The Dutch Connection

A European Machine Translation project for the Dutch language

by Catia Cucchiarini

NL-Translex is an MLIS (Multilingual Information Society Programme) project which is funded jointly by the European Commission, the Dutch Language Union, the Dutch Ministry of Education, Culture and Science, the Dutch Ministry of Economic Affairs, the Flemish Institute for the Promotion of Scientific and Technological Research in Industry and Systran S.A.

Catia Cucchiarini

Project Overview

The aim of this project is to develop Machine Translation (MT) components that will handle unrestricted text and translate Dutch from and into English and French.

This project is the result of an initiative taken a few years ago by the *Dutch Language Union (Nederlandse Taalunie–**NTU**)*. This is an intergovernmental organization established in 1980 on the basis of the Language Union Treaty between Belgium and the Netherlands, which has the mission of dealing with all issues related to strengthening the position of the Dutch language (refer to www.taalunie.org for further details on the NTU).

In 1997 the NTU, in cooperation with a number of partners from the Netherlands and Belgium, submitted the NL-Translex project proposal for the MLIS program. At that time, the idea was that the technology provider would be selected through a call for tenders. In December 1998 the proposal was approved by the European Commission and a contract providing the project with EUR 450,000 of Community funding was signed.

The possibility of using one's mother tongue is the only way to ensure that all citizens can fully participate in the information society.

With this project, the NTU and its partners pursue a number of language policy, cultural and economic objectives, such as:

- Strengthening the position of Dutch as a working language in the EU institutions by supporting the infrastructure required for that purpose;
- Strengthening the position of Dutch in linguistics and language technology in general and in automatic translation systems in particular;
- Ensuring the position of Dutch in new developments in automatic translation, e.g. the integration of automatic translation in the Internet and the integration of translation and speech technology in new systems for interaction between humans and machines;
- Contributing to the creation of a European multilingual language infrastructure in accordance with the aims of the MLIS program, which should enable Europe to safeguard its linguistic and cultural diversity;
- Guaranteeing a balance between the major languages and the languages of smaller communities in Europe;
- Promoting the use of modern tools by the public administration in the Member States;
- Making it easier for companies and institutions in the Dutch language area to tackle other language markets by providing the basis for MT aids which can be used to translate documentation and marketing material.

From the NTU's point of view, NL-Translex fits in the more general objective of strengthening the Human Language

Technology (HLT) infrastructure with a view to ensuring that all citizens with Dutch as their mother tongue can fully participate in what has come to be known as the information society. In this society, information and communication technologies (ICT) play a vital role in guaranteeing competitiveness in all branches of industry, trade and service provision. HLT is an essential part of many ICT applications. Thanks to HLT, it is possible for users to address computers in natural language, to process data from languages that they do not speak, and to continue using their mother tongue in all sorts of transactions. The possibility of using one's mother tongue is the only way to ensure that all citizens can fully participate in the information society. The availability of HLT is therefore a pre-requisite for the participation of a language, and of the citizens speaking that language, in the information society. This requires, among other things, the availability of MT components of sufficient quality.

Apart from the objectives related to strategy, economics and culture, all partners share the practical aim of developing MT modules that can be used effectively for translations between Dutch, English, and French. 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. For this reason, the MT modules to be developed will largely be tailored to those fields which are crucial to this institutional context, such as law and legislation, social security, agricultural policy, economic policy, etc.

The requirement that translators should be able to use the MT components effectively means that the MT system should make it possible for translators to carry out their work in less time. In this respect, the NL-Translex consortium has been very realistic and has not even considered the idea of having a system that produces final translations ready for use, the so-called FAHQT, Fully Automatic High Quality Translation, as this would be utopistic even with state-of-the-art technology. The idea is rather to have a system that produces raw translations that are subsequently post-edited by translators. Nevertheless, the combination of MT and post-editing should require less time than making translations from scratch. In other words, the aim in this project is to obtain FASQT, Fully Automatic Sufficient Quality Translation (J. Goetschalckx, C. Cucchiarini, J. Van Hoorde (2001) Machine Translation for Dutch: the NL-Translex Project. Why Machine Translation? Proceedings of the International Colloquium "Trends in Special language and Language Technology", Brussels, 29 & 30 March 2001, 261-280).

While the MT modules to be developed are intended primarily for use by EU institutions and by the translation services of official bodies in the Member States, they should also provide the basis for use in other types of text (e.g. engineering) and in other fields of application (e.g. the 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, Systran. Further tailoring to particular fields and users is therefore beyond the scope of this project.

