentée par Monsieur ...,** Director General**,** y, de otra, y **la firme "...",** con domicilio social en ... denominada en lo sucesivo "el proveedor", representada por el Sr **...**
según el poder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

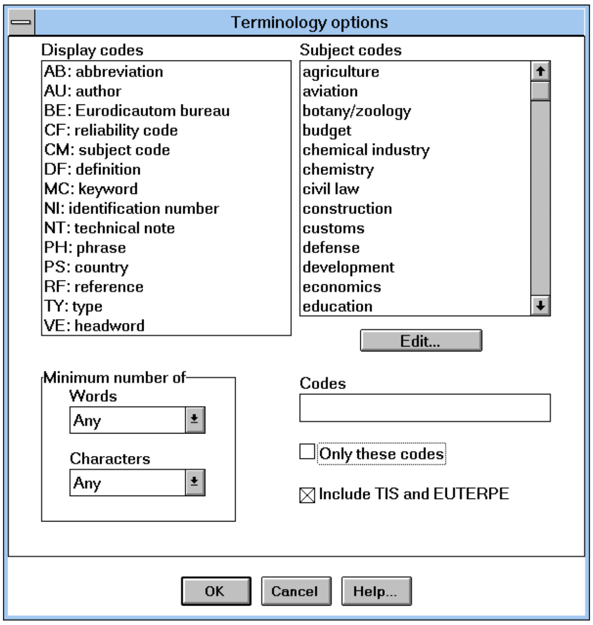
□ Danish
□ Dutch
□ English
□ Finnish
□ French
☒ German
□ Greek
□ Italian
□ Portuguese
□ Spanish
□ Swedish

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

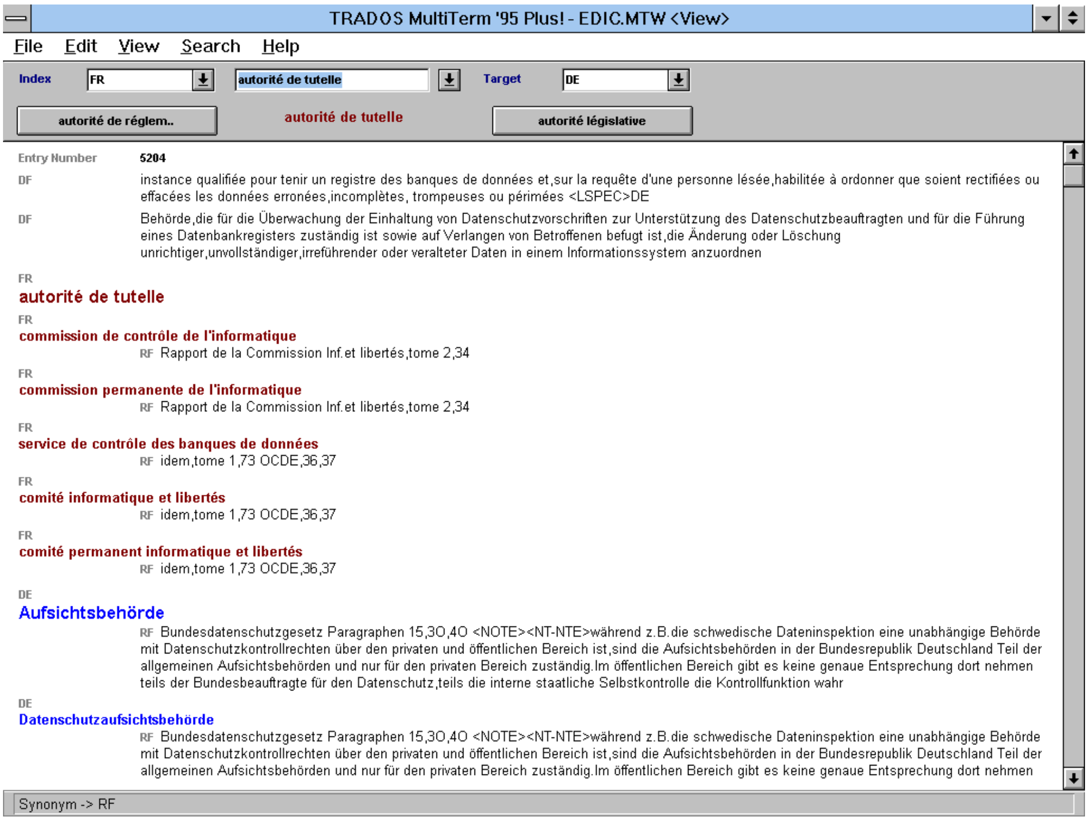
Characters: Any

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:



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.

