IDABC
Pan-European
eGovernment Services

Machine Translation

The linguistic policy of the European Union is built on the principle of multilingualism. Indeed, multilingualism is a legal requirement: whether EU legislation is directly applicable in all Member States or first has to be integrated into national legislation, it is essential that legislative texts are translated and published in all official EU languages. Moreover, multilingualism guarantees that EU institutions are accessible to all EU citizens.

However, a variety of measures are necessary if the principle is to be respected. These include the use of professional translators, language training for administrators, and the provision of language tools such as Machine Translation (MT), which can offer a fast and cost-effective means of obtaining raw translations of reports, minutes and e-mails.

The European Commission has been running an MT service for many years, and the results are used both by administrators in their day-to-day work and by translators as a basis for producing professional translations. The service is open to all EU staff as well as to public administrations in the Member States.

What is Machine Translation?

An MT system can be defined as a computer translation tool that works by breaking down sentences or other text segments in a given source language, analyses them in context and then attempts to recreate their meaning in a given target language, taking into account inflection, idioms and word order.

There are different kinds of MT systems, such as “rule-based”, “example-based”, and “statistical-based”. The first employs complex linguistic rules combined with dictionaries to analyse a text and then generate a translation; the other two make use of existing translations stored in bilingual corpora – great repositories of source texts aligned with their official human translations – and are largely based on past experience.

MT systems can also be “automatic” (users are not involved in the translation process) or “interactive” (if the computer has a problem with terminology, for example, it may ask users what their preferences are).

Moreover, some systems may prefer “controlled language”, where texts must be written in a clearly defined way using a restricted vocabulary, whilst others will happily accept any kind of text (full-text system).

ISBN 92-894-8944-8

The European Commission's Machine Translation Service

The European Commission's Machine Translation Service is built around EC Systran, a specific version of the SYSTRAN system originally developed by the World Translation Center (USA). Since 1976, SYSTRAN has been further developed and adapted by the European Commission for internal purposes.

EC Systran is a fully automatic, full-text, rule-based technology running on UNIX/Linux platforms at the Commission's Data Centre in Luxembourg. Hundreds of thousands of pages are submitted for machine translation every year, around 80% of which come from the Commission itself. Amongst the Member States, Spanish and German authorities (at both national and regional level) are the main users, but demand is also increasing in France.

Language Combinations

The choice of language combinations, or pairs, has been influenced by various factors such as the internal needs of the Commission, the translation quality expected from related languages, and the willingness of Member States to take part in co-financing projects, for example the EU's Multilingual Information Society (MLIS) programme.

The quality of MT results varies according to the pair selected and the contents of the source text. For example, the MT Service is more suited to reports and minutes than speeches or literary texts, and short sentences are easier to analyse than long ones. Moreover, the machine cannot "guess" what a misspelled word is meant to be, so it is useful to check the spelling of source texts before submitting them.

In general, the raw output is sufficient for browsing information, but if the results are to be distributed, the translation needs to be revised.

The MT service currently offers 28 language pairs:

Source Language (translation from)	Target Language (translation into)
Danish	English*
Dutch	English*, French*
English	Dutch, French, German, Greek, Italian, Portuguese, Spanish
French	Dutch, English, German, Greek*, Italian, Portuguese, Spanish
German	English, French
Italian	English*, French*
Portuguese	English*, French*
Spanish	English, French
Greek	English*, French
Swedish	English*

** = prototype under test*

Using MT

At present, EU staff request MT by means of an internal web interface, whilst Member State administrators have to use their standard e-mail software, a method which is less user-friendly. In both cases, the user must specify the source language and target language(s), and then attach a file or paste/type in the text for translation. The MT results are returned to the user's mailbox, normally within 20 minutes.

One of the tasks of the IDA(BC) MT project has been to develop a common access procedure that will be introduced in the near future.

What is IDA(BC) MT?

IDA(BC) MT is aimed at providing effective and user-friendly access to the Commission's MT service. The initiative is based on the principle

that more widespread access to such a tool will help European public administrations to overcome language barriers when sharing data and to reduce the overall costs associated with multilingual communication.

The project started in 2002 with a feasibility study that comprised an analysis of the Machine Translation service, a survey on the MT needs of European public administrations, and recommendations for meeting those needs.

The study highlighted a number of weaknesses in the Commission's MT service, such as a lack of user-friendliness and poor access for external users. Indeed, many national administrators were unaware that such technology was at their disposal. Other issues raised included improvement in translation quality, the integration of Member State terminology, and the extent of language coverage.

