Several directions for minority languages computerization

Vincent BERMENT
GETA, CLIPS, Joseph Fourier University
385 avenue de la Bibliothèque
Saint-Martin-d'Hères, France, 38400
vincent.berment@imag.fr

Abstract

Less than 1% of the languages spoken in the world are correctly "computerized": spell checkers, hyphenation, machine translation are still lacking for the others. In this paper, we present several directions that may help the computerization of minority languages as well as two projects where we apply some of these directions to the Lao language.

Introduction

During the last ten years, research has been driven and products have been developed to provide efficient linguistic tools for many languages. For example, Unicode is more and more a reality in today's operating systems and Microsoft Office XP contains proofing tools for more than 40 languages. However, for most of the world's people, the Information Era is still limited to using hardware and software that do not meet their needs in terms of language and script resources. Following the SALTMIL[1] terminology, we will call a minority language a language which has a smaller resource base than the major languages.

1 The available and the needed

According to the *Ethnologue*[2], more than 6800 different languages are spoken in the world. This number of languages shows that we are still far from having a software answer for all of them.

1.1 Commercial tools

First, we will notice that a trend in operating systems design and standardization allows the recent multilingual evolution. Windows (since Windows NT 3.1), MacIntosh (since MacOS 8.5) and Unix/Linux now support Unicode and many fonts are available, especially TrueType fonts[3] such as *Arial Unicode MS* which contains a large part of Unicode (51,180 glyphs but also 23 Mb that may slow our computers).
If we look now at Microsoft Office[4], one of the most widespread business suite, we observe that linguistic tools are available for 48 languages[5].

1.2 Research on minority languages

Though we may find that the coverage of several tens of languages in tools such as a word processor is a significant evolution because it covers most of the languages in terms of number of speakers, we also have to notice that it still covers less than 1% of them in terms of number of languages.
This question has been increasingly discussed in the recent years. The SALTMIL group was

[1] : Speech And Language Technology for MInority Languages (http://isl.ntftex.uni-lj.si/SALTMIL/) is a Special Interest Group of the International Speech Communication Association.

[2] : http://www.ethnologue.com/.

[3] : TrueType fonts can be used under Windows, MacOS X and Linux as well as, with limitations, under previous versions of MacOS.

[4] : http://www.microsoft.com/office/evaluation/indepth/multilingual/prooftools.htm

[5] : These languages are: Arabic, Basque, Brazilian Portuguese, Bulgarian, Catalan, Chinese Simplified, Chinese Traditional, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Gujarati, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kannada, Korean, Latvian, Lithuanian, Marathi, Norwegian, Polish, Portuguese (Portugal), Punjabi, Romanian, Russian, Serbian, Slovakian, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Vietnamese and Welsh.

created to “promote research and development in the field of speech and language technology for lesser-used languages, particularly those of Europe”. Since 1998, it has organized specific workshops at the LREC conferences.
Another definition of “minority language” is used to talk about non-indigenous minority languages. This definition may differ from the SALTMIL one. The Lancaster University (UK) has two ongoing projects related to such “minority languages”. The Minority Language Engineering (MILLE) project[1], “jointly based in the Department of Linguistics at Lancaster University and Oxford University Computing Services, seeks to investigate the development of corpus resources for UK non-indigenous minority languages[2] (NIMLs)”. The Enabling Minority Language Engineering (EMILLE) project[3], a joined project of Lancaster and Sheffield Universities, plans to “build a 63 million word electronic corpus of South Asian languages, especially those spoken in the UK”. Here, the considered languages are Bengali, Gujarati, Hindi, Punjabi, Singhalese, Tamil and Urdu which are, for some of them, already widely studied languages.

2 Difficult issues

After having verified that the need of script is covered by Unicode or at least by a *de facto* standard or simply by a font, one of the first difficulties generally met when starting with a new language is the lack of texts and dictionaries. This prevents, in particular, classical machine translation solutions from being immediately applied. Here raises a major problem: such resources are time consuming, in other terms expensive. So we need to find a way, in line with the often limited means of the minority languages populations, for getting resources or, alternatively, to build new methods, based on smaller linguistic resources.

[1] : http://www.ling.lancs.ac.uk/monkey/ihe/mille/1fra1.htm

[2] : A census done in the UK in 1991 stated that non-indigenous ethnic minorities formed about 6% of the Great Britain population (Somers 1997).

