NL-Translex: Machine Translation for Dutch

Catia Cucchiarini, Johan Van Hoorde and Elizabeth D'Halleweyn

Nederlandse Taalunie
Lange Voorhout 19, P.O. Box 10595, 2501 HN, The Hague, The Netherlands
c.cucchiarini@let.kun.nl, j.van.hoorde@ntu.nl, E.dhalleweyn@ntu.nl

Abstract

NL-Translex is an MLIS-project which is funded jointly by the European Commission, the Dutch Language Union, the Ducth Ministry of Education, Culture and Science, the Dutch Ministry of Economic Affairs and the Flemish Institute for the Promotion of Scientific and Technological Research in Industry. The aim of this project is to develop Machine Translation components that will handle unrestricted text and translate Dutch from and into English, French and German. In addition to this practical aim, the partners in this project all have objectives relating to strategy, language policy and culture. The modules to be developed are intended primarily for use by EU institutions and the translation services of official bodies in the Member States. In this paper we describe in detail the aims and structure of the project, the user population, the available resources and the activities carried out so far, in particular the procedure followed for the call for tenders aimed at selecting a technonolgy provider. Finally, we describe the acceptance procedure, the strategic impact of the project and the dissemination plan.

1. Introduction

The project proposal concerning NL-Translex was submitted under Action Line 3.2 of the Multilingual Information Society (MLIS) Programme by the Nederlandse Taalunie (NTU, Dutch Language Union) together with a number of other partners in the Netherlands and Flanders. In December 1998 the proposal was approved by the European Commission and on 31 December 1998 a contract was signed, so that in January 1999 the project officially started.

In this paper we first describe the aims, scope and organization of the project. Subsequently, we deal with the market and user population. We then pay attention to the resources that will be made available by the various partners in the project. In the following section we describe the activities that have been carried out so far, paying considerable attention to the procedure followed for the call for tenders. We then provide some details on the acceptance procedure that will be followed at the end of the project to determine whether the project has indeed produced what it purports to produce. Finally, we deal with strategic impact and dissemination.

2. NL-Translex: aims and structure

2.1 Aims of the project

The practical aim of this project is to develop translation modules between Dutch on the one hand and English, French and German on the other, which translators can use effectively. In particular, the aim is to facilitate the translation work of the European Commission's Translation Service (SdT) and the translation services of official bodies of the EU Member States in their communication with the EU and with one another. A further explicit objective is to improve comprehension within the European area for officials and citizens by providing an aid which will enable them, by means of machine translation, to process data from languages which they do not speak.

In addition to the above practical aim, the NTU and its partners all have objectives relating to strategy, language policy and culture.

The MT system or components that form the object of the present project are to be developed by a technology supplier that will be selected through a tender procedure.

2.2 Scope of the project

The project concerns the development or improvement of lexica and transfer modules for automatic translation from and into Dutch. In principle, the target performance standard is always the best available, i.e. post-editing. The translation combinations envisaged are: Dutch into English/French/German and English/French/German into Dutch. The final decision on the language pairs to be developed and the performance level of these pairs depends on:

1. the total available budget, particularly the level of the financial contribution from the technology supplier;
2. whether or not components already present in the system can be used (depending on the choice of technology supplier)

In view of the financial limitations, the above-mentioned language pairs are accorded varying degrees of priority. The top priority is translation from and into English. This was selected on account of market demands (technical translations, internet) which focus on ways of translating from and into English. The fact that English is increasingly becoming the dominant language in the European Union institutions also means that the pairs involving English have top priority for use of the components in the EU institutions.

The second priority is assigned to the pairs involving French. This priority was once again dictated by the likely size of the potential user group, particularly for use in the EU and in the context of Belgian institutions. French is still an important EU language in which many documents are produced. Furthermore, the Belgian federal government, the Flemish and Walloon regions and the institutions of the bilingual region of Brussels capital are also expected to make substantial use of combinations from and into French. The pairs with German as the source or target language have the lowest priority.

Development of these pairs depends on the level of co-financing by the technology supplier. One option might be to develop just one direction or to adopt a lower quality standard. If only one-way translation is developed, the Dutch-German combination is preferable.

