Creating rich online dictionaries for the Lao↔French language pair, reusable for Machine Translation

Vincent Berment
INaLCO, Paris
Vincent.Berment@inalco.fr

Abstract

In this paper, we present how we generated two rich online bilingual dictionaries — Lao-French and French-Lao — from unstructured dictionaries in Microsoft Word files. Then we shortly discuss the possible reuse of the lexical data for Machine Translation projects.

1 Introduction

The creation of a dictionary with a large coverage is a very difficult and time-consuming task, when starting more or less from scratch. At INaLCO (Oriental Studies Institute, Paris, France), where the Lao language is taught, two dictionaries were recently published almost at the same time, making an outstanding milestone because of their coverage and their quality compared to the previous ones. We detail hereafter how we transformed them into digital resources.

2 Available bilingual dictionaries containing the Lao language

Bilingual resources with Lao as one of the languages are relatively rare and often poor. The main ones to our knowledge are (list limited to general bilingual dictionaries with more than 300 pages ; this list simply shows how scarce are the Lao-NL bilingual dictionaries):

Authors	Date	Languages	Pages
SOUKBANDITH Bounmy	1983	English-Lao and Lao-English	719 p.
KERR Allen D.	1992	Lao-English	XX-1223 p.
PATTERSON William Lorenzo, SEVERINO Mario E.	1995	Lao-English	826 p.
SISAVEUY Souvanny	1996	English-Lao	901 p.
KANGPHACHANPHENG, Keo, VILAYSACK Vilayphan, KOUNLAPHAN Vongnathi	1996	English-Lao and Lao-English	1033 p., ill., 522 p.
BOUARAVONG Phone, CHIEMSISOURAJ Chanthaphilit, CHANTHAPHONE Vanhnolack	1999	English-Lao	508 p.
BOUARAVONG Phone, CHIEMSISOURAJ Chanthaphilit, CHANTHAPHONE Vanhnolack	2000	English-Lao and Lao-English	739 p.
MARCUS Russell	2000	English-Lao and Lao-English	416 p.
MINGBUAPHA Khamphan, BECKER Benjawan Poomsan	2003	English-Lao and Lao-English	780 p.
REINHORN Marc	1970	Lao-French	49-2150 p.
NGINN Somchine Pierre	1980	French-Lao	VI-910 p.
SOUKHAVONG Souphaphone, SOUKHAVONG Khamsay	1985	Lao-French	[14]-581 p.
SIMANA Suksavang	1994	Lao-French-English	429 p.
INTHAMONE Lamvieng	2011	Lao-French	1523 p.

Authors	Date	Languages	Pages
REINHORN Marc, BERMENT Vincent	2013	French-Lao	1729 p.
MOREV, L.N. VASILYEVA, V.H. PLUM, U.Y. (МОРЕВ Л.Н., ВАСИЛЬЕВА В.Х., ПЛАМ Ю.Я.)	1982	Lao-Russian	952 p.
MOREV, L.N., KEDAYTENE E.I., MITROKHIN V.I. (МОРЕВ, Л. Н., КЕДАЙТЕНЕ, Е. И., МИТРОХИНА В. И.)	1987	Russian-Lao	352 p.
SISAENGCHAN Thongsit, AMPHAI Vognobuntham	199?	Lao-Magyar	604 p.
LÊ Duy Luong	1992	Vietnamese-Lao	742 p.
PHAM Duc Duong, HOANG Tung Son, TRUONG Duy Hoa	1995	Lao-Vietnamese	835 p.
PROMPRAPHAN Waranon, SAYAVONG Somseng, McCARTHY Robert (Kasetsart University, Department of linguistics)	2000	Lao-Thai-English	XVII-762 p.
WIRAPHONG Misathan	2000	Lao-Thai	XXIV-428 p.
VIRACHIT Khamphanh, OUDOM Kikèo, PHONEKASEUMSOUK Kidèng	2000	Khmer-Lao	XII-1246 p.
HOUANGBINH Sisouvanh	2000	Lao-Chinese	1523 p.
INTHAVONGSA Kèo, et al.	2000	Lao-Japanese	XI-410 p.
SULAVAN Khamluan, KINGSADA Thongpheth, COSTELLO Nancy A.	1998	Katu-Lao-English	363 p.
PREISIG Elisabeth, SIMANA Suksavang, SAYGNAVONG Somseng	1994	Kmhmu-Lao-French-English	68-429 p.
TAYANIN Damrong, SVANTESSON Jan-Ölof, LINDELL Kristina, SAYAVONG Somseng, KINGSADA Thongpheth	1994	Kmhmu-Lao	501 p.

