Several Ways to Use the Lingwarium.org Online MT Collaborative Platform to Develop Rich Morphological Analyzers

Vincent Berment[1](✉), Christian Boitet[2](✉),
Jean-Philippe Guilbaud[3](✉), and Jurgita Kapočiūtė-Dzikienė[4](✉)

[1] INaLCO, 65 rue des Grands Moulins, 75013 Paris, France
Vincent.Berment@inalco.fr
[2] GETALP, LIG, UGA, 700 avenue Centrale, 38041 Grenoble, France
Christian.Boitet@imag.fr
[3] GETALP, LIG, CNRS, 700 avenue Centrale, 38041 Grenoble, France
Jean-Philippe.Guilbaud@imag.fr
[4] Vytautas Magnus University, Kaunas, Lithuania
Jurgita.K.Dz@gmail.com

Abstract. We will demonstrate several morphological analyzers of languages for which morphological analysis is very difficult, and/or that are under-resourced. It will cover at least French, German, Khmer, Lao, Lithuanian, Portuguese, Quechua, Spanish and Russian. These morphological analyzers all run on the collaborative platform lingwarium.org that supports the ARIANE-H lingware development environment. Some will also be presented as stand-alone Windows applications.

1 Introduction

The online platform lingwarium.org was opened in July 2016. It provides a means for geographically scattered groups of language experts to develop new machine translation systems collaboratively, especially for under-resourced languages. The main linguistic programming toolkit is ARIANE-H, the version of ARIANE-G5 recently produced by Vincent Berment [1]. Lingwarium.org also offers other tools such as MOTOR, dedicated to the word segmentation of texts in languages using an unsegmented writing system, such as many Asian languages (Burmese, Khmer, Lao, Thai…). It also contains some programs used to speed up the development process.

The present paper details the different approaches used under lingwarium.org to develop rich morphological analyzers (as first steps of MT systems), using the ARIANE-H toolkit and some other tools. The demonstration will include morphological analysers for several languages, including the ones detailed in this paper: French, German, Khmer, Lao, Lithuanian, Portuguese, Quechua, Spanish and Russian.

A. Gelbukh (Ed.): CICLing 2017, LNCS 10761, pp. 81–86, 2018.
https://doi.org/10.1007/978-3-319-77113-7_6

2 Word Segmentation

The MOTOR word segmenter relies on the minimum matching algorithm that computes the segmentation of a text which contains the smallest possible number of words. In case it finds several solutions, it outputs the first one.

To run its algorithm, MOTOR only needs a list of words (word forms) for the language to be treated. MOTOR is currently used operationally in analysers for Burmese (27,493 words), Khmer (85,655 words), Lao (50,078 words), Thai (20,574 words) and old Tibetan (26,730 words). We also tested it with Japanese for a limited corpus (see below, the "Little Prince" project).

3 Tokenization, Stemming and POS Tagging

Another important operation in morphological analysers is to compute a lemma for each word of the texts to be analysed. In LINGWARIUM, this task is handled by writing inflectional and compositional rules in ATEF. ATEF is the SLLP (specialized language for linguistic programming) of the ARIANE framework used for writing morphological analysers.

Though this language is quite easy to use, a number of tools have been developed to simplify the task of the lexicographers. These tools can generate ATEF code from simple tables, typically Excel sheets or database tables, or from other frameworks such as NOOJ.

For **Lithuanian**, we took all the distinct words (word forms) from an extract of the "*Corpus of the Contemporary Lithuanian Language*", created at Vytautas Magnus University [2]. The corpus extract we used contains about 1 million running words and covers different domains: fiction texts, newspaper texts, legislative texts, parliamentary transcripts, etc.

These word forms have been associated with their lemmas and grouped into 17 parts-of-speech: nouns (16,321 distinct lemmas); adjectives (4,937); adverbs (2,017); numerals (78); several verb forms differing in their inflection as verbs (11,831), participles (11,831), half participles (11,751), adverbial participles (lith. padalyviai) (11,831), adverbial participles1 (lith. būdiniai) (11,751); pronouns (43); particles (117); interjections (59); onomatopoeias (40); conjunctions (62); prepositions (73); abbreviations (109); and acronyms (156).

For each lemma, stable and unstable parts (changing due to inflection) are indicated. Where possible, word forms have been annotated with values of several attributes: polarity (positive, negative), degree of comparison (comparative, superlative), reflexivity (non-pronominal, pronominal), gender (masculine, feminine, neuter), number (singular, plural), and case (nominative, genitive, dative, accusative, instrumental, locative, vocative). The same morphological information has also been associated with the appropriate list of affixes (suffixes and endings) that vary to produce inflected forms.