The field envisaged is that of directives, legal texts and official languages, particularly written communication from Member States to the EU, from the EU to Member States and between the Member States themselves. One area to be given special attention within that field is correspondence on social law, including that between administrative agencies in different states regarding implementation of the social security legislation. The specific terms and expressions relating to these fields will be collected from corpora available from the intended user organisations. These corpora include terminology lists and digitally stored text archives (largely "parallel corpora"). The specific terms in those corpora can be listed by headwords using existing computer facilities. Such aids have already been developed at universities in our language area. The user organisations include not only the Member States' bodies mentioned in this application but also the Commission's Translation Service (SdT). The SdT also has extensive resources which can be used directly, such as Eurodicautom

2.3 Parties involved in the project

This project is coordinated by the NTU and is a Dutch-Flemish effort involving various public bodies as will be explained below.

2.3.1 Financing partners

The financing project partners together provide the financial contribution which will be matched by the contribution from the European Commission. The following bodies are involved in the project as financing partners

- Ministry of Education, Culture and Science, Department of Research and Scientific Policy (OWB), Netherlands
- Ministry of Economic Affairs, Directorate General of Technology Policy (ATB), Netherlands
- Flemish Institute for the Promotion of Scientific and Technological Research in Industry (IWT), Flanders
- NTU, inter-governmental organisation (Belgium-Netherlands) for common policy on the Dutch language and literature.

The above organisations are mainly interested in the implementation of the project for language policy and economic reasons. They themselves only have an occasional, small-scale need for translation into or from the pairs of languages envisaged. However, they wish to help ensure that the Dutch language plays a full part in new technologies, including automatic translation, and to increase the awareness and use of technological facilities by the maximum number of potential users, including other government bodies. In conducting this project, they hope to achieve objectives principally in the field of language policy and culture, and the specific aims of the MLIS action programmes.

2.3.2 User organisations (non-financing partners)

The following partners are involved in the project as user organisations from the public sector:

- Ministry of Foreign Affairs, Translation Branch, Netherlands
- Ministry of the Flemish Community, Coordination Department, Chancellery and Information Service, Chancellery Section, Translation Service, Flanders

The above organisations are co-operating in the project because they themselves have translation requirements involving the pairs of languages envisaged and could use the components developed to meet those requirements. They have text material such as digital text corpora and specific terminology and expressions which can be used to construct lexica; they are willing to have real text translated by the MT system during the development phase, and have those texts critically assessed with a view to optimising the components developed. They will not make any direct financial contribution to the project, but will contribute expertise and labour effort.

These two bodies translate texts concerning virtually all areas of government policy, many of them for bodies in other Member States and/or EU institutions. The Translation Branch of the Netherlands Ministry of Foreign Affairs is also one of the principal suppliers of Dutch terminology for Eurodicautom, the EU's terminological data bank.

2.3.3 Associates

The Sociale Verzekeringsbank (SVB) is an associate in this project. This is a social security organisation responsible for implementing a number of items of Dutch legislation. The SVB translation agency primarily serves the SVB, but also carries out work for other social legislation administration agencies. Much of the translation requirement concerns correspondence with similar social security administration agencies in other countries, including EU Member States

The SVB is an associate of this project because it is not, strictly speaking, part of the public administration. However, it is a government body and is not liable to VAT. Like other user organisations, the SVB Translation Agency is co-operating in the project because it has substantial translation requirements involving the pairs of languages envisaged and could use the components developed to meet those requirements. The SVB has text material such as digital text corpora and specific terminology and expressions which can be used to construct lexica and is willing to have documents translated by the MT system during the development phase, and have the translations critically assessed with a view to optimising the components developed. This party will not make any direct financial contribution to the project, but will contribute expertise and labour effort. This is important for attaining a high standard of automatic translation in the case of texts relating to social security, one of the key areas of communication between public and other bodies of the various EU Member States

3. The market and the user population

The NTU and its partners are aiming at maximum versatility with this project. The modules to be developed are intended primarily for use by EU institutions and the translation services of official bodies in the Member States. They will therefore be largely tailored to those fields which are crucial to this institutional context, such as the law and legislation, social security, agricultural policy, economic policy, etc. At the same time, these modules should provide the basis for use in other types of text (e.g. engineering) and in other fields of application (e.g. internet). To achieve good quality, the modules will need to be further adapted to the specific requirements of the fields in question. This is a job for the market operator involved in the project, i.e. the technology provider. Further tailoring to particular fields and users is therefore beyond the scope of this project.