Figure 1: Bilingual dictionaries including Lao (from Bernard Gay, 2003 [1], with additions)

The number of pages provides (to a certain extent) a possibility to compare the quantity of lexical information between the dictionaries, but it does not allow evaluating their quality. The content is actually often limited to an entry, a part of speech, a pronunciation and only one word (thus one sense) as the translation. Moreover, some dictionaries are obviously partial or integral plagiarisms.

As for the available online bilingual dictionaries, the main ones are with English:

- http://sealang.net/lao/dictionary.htm
    - Lao ↔ English (both directions)
    - Derives from Kerr and from Patterson/Severino dictionaries
- http://www.seasite.niu.edu/Lao/LDictionary/default.aspx
    - Lao → English
- https://translate.google.fr
    - Lao ↔ English (both directions, through English for the other languages)

and for French:

- http://laosoftware.com/
    - Lao ↔ French (both directions)
    - Relies on Paul Jadin’s dictionary

3 From unstructured dictionaries to clean databases with fine structures

3.1 The original dictionaries in Microsoft Word files

Recently, two relatively rich dictionaries (~40,000 entries, ~60,000 word-senses, ~15,000 expressions, many details including POS, examples, glosses, special plural forms, synonyms...) between Lao and French ([2], [3]) brought the opportunity to provide them as digital resources, as we were allowed to use their original Microsoft Word files. In both cases, the dictionary was made of one file per initial letter. Altogether, the authors spent about 40-50 years to produce these two dictionaries.

3.2 Step 1: Parsing the Word files

The first step towards constructing the database was to parse as finely as possible the Word files, in order to discover their fine-grained structure ("microstructure"). We used Claude Del Vigna's "saint-jean" compiler that generates parsers in C++. This task was quite complex as this structure could sometimes lack regularity or rigor. Actually, this step also included manual modifications in the files when parsing failed, in order to make the structure rigorously regular. A simple example among hundreds or thousands: the POS could vary from entry to entry, giving (for "*verbe transitif*") sometimes "`vt`", "`vt.`", "`v.t.`", or even "`v .t.`" (with a white space inside). This step has certainly been the longest one, due to this iterative process, and also because the structure discovery itself was also iterative. The rarest types of lexical information drove us to modify the parser every time they occurred, and also to reparse the parts already successfully parsed with the previous version of the parser.

In order to exploit the style information available in the Word files, we chose to embed the parser in a Word addin[1]. A Word addin is a DLL library that is automatically loaded when Microsoft Word starts. This library must be placed in a specific directory (configurable in Word) and have "`.wll`" as file extension (instead of the usual "`.dll`"). Doing so, we could use the font name, size, and style to characterize the different elements. For example, the legacy (non-Unicode) fonts used to write the parts in Lao language are never used for French, so these parts could be assigned the categories of either `entry` (Lao→French) or `translation` (French→Lao) or `example` in Lao... This was indeed very useful and even simply made the parsing possible.

3.3 Step 2: Generating the lexical database

The WLL was written in C++ and was compiled and linked with the SQLITE[2] code to generate the lexical database. The generated tables are directly derived from the dictionary structure. Here is an example for the French→Lao dictionary (without the tables used to describe the examples).