The data have then been compiled automatically into a lexical database that can be used directly to produce the "lingware files" that make up the Lithuanian analyser in ARIANE-H. Basically, this database was obtained by transforming:

Table 1. Extract from the dictionary table

Id	Lemma	Morphological information	Paradigm
1	abatinis	FSAdjP	ADJ001
2	abdominalinis	FSAdjP	ADJ001
3	abejingas	FSAdjP	ADJ002
4	abejotinas	FSAdjP	ADJ002
5	abiotinis	FSAdjP	ADJ001
6	abipusis	FSAdjP	ADJ001
7	abonentinis	FSAdjP	ADJ001
8	abraomiškas	FSAdjP	ADJ002
9	abrazinis	FSAdjP	ADJ001
10	absoliutus	FSAdjP	ADJ004
…	…	…	…

- lemmas and morphological information into a *dictionary table* (see Table 1) containing lemmas and associated morphological information (expressed using so-called *ATEF formats* that are simple property lists, or *decorations* in Ariane terminology),
- endings and their associated morphological information into a *paradigm table* (see Table 2).

Table 2. Extract from the paradigm table

Id	Ending	Morphological information	Paradigm	Nb Char[a]
1	is	FAD1MSNN	ADJ001	2
2	io	FAD1MSNG	ADJ001	2
3	iam	FAD1MSND	ADJ001	2
4	į	FAD1MSNA	ADJ001	2
5	iu	FAD1MSNI	ADJ001	2
6	iame	FAD1MSNL	ADJ001	2
7	iam	FAD1MSNL	ADJ001	2
8	i	FAD1MSNV	ADJ001	2
9	iai	FAD1MPNN	ADJ001	2
10	ių	FAD1MPNG	ADJ001	2
…	…	…	…	…

[a] The "Nb Char" column contains the number of characters that have to be removed from the end of the lemma to build the radical that will be put in the ATEF dictionaries.

Here are several examples of how this is made for several other analysers (examples given for the inflectional analysis).

For **French**, we transformed two tables of a database built by Sylviane Chappuy, which contained (1) a list of words with their morphological paradigms, and (2) the paradigms themselves (the endings for each existing person, gender, number, tense…). This morphological analyser has been developed in the Traouiéro ANR project [3].

For **Russian**, we started from the NooJ lexical data built by Vincent Bénet [4], which contains Zaliznyak's dictionary.

The ATEF “variables” file DVM + DVS was derived from the _properties.def file. For example:

NooJ: “*A_Forme = fc | fl | adv;*” → ATEF: “*A_Forme: = (fc, fl, adv).*”.

The ATEF radicals file was derived from the NooJ dictionary file. For example:

NooJ: “*багреный,A + FLX = новый*” → ATEF: “*багрен ==P1 (A,багреный).*”,

where багрен is the radical obtained by removing a number of characters corresponding to the highest <Bi> in the новый paradigm, P1 is the *morphological format* (it triggers the analysis rules) corresponding to the новый paradigm, A is the *syntactic format* (the combination of P1 and A contains the lexical information of the NooJ entry) and багреный is the *lexical unit* or LU. In many analysers for MT, the LU is a derivational class, but in this analyser, it is simply the lemma[1].

The grammar rules (GRAM component) and the endings dictionary are derived from the NooJ paradigms file _russe-morph.nof. The other ATEF files — the morphological formats file FTM (these formats trigger the rules) and the syntactical formats file FTS (which contain the lexical information) — are also derived from the NooJ dictionaries.

For **Quechua**, we started from the lexical data built by Maximiliano Duran. For many years, Duran compiled a bilingual dictionary between the Ayacho dialect, an agglutinative and under-resourced language, and French. We derived the radical file from this data, and the other ATEF files were written manually from the information detailed in his PhD thesis: parts of speech, suffixes… [5].

For **German**, Jean-Philippe Guilbaud directly writes in ATEF [6]. In June 2016, his analyser contained 18,219 verbs, 142,321 nouns and 21,747 adjectives, totalising 182,725 different lemmas.

For **Portuguese**, Paltonio Daun Fraga has also written the system directly in ATEF. From the Portuguese system, he derived a **Spanish** analyser in a very short period of time (less than six months) that even outperforms the Portuguese one.

For several Southeast Asian languages, a group of language experts scattered in many places around the world joined their efforts to develop a set of small but consistent analysers for Khmer, Lao, Myanmar, Thai, Tibetan and Vietnamese. The linguistic scope of this project is limited to the text of Saint Exupéry's “Little Prince”. Going beyond this reduced perimeter, Vincent Berment and Guillaume de Malézieux are developing morphological analysers for **Lao** and **Khmer** with broader coverage.

[1] The RUS-FRA MT system built in the 70's by N. Nédobejkine in Ariane-G5 contains a very good MA for Russian, where the LUs are indeed derivational families. Its 13000 LUs correspond to about 40,000 lemmas, themselves corresponding to about 400,000 different accented word forms.