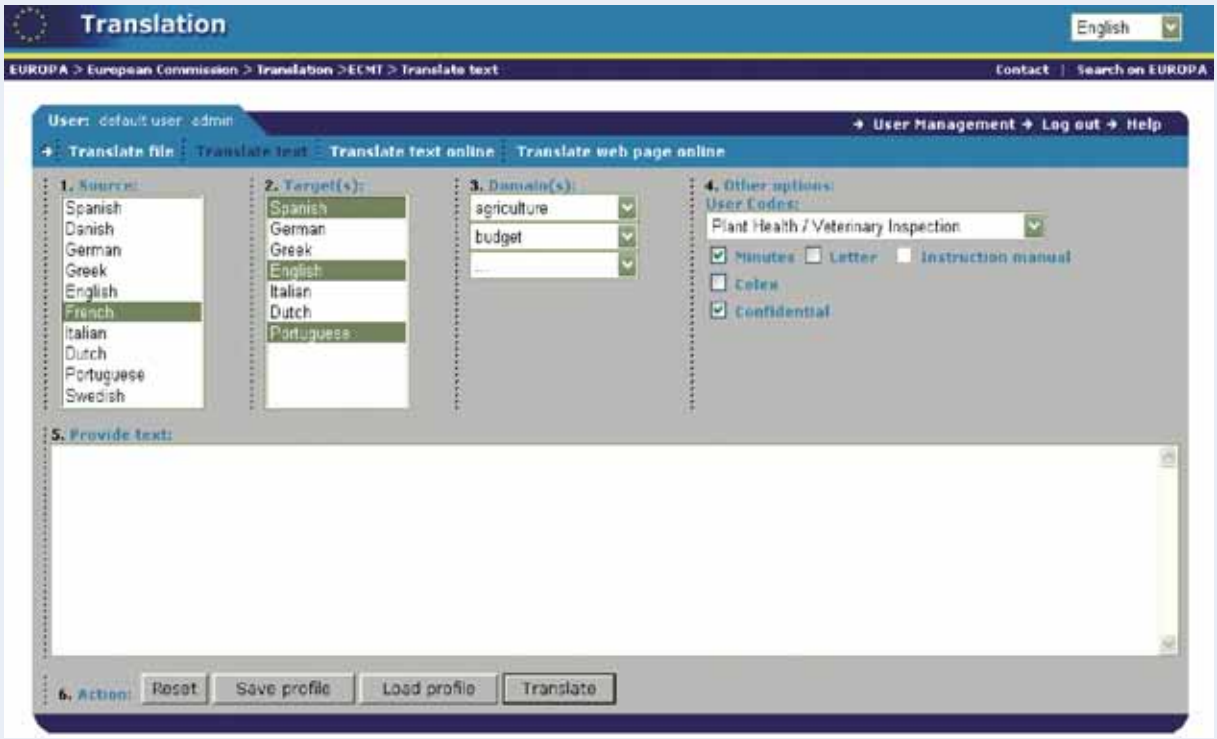
Following on from the study, terminology provided by national administrations was added to the MT dictionaries on a trial basis for translation between French, German and English (2003).

Work also commenced on upgrading the MT architecture. Improvements were carried out in two stages (2003/2004). Firstly, a common access procedure was developed so that all users could request machine translation via the web, regardless of whether they worked for EU or Member State administrations. At the same time, the user interface was completely overhauled and a new function was added – real-time translation. This means that MT results can now be returned directly to the user's screen as well as to his or her mailbox.

> Machine-produced translations are certainly less reliable than human translations, but in a multilingual working environment they can be of great help when administrators need to understand the gist of texts written in languages that they do not know, or do not know well.

Secondly, the interface was placed on a web services platform so that national administrations could integrate it with their own websites and adapt it to the "look-and-feel" of their informatics environment.

The new service will be accessible to all European public administrations through the TESTA[1] network and the Internet (for administrations not connected to TESTA) in the first half of 2005. Users will find that the service offers a variety of options, some of which can influence the final translation results:



The new MT interface

- *Domains*: these are subject-field dictionaries which, when selected, take priority over the general dictionaries. In the Nuclear Domain, for example, the French word 'coeur' is translated into English as 'core' rather than the more general 'heart'. Similarly, 'arrêt' becomes 'shut-down' instead of 'stop'. In total, 36 Domains are available and the user can choose up to three. The depth of Domain terminology varies according to the language pair.

- *User Codes*: a special code is given to a user or a group of users for accessing a set of customised terminology. A User Code takes precedence over both the general dictionaries and Domains.

- *Text Types*: (limited to certain language pairs): The selection of a Text Type ("Letter", "Minutes", or "Instruction manual") instructs the system to translate a text using specific stylistic conventions. For example, English minutes written in the past tense will be converted to the present tense for French, Spanish, and Italian; translation from French into English offers the same service, but in reverse.

- *Confidentiality*: this guarantees that the document to be translated remains on the server only during the time needed for processing, that it is not archived and is not accessible to system managers.

- *CELEX title extraction*: if this option is selected, any text references to EU legislation will be extracted automatically and looked up in the CELEX legislative database. The full title of the relevant piece of legislation will then be retrieved for both source and target languages and sent as a separate attachment with the machine translation results.

Finally, IDA MT commissioned a study on MT products for the 10 official EU languages that are not yet covered by the Commission's service – Finnish, Czech, Estonian, Hungarian, Latvian, Lithuanian, Maltese, Polish, Slovak, and Slovenian. With future accessions in mind, Bulgarian and Romanian have also been included. In addition to a market survey, the study involves a linguistic and technical evaluation of products (quality for browsing purposes, adaptability, user options, etc.) and a cost/benefit analysis. The results of the study should be available by the end of 2005.

How to participate

The address of the new MT interface will be posted on the IDABC website and elsewhere once it goes online in 2005.

In the meantime, national administrators can register for e-mail access by contacting the Commission's MT Help Desk at systran-helpdesk@ cec.eu.int.

Users are welcome to send feedback on translation errors they discover.

Machine Translation (MT) is not aimed at replacing human translators. In fact, computers cannot compete with them since they don't have the world knowledge and experience that humans acquire. Machine Translation is simply a complementary tool that can offer a quick, rough translation when time is at a premium.

About us

The European Commission's Machine Translation service is managed by the Directorate-General for Translation (DGT). IDA(BC) MT was developed and funded under the IDA programme. It will continue as an infrastructure service under the IDABC Horizontal Measures.

More information about IDABC and IDA(BC) MT can be found at www.europa.eu.int/idabc or by sending an e-mail to idabc@cec.eu.int.

[1] Information on TESTA can be found in the TESTA leaflet of this series or at www.europa.eu.int/idabc

NB-65-05-149-EN-C

European Commission
ENTERPRISE AND INDUSTRY DIRECTORATE-GENERAL