Machine Translation, Terminology and the African Languages in South Africa: An Overview

Milde Jordaan-Weiss

Abstract

The South African language policy changed in 1994, increasing the number of official languages from two to eleven. These are Afrikaans, English, isiZulu, isiXhosa, isiNdebele, siSwati, Sesotho, Sepedi, Setswana, Tshivenda and Xitsonga. IsiZulu tops the list by being the home language of more than 20% of the population.

In South Africa, where many opportunities for translators exist, terminology forms an important link in the translation chain. The National Terminology Services (NTS) is the only national office for terminology work in this country. Co-operation between MT developers and the NTS is probable in future, concerning the verification of terms in the African languages.

The company EPI-USE Systems is the only local developer of MT in South Africa. It has close ties with the University of Pretoria and offers inter alia a full translation service. Their product Translator Professional contains ten language pairs, four of which are African languages officially used in South Africa. All of these can be linked to optional domain dictionaries containing various terminologies. The product will shortly be available in Europe.

Milde Jordaan-Weiss

Milde Jordaan-Weiss graduated from the University of Pretoria, where she is presently completing her doctorate in History. She started working at the NTS in 1992 in the Section for Systems Development and Research and is currently an Assistant Director and head of the Section for Computer Facilities.

National Terminology Services (NTS)

The NTS is part of the South African Central Government's Department of Arts, Culture, Science and Technology. It facilitates the development and modernisation of technical and scientific terminology in all the official languages. Twenty-eight terminologists handle the excerpting, creation, documentation, standardisation and dissemination of terminology. The management of the National Termbank, which contains approximately 600 000 bilingual entries, is also the responsibility of the NTS.

Ms Milde Jordaan-Weiss
Head: Computer Facilities, National Terminology Services
Department of Arts, Culture, Science and Technology
Private Bag X894
0001 Pretoria
Republic of South Africa

Tel +27 12 314 6165 Fax +27 12 325 4943
EPI-USE's Web page: http://www.epiuse.co.za
NTS e-mail: mj-weiss@acts2.pwv.gov.za

1 Introduction: South African Languages

1.1 History of languages

Two official languages were recognised when the Union of South Africa was proclaimed in 1910. These two, English and Dutch, reflected both the composition and the history of the white part of the population. Afrikaans, the only Western European language indigenous to Africa, which replaced Dutch as an official language in 1925, developed from 17th century Dutch, influenced by inter alia Khoisan, French, Portuguese, English and Malay. English came to the country at the beginning of the 19th Century when the British took over the Cape. As a world language and the language of the conqueror, it quickly took root in this country.

In 1994 a new political dispensation changed the language policy in South Africa. The nine Bantu or African languages indigenous to South Africa also became official. They can be divided into four different families. In the Nguni family we find isiZulu, isiXhosa, isiNdebele and siSwati. The Sotho family includes Sesotho, Sepedi, and Setswana. The two other languages, Tshivenda and Xitsonga, form families on their own.

1.2 Statistics

IsiZulu is the home language of the largest group of people in South Africa. It is spoken by 8,5m people or 21.6% of the population. Second is IsiXhosa with 6,8m mother tongue speakers (17.4%) and third Afrikaans, which is the home language of approximately 6,1m people (15.6%). Next in line is Setswana with 3,6m (9.1%) and Sepedi with 3,4m (8.7%). English is spoken at home by less than 3,4m (8.6%), Sesotho by 2,6m (6.7%), Xitsonga by 1,3m (3.5%), Siswati by 926,000 (2,3%), IsiNdebele by 799,000 (2%) and Thivenda by 763,000 (1.9% of the population).

1.3 Other languages

There are other languages spoken in South Africa in an unofficial capacity. There are various communities using their own languages, like Germans, Portuguese and Greeks, while languages like French, Spanish and Dutch are spoken as well. There is also a strong Asian presence in the country. Some of these languages are taught at schools and universities.

2 The National Terminology Services as facilitator of terminology development in South Africa

The availability of terminology in various subject fields comes to mind when one talks about translation and Machine Translation (MT). Since the fifties, terminology development in South Africa's official languages has been the mission of the National Terminology Services (NTS). This office falls under the jurisdiction of the Department of Arts, Culture, Science and Technology of the Central Government. Until 1994 the NTS only attended to the two official languages, Afrikaans and English. Since then, the nine African languages mentioned above were added.

The NTS excerpts, documents, creates, standardises and disseminates terminology. By assisting terminology projects from outside collaborators as well, it facilitates terminology development on a national level. Although it is geared to handle any subject field, topics related to education, training, health, building, etc. receive

priority attention at present. Of the 23 terminologists working on projects, ten are mother tongue speakers of the different African languages.