4 Named Entity Extraction

It is easy, with the ATEF language, to describe exhaustively all closed classes. By cons, if the affixes dictionaries may contain the full list of endings, prefixes and suffixes of the concerned language (grammatical morph[eme]s), the lexemes of the language constitute an unbound set, hence the lexical dictionaries can never be exhaustive. The "unknown word problem" is a recurring unavoidable phenomenon.

To handle it, we use the possibility offered by ATEF to write a whole subgrammar to handle unknown words. That subgrammar is triggered by the obligatory MODINC morphological format, and must contain at least a special rule, MOTINC, that is guaranteed to produce at least one result (it unconditionnally produces as LU value the input form itself and stops). When the analysis of a form fails, ATEF restarts it in a special configuration, as if the empty string had been segmented as a prefix, and would be associated with the MODINC morphological format (and hence all rules callable by it) in the dictionaries.

The MODINC subgrammar can be very simple (containing then only the MOTINC rule), or it can implement an elaborate strategy, for example to handle some classes of proper nouns, acronyms, neologisms, etc. For example, a verbal neologism such as "lispified" (transformed into LISP) can be assessed to be the participle past of an unknown verb "lispify", thanks to a normal Markov rewriting method that produces the hypothetical lemma with a few extra ATEF rules and a dictionary of special affixes obtained by a systematic transformation of the subset of normal affixes which are supposed to intervene in the inflectional morphology of unknown words [7].

5 Chunking, Parsing, and Coreference Resolution for Disambiguation

In order to process separate particles (such as the particle "*an*" in the German verb "*ankommen*") and also to disambiguate to some extent the output of the lemmatizer, we can use a sequence of two specialized modules after the ATEF phase: a first module written in EXPANS and a second one written in ROBRA[2].

The EXPANS module contains a dictionary whose entries are the base verbs accepting separable particles (e.g. "*kommen*"). For each such base verb, the dictionary provides a tree containing as many leaves as there are possible combinations of "particle + base verb" (e.g. "*an*" + "*kommen*"). Each leaf is actually a decorated structure containing the new value of lexical unit corresponding to the combination (e.g. "*ankommen*") together with a tactical variable used for coding the particle.

Then, the next (ROBRA) module executes a grammar that looks for the separable particles in the sentence and compares them with the expected values of particles for the processed verb tree. When the correct candidate is found, the others leaves are removed from the tree. This disambiguation process, able to recognize compound

[2] EXPANS and ROBRA are specialized languages of ARIANE, just as ATEF.

words and verbs with separate particles, is implemented by Jean-Philippe Guilbaud in his German morphological analyzer (AMALD).

6 Access Through an API

LEXTOH. Ying ZHANG has developed LEXTOH, a middleware to call morphological analysis web services, and then normalize, merge and filter the results.

7 Conclusion

Reusing software and relying on a community help make the efforts for developing new morphological analysers more efficient. Beyond the most advanced analysers presented in this paper, several prototypes are currently being developed for Ngazidja (the Comorian dialect of Gran Comoros), Swahili, Somali, and Breton. The “Little Prince” project is another approach to help language experts developing new systems, especially for the under-resourced languages on which we are focusing.

References

1. Berment, V., Boitet, C.: Heloise — An Ariane-G5 compatible environment for developing expert MT systems online. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING), Mumbai, 9 p (2012)
2. Marcinkevičienė, R.: Tekstynų lingvistika: teorija ir praktika [Corpus Linguistics: Theory and Practice]. Darbai ir Dienos **24**, 7–64 (2000)
3. Chappuy, S., Guilbaud, J.-P., Berment, V.: T7o — Lemmatiseur du français FR4 en ATEF avec 100 000 lemmes. L4.2.b, deliverable L234.1, Traouiero ANR project, 8 p, 15 June 2011
4. Bénet, V.: Conception et réalisation de ressources lexicales et grammaticales pour le russe. Semaine NOOJ Inalco, 40 p, 31 January 2012
5. Duran, M.: Dictionnaire electronique de verbes français-quechua pour le TAL. Thèse de doctorat en linguistique, 286 p (2016). (should be defended before the COLING 2016 conference)
6. Guilbaud, J.-P., Boitet, C., Berment, V.: Un analyseur morphologique étendu de l’allemand traitant les formes verbales à particule séparée. TALN-RÉCITAL 2013, Les Sables d’Olonne, 9 p, 17–21 June 2013
7. Guilbaud, J.-P., Boitet, C.: Comment rendre une morphologie robuste du français encore plus robuste en traitant finement les mots inconnus avec les données disponibles. TALN 1997, Grenoble, 12 p, 12–13 June 1997