The applications and functions which the project should be able to handle include:

- automatic translation of official texts of the EU and Member States as input for post-editing by human translators
- automatic translation of texts as end products for browsing by non-translators
- integration of automatic translation into overall interface systems for human translators, with automatic translation being integrated with other translation aids such as translation memories and terminology databases
- automatic translation on the internet, e.g. by integrating translation systems and components into internet search engines
- use of the same components for applications other than automatic translation, e.g. for restricted grammar checkers and for automatic indexing programs.

The fact that key intended user organisations are involved in the development phase should ensure that the project produces the optimum results for application in practice. Organisations not involved will be urged to familiarise themselves with the facilities by using the information channels of the government services of the EU Member States. In addtion, the project should also produce tools for other potential users of automatic and semi-automatic translations, such as:

- translation services of major multinationals requiring translation of technical documents (e.g. software producers);
- translation agencies carrying out assignments for third parties;
- providers of internet facilities, such as navigation systems, search engines, editors, servers and web managers such as NL-NET etc.;
- providers of on-line translation assistance for small translation services and freelance translators.

4. Available resources

Apart from the financial contribution, the project partners will also provide resources and labour effort. The resources available to the project in this way include:

- The Dutch Reference Database (RBN) of the Committee for Lexicographical Translation Facilities, a committee set up by the Netherlands and Flemish Ministers for Education, which is legally accountable to the NTU. The RBN is a database for producing bi-lingual or multi-lingual dictionaries with Dutch as the source language and has around 45,000 entries, selected on the basis of frequency in modern source material comprising texts not specific to a particular field. The database is linguistically enhanced not only by morphological information (word types, inflection) but also with semantic categories.
- The database of the Dutch Orthographic Dictionary developed for the NTU by the Institute of Dutch Lexicology (INL). The Dutch Orthographic Dictionary contains 110,000 entries selected on the basis of frequency and distribution in general post-war source material. The database is enhanced with morphological information such as word classes and inflection. If necessary, the expanded word list file developed for the Electronic Dutch Orthographic Dictionary, which contains many more word forms, can be used subject to certain conditions.
- Terminology files of the participating user organisations such as the Sociale Verzekeringsbank.
- Digital files containing text material in different languages of the participating user organisations and possibly their sister organisations in other countries, including a high proportion of parallel texts, i.e. texts in Dutch and their translation in the other languages envisaged, namely English, French and German.
- On conditions to be decided, use of the text corpora of the Institute of Dutch Lexicology in Leiden produced and linguistically enhanced partly with the support of the NTU.
- OMBI, an editor for producing bilingual or multi-lingual dictionaries, developed by the Lexicographical Translation Facilities Committee for dictionary projects.
- Possibly, on conditions to be decided, other tools and aids available to the NTU's partners, such as term extractors, pre-processors, etc.

The user organisations, in particular, will also provide labour effort:

- representatives of the Users' Advisory Group who attend the advisory group meetings, consult their colleagues at work and produce preparatory consultation documents;
- testers in the translation field who have real text translated by the system (prototype) in the development phase, assess the quality of the translation and of the interface and contribute their experience to the development process.

5. Activities carried out so far

5.1 Preparatory phase

After the European Commission approved the project proposal under the MLIS action programme, contracts were signed by the European Commission and the NTU as the coordinator on behalf of the proposer, after which a number of fundamental structural arrangements were made. In sequence, these were:

- formation of a project management board

- appointment of a project manager
- formation of an advisory group of experts
- formation of an advisory group of users

The project management board is made up of representatives of all the partners from the Dutch language area and representatives of the European Commission (DG XIII) and the SdT. This board takes all fundamental decisions regarding the project, such as the appointment of a project manager, drawing up the tendering procedure and the contract documents, and placing the contract on the basis of the proposals submitted.