- `Vedette` (the main table with the lemma and an index for linking the other tables)
  - `NumeroEntree` (entry id)
  - `Vedette` (lemma)
- `Entree` (table containing miscellaneous information for the entry)
  - `NumeroEntree` (entry id)
  - `NumeroDeSens` (sense id French)
  - `CMS` (POS)
  - `Correlat` (reference to other entries in the dictionary)
  - `Exemple` (example)
  - `Pluriel` (special plural forms)
  - `Locution` (in case the entry is part of a frozen expression)
  - `CommentaireParentheses` (gloss)
  - `LocutionEtoile` (gloss in case the entry is part of a frozen expression)
  - `CMSLocutionEtoile` (POS in case the entry is part of a frozen expression)
  - `Synonyme` (synonym)
- `Renvoi` (table linking an entry to another, for example in case of multiple spellings)
  - `NumeroEntree` (entry id)
  - `NumeroDeSensLao` (sense id of the Lao translation referred to)
  - `Renvoi` (lemma of the reference entry)

[1] https://support.microsoft.com/en-us/kb/190057 (see http://www.wordaddins.com/ for recent versions of Word).

[2] https://sqlite.org/

- `Traductions` (table containing the translations)
    - `NumeroEntree` (entry id)
    - `NumeroDeSensLao` (sense id Lao)
    - `Traduction` (translation)
    - `Commentaire` (comment in Lao associated to the translation)
    - `Abreviation` (abbreviation in Lao associated to the translation)

3.4 Step 3: Cleaning the lexical database

A meticulous verification step followed the generation of the database. Some errors still remained and had to be fixed. Then, we still had to transform into Unicode the parts that were initially written with non-Unicode fonts. This was the case for the parts written in Lao as well as the IPA transcriptions (for the Lao→French dictionary only).

Nota: An ongoing work is currently being done by Lamvieng Inthamone to refine the structure of the Lao-French dictionary, so that the two dictionaries will be at the same level. This is done in an Excel file extracted from the cleaned database.

3.5 Step 4: Creating the software for the online dictionary look-up

The last step was to make the dictionaries available online. As we wanted to provide the users with the possibility, for the Lao→French direction, to submit strings containing more than one word (it is not always easy to know where a word starts and ends, as there are no spaces between words in Lao), the first thing to do was to embed a word segmenter in the translation process. We chose the general-purpose segmenter MOTOR (see [4], [5]) and embedded it as the initial phase of the translation pipeline from Lao to French. In order to make the translations consistent with the segmentation, the segmenter was configured with the list of entries of the Lao→French dictionary. Then, when a string contains several Lao words, several requests are done in AJAX towards the server.

The dictionaries are available at http://laosoftware.com/HeloiseTest/Dicolofr/indexnew.htm.

4 Conclusion

The Lao-French and French-Lao dictionaries are available online since early 2016 and the first feedbacks are very positive. Now, the next step will be to use the associated lexical databases to bootstrap the creation of dictionaries for Lao→French and French→Lao machine translation systems. This can be done semi-automatically by associating the lemmas in French with the lexical units of the existing analysis and generation modules of French. The examples in the dictionaries, which are most generally multi-word expressions, can be associated to the existing MWEs of the analysers or added when absent.

Another interesting use of the lexical databases would be to link the Lao words to words in other languages, thanks to existing dictionaries between French and other languages, or using interlingual lexical units such as WordNet synsets[3] or UNL Universal Words[4]. A further possibility would be to build MT systems between Lao and many other languages using UNL graphs as pivot representations.

References

[1] Gay, B. (2003). Les sources contemporaines du lao / Contemporary sources on Lao : 1976-2003. ACRS (Singapour) / Institut de Recherches sur la Culture (Laos) Comité National des Sciences (Laos) 2003. 1385 p.

[2] Inthamone, L. (2011). Nouveau Dictionnaire Lao-Français. You Feng. 1523 p.

[3] Reinhorn, M. ; Berment, V. (2013). Dictionnaire Français-Lao. You Feng. 1729 p.

[4] de Malézieux, G. ; Bosc, A. ; Berment, V. (2014). RBMT as an alternative to SMT for under-resourced languages. WSSANLP 2014, 23 August 2014, Dublin.

[5] Berment, V. (2014). Some thoughts on how to address commercially unprofitable languages and language pairs. keynote speech, WSSANLP 2014, 23 August 2014, Dublin.

[3] Here is the WOLF (*WOrdnet Libre du Français*) French WordNet: http://alpage.inria.fr/~sagot/wolf.html.

[4] Resources are here: https://github.com/dikonov/Universal-Dictionary-of-Concepts/tree/master/data/csv.