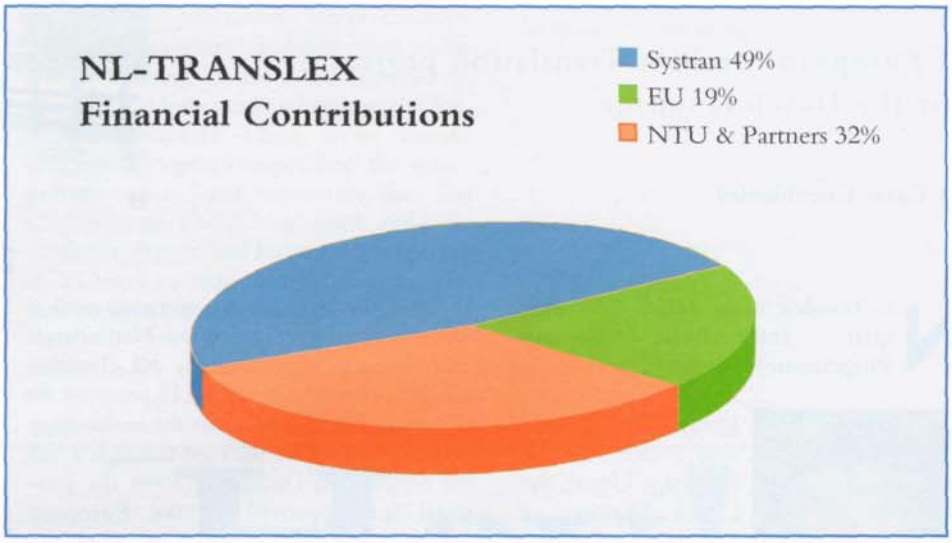

Figure 1 Specification of financial contributions.

Project Consortium

There are two types of participants in this project:

1. Financing partners (see Figure 1):

- Ministry of Education, Culture and Science, Department of Research and Scientific Policy (OCW), the Netherlands;
- Ministry of Economic Affairs, Directorate General of Technology Policy (ATB), the Netherlands;
- Flemish Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen), Belgium;
- NTU, inter-governmental organization (Belgium-Netherlands) for common policy on the Dutch language and literature;
- Technology provider: a company (selected through a call for tenders: Systran S.A.) that supplies the MT system and is willing to invest jointly in developing Dutch components.

2. User Organizations (UOs)

- Ministry of Foreign Affairs, Translation Branch, the Netherlands;
- Ministry of the Flemish Community, Coordination Department, Chancellery and Information Service, Chancellery Section, Translation Service, Flanders, Belgium;
- Translation Service of the European Commission (SdT);
- Sociale Verzekeringsbank (SVB), the Netherlands.

First Phase: Call for Tenders

An interesting feature of this project is that users have been involved right from the beginning. Not only are user organizations represented in the management board, but there is also an advisory group of users that was set up soon after the project started. During the preparation of the call for tenders, the task of this group was to indicate the wishes and requirements of the user organizations and later to evaluate the test translations made with the systems of the tenderers. The role assigned to this user group is in line with the objectives of the project, i.e. developing an instrument that can effectively be used by translators. In addition to this user advisory group, an advisory group of experts was also set up.