[3] : Enabling Minority Language Engineering http://www.emille.lancs.ac.uk/.

3 Directions

3.1 Generalized Linguistic Contribution

Our point of view is that linguistic resources can be efficiently obtained by a collaborative work on the web (Boitet 1999), replacing a local development team with a free and potentially much bigger distributed team. This idea of a “generalized linguistic contribution” on the web, already present in an early *Montaigne* project (1996), has recently been implemented at GETA for the Lao language in a revisited version (see § 4.2). It has also been applied by Oki to the Japanese language[4] (Shimohata 2001) and by NII/NECTEC to a Japanese-Thai dictionary[5]. At another (a meta-) level, the Open Language Archives Community[6] (OLAC) provides a collaborative platform for "creating a worldwide virtual library of language resources". Founded in December 2000, this recent project already gathers more than twenty participants (*data providers*) which resources can be accessed by using a *service provider* such as "the Linguist"[7].

3.2 Dictionary recycling

An alternative solution for building electronic dictionaries is to reengineer the document files made with a word processor to produce a paper lexicon or dictionary. When the files are not available, Optical Character Recognition (OCR) can sometimes be used to create it. There, recycling tools have to be applied to transform the original irregular format into a format that is suitable for automated tasks (Nguyen 1998).

3.3 Using analogy between languages

Another interesting direction is to take party of the similarities between languages, in particular in machine translation projects (Paul 2001). Here are several recent examples in the minority languages area.
In Europe, machine translation projects between Spanish and two languages closely related to Spanish — Catalan (Canals-Marote et al. 2001) and Galician (Diz 2001) — are already working.

[4] : http://www.yakushite.net/.

[5] : Saikam project, http://saikam.nii.ac.jp/.

[6] : http://www.language-archives.org/.

[7] : http://www.linguistlist.org/olac/.

In Asia, an example with languages from different families[1] such as Japanese and Uighur shows that syntactical closeness can be sufficient to obtain good results (MAJO system, Mahsut et al. 2001).
These machine translation projects based on analogy generally use a relatively low level transfer module and present satisfying response times thanks to the use of finite state algorithms.

3.4 International pivot-based projects

The achievement of good quality machine translation for minority languages can be boosted by the adoption of a pivot approach. In such an approach, the development of one interface (with a pivot) gives access to all languages. International pivot-based projects such as Papillon[2] and UNL[3] provide examples of such pivot-based projects including minority languages. For example, after an initial period where only major languages were involved, less computerized languages such as Mongolian and Latvian have been looked at in the UNL project.

3.5 CMU approach

The Language Technologies Institute of the Carnegie Mellon University developed an original approach of machine translation for the AVENUE project (Probst & Levin 2002). This multi-engine system, based on a corpus-based machine translation (CBMT), uses both EBMT and SMT[4] as well as an elicitation tool[5] that learns transfer rules from a small and controlled corpus. This elicitation tool, currently being applied to Mapudungun, a language from Chile, seems to be well suited to the minority languages because of its low need of linguistic resources.

4 Ongoing projects at GETA

Hereafter are presented two developments we currently undertake at GETA in Grenoble to apply the ideas presented here[6]. Both are applying these ideas to the Lao language. Lao is spoken in Laos by about 4 million people and in Thailand by more than 10 million people[7].

4.1 PapiLex[8]

4.1.1 Principles

PapiLex, a Lao lexical base developed in the context of the pivot-based *Papillon* project, follows the fundamental rules of this project:

- lexical base in XML format,
- use of the explanatory and combinatorial lexicology (ECL) concepts[9] (from which the core monolingual Papillon XML schema is directly derived),
- use of Unicode for the characters encoding.

PapiLex is a mockup aimed at giving a help in evaluating the *Papillon* project difficulties. The dictionary structure contains eight fields, derived from the ECL:

- Lexical item,
- Part of speech,
- Semantic formula,
- Government pattern,
- Lexical functions,
- Examples,
- Idioms,
- Interlingual meaning.

4.1.2 Architecture

PapiLex has been developed using HTML and Perl. The Perl scripts handle the interaction with the XML base. The interface with this base relies on DOM, the Document Object Model standardized by the W3C. We used a DOM package for Perl which can be found on the perl.com site. The parsing set used on the web server is the one which is included in ActivePerl 5.6.1 for Windows. It is made out of the Larry

---

[1] : Uighur is a Turkic language and Japanese is considered as independent (Katzner 1995).