In view of its complexity and scale in terms of duration, budget and number of players involved, the project needed a central figure to implement and advise on the project as a whole. The project manager reports to the board and is responsible for coordination and support for the various advisory groups, conducting the tendering procedure and following up implementation by the technology supplier. The first author has been appointed as project manager.

The advisory group of experts consists of experts in the field of machine translation, natural language processing etc., e.g. people from the scientific or academic world and people with proven experience in an industrial environment. They have advised the board in drawing up the technical specifications for the tendering procedure and have assessed the responses and proposals received.

The advisory group of users indicated the wishes and requirements of the user organisations during the preparation of the project, e.g. in relation to the technical specifications and the assessmen of the proposals received from the user's point of view.

5.2 Call for tenders

After the above structures were set up, the project management board worked out a procedure for the call for tenders aimed at the selection of a firm or consortium to supply the technology for the translation system and willing to invest jointly in developing Dutch components in that system. In the following subsections we describe the requirements and the procedure characterizing this call for tenders.

5.2.1 Requirements

The technological specifications were defined when drawing up the documents for the call for tenders and following consultation with the advisory groups of users and experts. The principal requirements or expectations as regards technological aspects were:

- The system to be developed should be used as an aid for human translators in their normal computer operating system and PC word processing environment. The system should preferably not require anything other than access to an up-to-date work station and possibly a modem. The system should be capable of being incorporated in current word processing environments and/or supporting the text formats of these word processors, e.g. MS Word and Word Perfect.
- Possibility of integrating the system with other modern aids for translators, such as terminology banks and translation memories.
- The user interface should be as user-friendly as possible and require a minimum level of specific knowledge of controls and interaction.
- The components envisaged must be compatible or capable of being integrated with the existing MT environment at the European Commission.
- The components to be developed should be structured files which are as separate as possible from the (program code of the) translation engine and should offer maximum guarantees as regards potential re-use: a) in newer machine translation systems and b) in applications other than machine translation, e.g. as a basis for electronic dictionaries on CD-ROM.
- The proposers would prefer an MT system which has already proved its practical value in real life situations.

Apart from technical and technological specifications, other criteria were also considered such as:

- The firm's proven experience in developing similar components, references from existing customers using the system.
- Stability of the commercial partner, direct or indirect financial, manpower and organisational scope for executing projects of such a size.
- Willingness to make a financial contribution, and the level of that contribution.
- Willingness to grant users' licences on particularly favourable terms to the organisations taking part and to the public administration sector in the EU in general.
- Willingness to accept the NTU's ownership/joint ownership of the components developed.
- Once the components have been developed, the firm's willingness to take on their technical maintenance on its own account and at its own risk.
- (Possible) willingness to work with other firms or technology suppliers such as publishers of dictionaries and lexicographical products, and with centres of scientific expertise, particularly in the Dutch language area.

5.2.2 Tender procedure

On the basis of the information provided by a legal adviser the NL-Translex management board decided to launch a call for tenders based on the negotiated procedure. In this procedure two phases can be distinguished:

1. The announcement phase

An announcement is published in official journals and the candidates are expected to reply by sending an expression of interest. Subsequently, the candidates who reply are judged on the basis of suitability criteria which primarily concern the company rather than the product they are going to offer.

2. The offer preparation phase

The candidates who are admitted to this phase receive a copy of the Specifications that will form the basis for preparing the offer.

In the period March-April 1999 the project coordinator, in consultation with the project management board and the legal adviser, prepared the documents for the first phase of the tendering procedure and the call for tenders was officially published at the beginning of May 1999.

Four candidates replied to the call for tenders by sending a filled-in application form. The four applications were checked to determine whether they met the suitability criteria. Since all four applications passed the eligibility check, all four candidates were admitted to stage 2 of the tendering procedure.

Subsequently, the Specifications were prepared by the project coordinator in consultation with the management board, the advisory group of experts, the advisory group of users and the legal adviser and were then sent to the four candidates.

On 18 November 1999 the submitted offers were officially opened and registered. Before this meeting a weighting scheme for scoring the offers was drawn up by the project coordinator in co-operation with the advisory groups and the management board. Subsequently a second eligibility check was carried out. For each offer its conformity to the necessary conditions in the Specifications was verified.