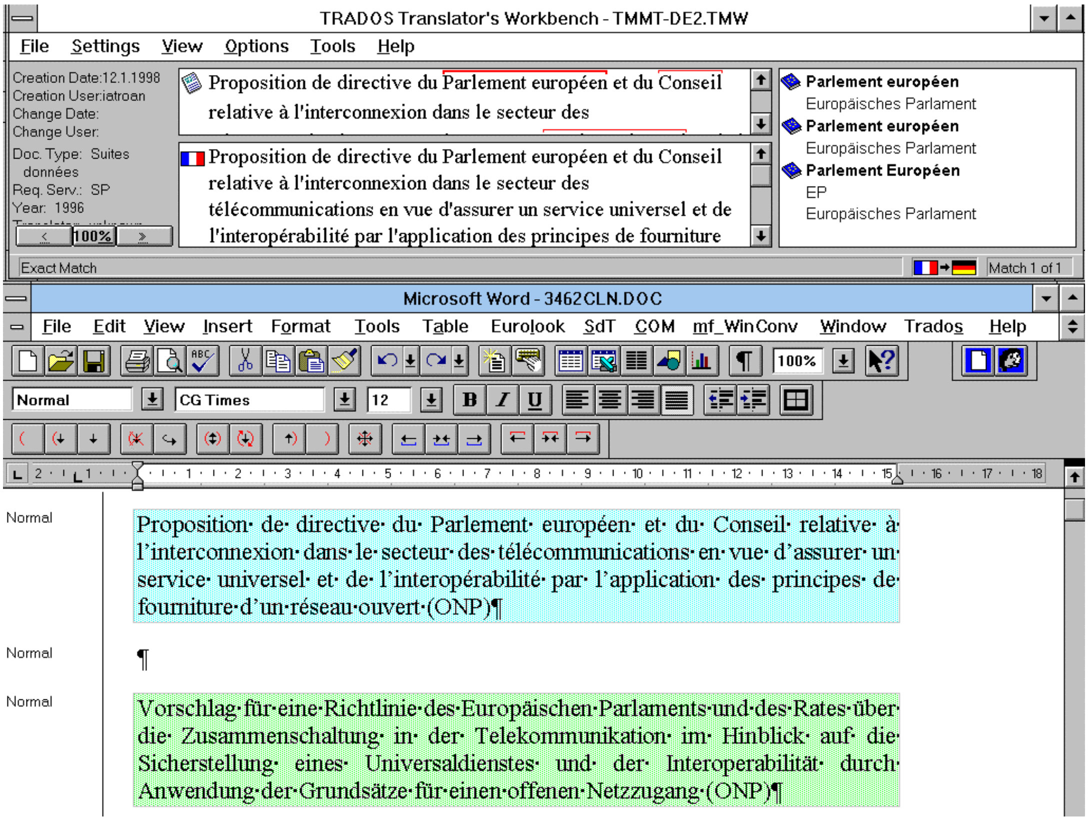
4.1. Translator's Workbench

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
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

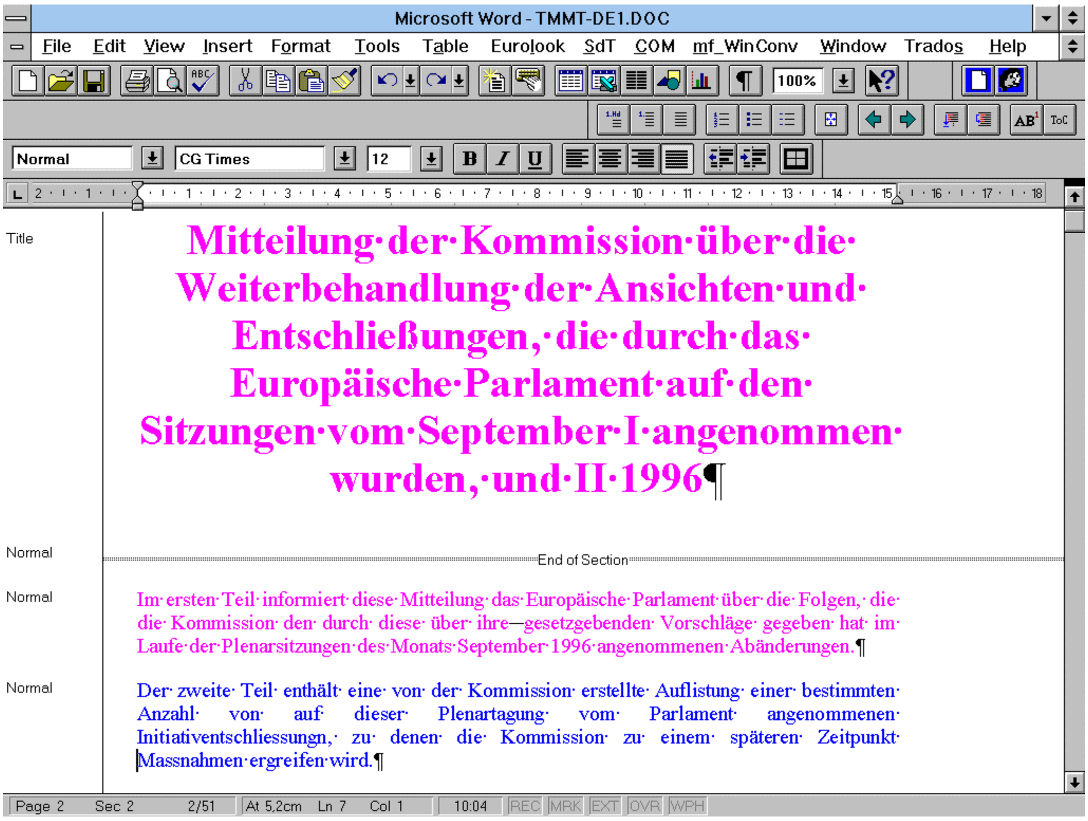
Proposition de directive du Parlement européen et du Conseil relative à l'interconnexion dans le secteur des télécommunications en vue d'assurer un service universel et de l'interopérabilité par l'application des principes de fourniture d'un réseau ouvert (ONP)¶

¶

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.