[2] : http://vulab.ias.unu.edu/papillon/.

[3] : http://www.unl.ias.unu.edu/.

[4] : Example-Based and Statistical MT.

[5] : Called iRBMT = instructible Rule-Based MT.

[6] : In our works, we focus on "minority languages" taken in the SALTMIL definition sense.

[7] : In the Isan area of Thailand where Lao is spoken, Thai scripts are used and also the language itself is somehow different from Lao spoken in Laos. There is also an important Lao diaspora in France, Australia and USA. See www.geocities.com/lao_thai2000.

[8] : http://cams-atid.ivry.cnrs.fr/papilex/.

[9] : On this matter, see André Clas, Igor Mel'čuk and Alain Polguère's book, Introduction à la lexicographie explicative et combinatoire, Duculot 1995

Wall and Clark Cooper "XML::Parser" package and of "Expat", the James Clark's XML parser.

4.2 Montaigne project[1]

4.2.1 Initial specifications

Basically, the Montaigne project's idea is to offer a free collaborative work facility on the web for development of linguistic resources and machine translation tools. Though its ambition is generic, the project started with an application to Lao.

In this early form, the web site mainly offers three kinds of services:

- Lao-French translations,
- Transcriptions of Lao,
- Lexicographic creation.

Contrarily to the two first items (translations and transcriptions) which are open to all visitors, the lexicographic creation access is limited to registered skilled users. Each registered user has his own space where he can save his private words and texts.

The linguistic structure of the dictionary follows the ECL concepts so it can easily be exported toward Papillon. However, additional fields have also been added in order to derive other applications from the database as for example paper dictionaries or machine translation.

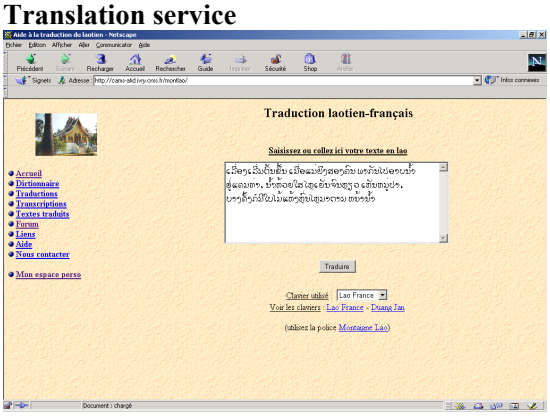
In order to start the process, a first dictionary of 1038 words has been entered, simply deriving from a paper dictionary done by Lamvieng Inthamone in Word format. So this initial dictionary does not meet yet the ECL concept of lexical item required for exporting the dictionary toward Papillon. An "ECLization" of this base dictionary is then currently being handled by a group of Lao students from Inalco[2] (*Institut National des Langues et des Civilisations Orientales*, located in Paris). This team will produce an ECL-compliant dictionary that will replace the current one at the end of their task.

4.2.2 Architecture

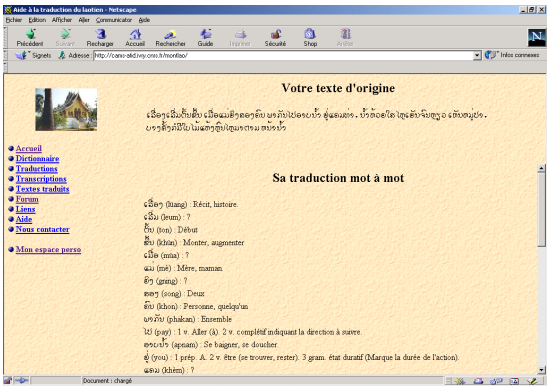
The architecture is based on HTML, SSI, PHP, JavaScript and compiled C code used as CGI. The dictionary is stored as a MySQL database table as well as the contributors' profiles. C code is used for segmenting Lao texts into words[3] and for sorting the dictionary[4]. It uses a syllable recognition technology (Berment 1998) and a longest matching algorithm (e.g. Meknavin et al. 1997). Unlike PapiLex, the Montaigne Lao project uses non-Unicode fonts. This is mainly due to the unavailability of Unicode fonts for Lao that would actually work. Text input is possible with the two currently used Lao keyboard layouts thanks to JavaScript and to *TextArea* or *Input* HTML forms controls.