The offers were then sent to the advisory groups of experts who evaluated them on the basis of the criteria mentioned in the Specifications. At the same time the candidates were asked to make their existing MT systems available for translation tests for their best language combinations. Tests were then carried out with the various system versions that had been made available and the test translations were evaluated by the representatives of the various user organizations.

On the basis of the scores assigned by the evaluators to the offers and to the test translations, a final ranked list of the retained offers was drawn up. Negotiations were then started with the first three candidates on the list. At the moment negotiations are going on with a selected firm on detailed conditions within the framework of the contract documents and the proposal submitted, including drawing up the contract, stipulating the detailed conditions (price, term, etc.), licences, ownership, maintenance and control of the components to be developed.

The project will of course be implemented by the firm or consortium in accordance with the contractual provisions following award of the tender. Implementation will be followed up by the project manager. If necessary, she can ask the advisory groups of experts and users for their opinion and in that way direct or adjust the course of the project. In this phase, the users are expected to make the main contribution because they will be involved in assessing the translation quality in the prototype phase. The advisory group of experts can play an advisory role in deciding on and implementing an acceptance procedure. The project manager will report on the progress of the project to the management board. The board's task is to supervise proper implementation of the project in accordance with pre-set objectives, specifications and target results within the framework established for that purpose, including the financial framework.

6. Acceptance procedure

The aim of the project is to obtain a machine translation system that will make it possible for translators to carry out their work in less time. In order to determine whether this objective has been achieved and therefore whether the technology supplier has lived up to the expectations an experiment will be carried out. In this experiment we will try to determine whether the combination of machine translation and post-editing by translators requires less time that making translations from scratch.

To this end we will use a large text corpus (thousands of pages) from which a number of texts will be selected at random. These texts will be translated by a group of translators who have experience with MT. A subset of the translators will translate the texts from scratch. The other subset will post-edit the translations produced by the MT-system. We will then check how much time was necessary to make the translations from scratch and how much time was required for the combination of MT and post-editing All translations will then be checked by experts who will not be told how the various translations were obtained. By weighting the two criteria, i.e. time required and quality achieved, we will then determine whether the developed MT system does indeed comply with the requirements.

7. Strategic impact

The project makes a substantial contribution to the objectives of the MLIS Action Programme, particularly the following sub-objectives:

- maintaining language diversity in Europe;
- guaranteeing a balance between the major languages and the languages of smaller communities in Europe;
- promoting and establishing cooperation between the EU and the Member States for the purpose of developing basic components for the multi-lingual information society;
- developing an advanced translation industry by tendering;
- creating a trained, advanced community of users by involving large user organisations from the public sector of the Member States concerned and by judicious dissemination of the project results to other potential users.

The NTU and its partners also have their own aims in carrying out the project. These objectives are not only practical but are also related to language policy and culture. The main language policy and cultural objectives of the project are:

- to strengthen the position of Dutch as a working language in the EU institutions by supporting the infrastructure required for that purpose;
- to strengthen the position of Dutch in linguistics and language technology in general and in automatic translation systems in particular;
- by judicious choice of system, to ensure the position of Dutch in new developments in automatic translation, e.g. as regards the integration of automatic translation in the internet and the integration of translation and speech technology with

a view to new systems for interaction between humans and machines;

- to make a substantial contribution to the creation of a European multilingual language infrastructure in accordance with the aims of the action programme for the "Multilingual Information Society" (MLIS) which should enable Europe to safeguard its linguistic and cultural diversity in the information society of tomorrow (e.g. integration of translation systems on the internet);
- to promote the use of modern aids by the public administration of the Member States concerned in dealing with their translation requirements;
- to make it easier for firms and institutions in the Dutch language area to tackle other language markets by providing the basis for automatic translation aids which can be used to translate documentation and marketing material.

8. Dissemination

The project results will mainly be disseminated and circulated via the internal channels available to the authorities taking part, such as their own websites, information services and officials of the various departments and the periodic publications of those departments. The project results will be publicised outside the public sector in the general press channels and focal points for awareness of the MLIS programme, active in the Netherlands and Flanders and in which the NTU has a structural and financial involvement. The technology provider will then be responsible for dissemination of the project results in the commercial sector.