3 The role of the NTS in MT development

The primary function of the NTS is terminology development and dissemination. However, as national body in this field it is logical that terminology verified and disseminated by the NTS will be used when MT programs involving technical and scientific domains are developed.

As MT development in South Africa is fairly recent, the exact extent to which the NTS will be involved in this development still has to be determined. As part of the Civil Service, giving free information and assistance to the public, terminology for MT development will have to be made available to all interested parties. At present, only one company has shown an interest in making use of this service.

4 EPI-USE Systems as developer of Machine Translation (MT) Software

At the University of Pretoria a research group existed at the Department of Computer Science. One of its projects was MT. About 15 years ago, a private company, called EPI-USE Systems, was formed by some of the lecturers involved in this research.. Contact between the NTS and this company dates back many years, when the NTS developed its own database program for the documentation of terminology in the official languages, which were then Afrikaans and English. This program, called Lexikon, was distributed free of charge to clients, and EPI-USE Systems considered using this database for its MT software.

This company offers various services, including translation, Network Management and SAP (R3) consulting. The component handling translation is called the Lexica Division and the software they use was developed by the company itself. It was called Lexica at first, but recently changed its name to Translator Professional. Various Windows packages are being marketed. Translator Professional has a stand-alone version using ASCII text and an integrated version to be used with MSWord 6. A scaled down version of the program is available (Translator), as well as a bilingual glossary for Afrikaans and English, called Glossit.

The minimum hardware requirement for using the program is a 486DX2-50 with 8Mb of RAM, but a Pentium with 16Mb RAM is recommended.

Translator offers the following language pairs in the European languages:

English to Afrikaans/ Afrikaans to English (50 000 concepts)
English to German/ German to English (60 000 concepts)
English to French/ French to English (90 000 concepts)
Portuguese to English (70 000 concepts)
Spanish to English (25 000 concepts)
Italian to English (35 000 concepts)

Of special interest are the African language pairs that have been developed. Swahili, although not an official language in South Africa, is one of the better developed Central African languages. Translator offers Swahili to English with 25 000 concepts. Of the official South African languages, the following are available:

Setswana to English (20 000 concepts)
Sepedi to English (20 000 concepts)
English to IsiZulu (45 000 concepts)
English to IsiXhosa (40 000 concepts)

The development of the language pair Sesotho to English is in progress, and English to Sesotho will probably follow.

Several bilingual domain dictionaries (Afrikaans and English) are available optionally for use with the MT software. These domains are Industrial, Office, Military and Academic. Within each domain, several subject fields are covered, for example the Industrial domain includes a Technical (30 000 concepts), Electrical (3 000 concepts), Chemical (7 000 concepts) and Engineering (9 000 concepts).

Although the cost of development is R100 000 (approximately $25 000) per language direction, the company markets its products at a competitive price. Translator Professional will be marketed in Europe shortly, where it will probably sell for about $100. The lighter Translator version, which is only available in South Africa, sells for about $76. Training for these products is done in a 5 hour session at additional cost.

5 Proposed co-operation between the National Terminology Services (NTS) and EPI-USE Systems

In its quest for a new Terminology Management System (TMS) able to handle eleven languages, the NTS realised that such a TMS would probably have to be imported. As it turned out, the German program MultiTerm, developed by Trados of Stuttgart, satisfied most of the NTS requirements. Local vendor support was an important factor in the final decision. As Trados had no local representative, the NTS considered EPI-USE Systems to fulfil this role. EPI-USE Systems imported the software (which the NTS as a Civil Service organisation was not allowed to do directly), installed it and provided training for all staff members.

On an MT level, EPI-USE Systems would like to co-operate with the NTS as well. As the leader in terminology work in South Africa, the NTS will probably in future verify African language terms documented by the EPI-USE Systems team. In return, the NTS will have the use of these documented terms for further processing.

6 The future

The demand for translations into the African languages is increasing. The NTS, with its limited staff resources, is currently working on a large number of terminology projects in all the official languages. These terms will eventually be accessible via electronic medium. However, if the standardised terms could also be incorporated into domain dictionaries linked to MT software, the dissemination of African language terminology to translators and other users will be facilitated. We all realise that MT is not an instant solution to all translation problems. However, in the current South African language environment MT into the African languages could be a tremendous help in promoting multilingualism.

Bibliography

WEBB, V (red) Afrikaans na Apartheid. Pretoria, 1992

1VN Webb (red). Afrikaans na Apartheid, p 28