4.2.3 Several views

Translation service



Original text input page (Lao)



Word for word translation page (French)

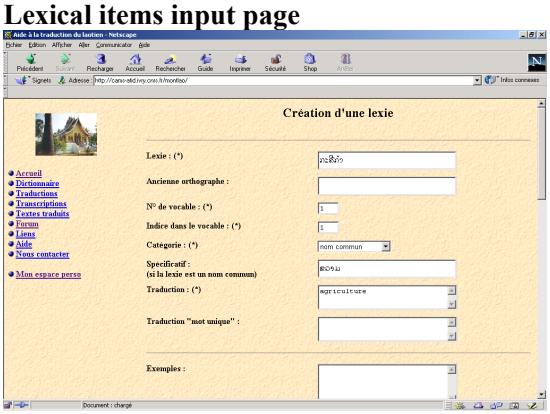
[1] : http://cams-atid.ivry.cnrs.fr/montlao/.

[2] : http://www.inalco.fr.

[3] : Lao is written from left to right with an alphabet deriving from Indian scripts. A major characteristics of Lao writing is that words are not separated with spaces, like Khmer, Thai or Burmese writings.

[4] : Another important characteristics of Lao writing is that some vowels are placed before the consonant. This contributes to make the automatic sort of Lao dictionaries more complex.

Lexical items input page



Ordered list of lexical items

Conclusion

In the close future, we plan to develop the Montaigne project in two directions.

First, the current prototype will become a full scale production tool. For that, the Lao-French translations and the lexicographic creation will be linked together so that a registered user can modify a translation. This will update his private dictionary and the altered word will be submitted to the Linguistic Management Team for updating the common dictionary. Analogy between Lao and Thai languages will also be looked at.

The second anticipated milestone is to develop the project toward its initial generic aim: a free collaborative work facility on the web for development of linguistic resources and machine translation tools for any minority language. This includes:

- gathering a free and structured set of generic tools (lemmatizers, segmenters, speech tools, ...) and making them available on the web site,
- offering a collaborative environment for each candidate language, derived from the Lao experimental one.

References

Berment Vincent. (DEA dissertation, Inalco 1998) *Prolégomènes graphotaxiques du laotien.* 160 p.

Boitet Christian. (MT Summit 1999) *A research perspective on how to democratize machine translation and translation aids aiming at high quality final output*. 10 p.

Canals-Marote R., Esteve-Guillén A., Garrido-Alenda A., Guardiola-Savall M.I., Itturaspe-Bellver A., Montserrat-Buendia S., Ortiz-Rojas S., Pastor-Pina H., Pérez-Antón P.M., Forcada M.L. (MT Summit 2001) *The Spanish-Catalan machine translation system interNOSTRUM*. 4 p.

Diz Gamallo Inés. (MT Summit 2001) *The importance of MT for the survival of minority languages: Spanish-Galician MT system.* 4 p.

Katzner Kenneth. (Routledge 1995, 3rd edition) *The Languages of the World*. 378 p.

Mahsut Muhtar, Ogawa Yasuhiro, Sugino Kazue, Inagaki Yasuyoshi. (MT Summit 2001) *Utilizing Agglutinative Features in Japanese-Uighur Machine Translation*. 6 p.

Meknavin Surapant, Charoenpornsawat Paisarn, Kijsirikul Boonserm (Natural Language Processing Pacific Rim Symposium 1997) *Featured-based Thai word segmentation.*. pp 41-46.

Nguyen Hai Doan. (PhD dissertation, UJF Grenoble 1998) *Techniques génériques d'accumulation d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multi-lingues hétérogènes*. 168 p.

Paul Michael (MT Summit 2001) *Translation Knowledge Recycling for Related Languages*. 5 p.

Probst Katharina, Levin Lori (Proceedings of TMI 2002) *Challenges in Automated Elicitation of a Controlled Bilingual Corpus*. 11 p.

Shimohata Sayori, Kitamura Mihoko, Sukehiro Tatsuya, Murata Toshiki. (MT Summit 2001) *Collaborative Translation Environment on the Web*. 4 p.

Somers Harold. (Translating and the Computer 19, Papers from the ASLIB Conference 13/14 November 1997) *Machine Translation and Minority Languages*.