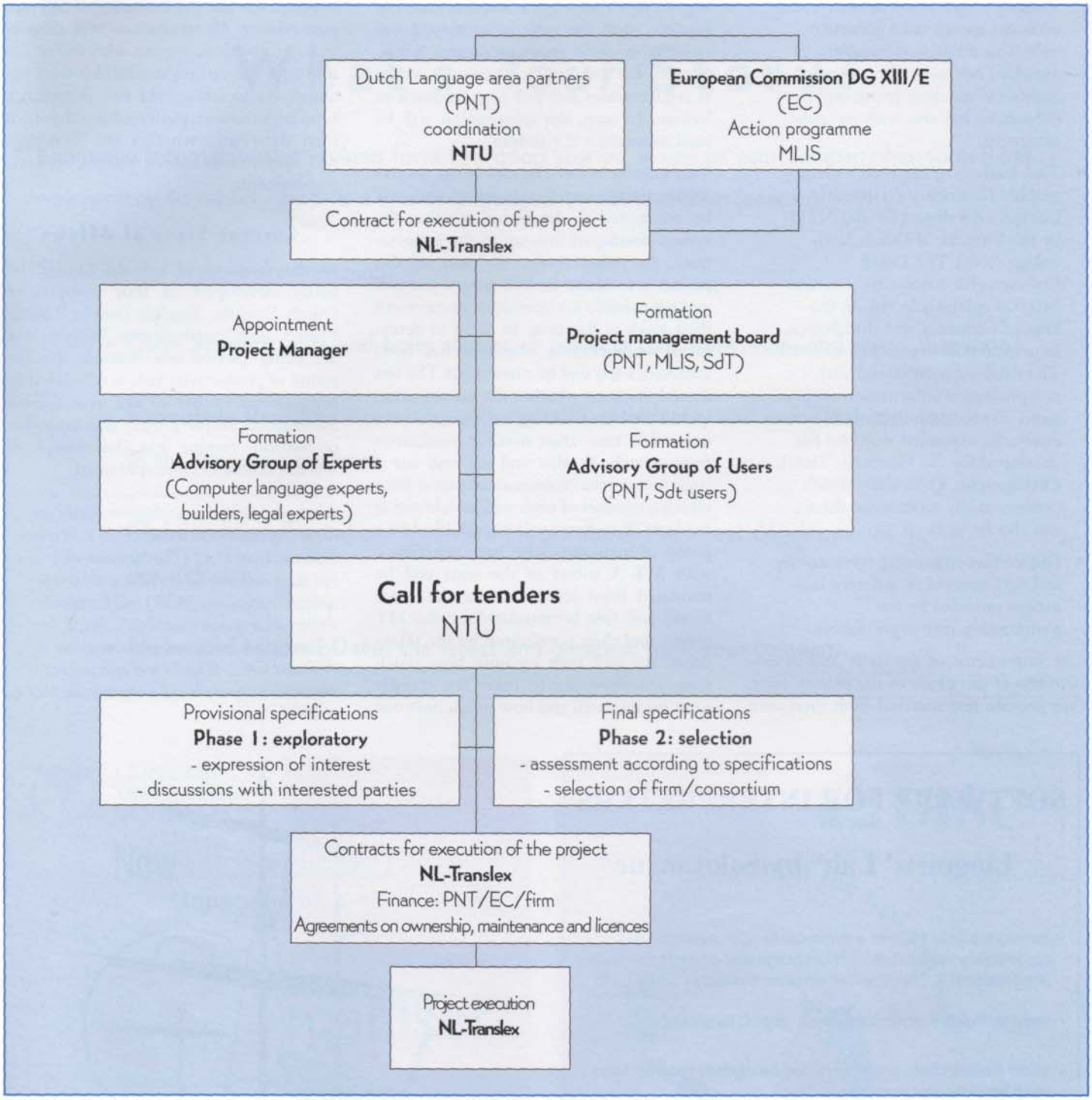

Figure 2: Activities carried out up to the selection of the technology provider.

They helped the project manager in drawing up the technical specifications for the call for tenders and evaluated the offers submitted by the candidates in the tender procedure.

After careful preparation, a call for tenders was eventually launched in May 1999. From among the four candidates that applied for this project, Systran Luxembourg turned out to be the most suitable one. Figure 2 gives a schematic indication of the activities that were carried out up to the selection of the technology provider.

Second Phase: Development

In July 2000 the company that was selected after the tendering procedure, Systran S.A., started developing the MT system for Dutch. During the negotiations it was decided that Systran would develop translation components for Dutch-English, English-Dutch, Dutch-French, and French-Dutch in this order of priority. In the development phase another interesting feature of the project emerged: the consortium's intention to make maximal use of existing resources. These are:

- The Dutch Reference Database (RBN) of the Committee for Lexicographical Translation Facilities (CLVV), a committee set up by the Dutch 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 enriched not only with morphological information (word types, inflection), but also with semantic categories.

- The database of the Dutch Orthographic Dictionary ("Groene Boekje") 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 also be used.
- Digital files containing terminology and text material in different languages provided by the participating user organizations.

The importance of the users' role is evident also in this phase of the project. First, they provide text material from their own organizations so as to ensure, as much as possible, that the system developed can cope with their own translation needs. Second, the users will evaluate the system at regular times and will give feedback to Systran. In turn, this information will be used to improve the system.

Two months before the end of the project an acceptance test (productivity test) will be conducted to determine whether the system developed lives up to the expectations. As stated above, the aim of this project is to obtain an MT 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, a test will be carried out. The test should dermine whether the combination of MT and post-editing by translators requires less time than 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 texts will be translated from scratch, while the other subset will first be translated by the MT system and then post-edited by the translators. We will then measure 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 evaluated by three 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 our requirements.

Current State of Affairs

At the moment of writing, Systran has partly developed all four components Dutch-English, English-Dutch, Dutch-French, and French-Dutch. Progress tests are being carried out whereas the first round of productivity tests is scheduled for September/October of this year. Further information on the project and an on-line translation engine can be found at: www.systranlinks.com/systran/nl.

After graduating in translation in Triest (Italy), Catia Cucchiarini obtained a PhD in Phonetic Sciences from KUN (The Netherlands). She has done research on phonetics, automatic speech recognition (ASR) and Computer Assisted Language Learning (CALL). At present she is employed as a researcher in ASR and CALL at KUN and as a project manager for Speech and Language Technology at NTU.