THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 12, Number 2/2011, pp. 151–158

FRENCH TEXT PREPROCESSING WITH TTL

Amalia TODIRAŞCU[1], Radu ION[2], Mirabela NAVLEA[1], Laurence LONGO[1]

[1] LILPA, Université de Strasbourg, France
[2] RACAI, Romanian Academy, Romania
E-mail: todiras@umb.u-strasbg.fr

In this paper we present some experiments on the building of French resources for the TTL POS tagger (Ion, 2007). TTL is a collection of interconnected text preprocessing modules (sentence splitter, tokenizer, tagger, lemmatizer and chunker) with resources for Romanian and English but with no resources available for French. We show how we develop the required POS tagging training corpus and that the average POS tagging accuracy for French exceeds 97% when TTL is trained on this corpus.

Key words: POS tagging; Lemmatization; Tokenization; Chunking; Language models for French; Statistical language processing; Web services; Annotated corpora.

1. INTRODUCTION

Part Of Speech (POS) tagging is an elementary processing step for any Natural Language Processing (NLP) application such as document indexing, text summarization or machine translation. The quality of the results of (POS) tagging and lemmatization is very important for the processing steps that are next in line. Tagging errors will propagate at more elaborate processing steps such as syntactic or semantic analysis and it is thus important to obtain a high rate of precision in order to avoid this error propagation. Most of the work in POS-tagging relies on the availability of high-quality training data and concentrates on the engineering issues to improve the performance of learners and taggers. Building a high-quality training corpus is a huge enterprise because it is typically hand-made and therefore extremely expensive and slow to produce. A frequent claim justifying poor performance or incomplete evaluation for POS taggers is the dearth of training data. Efforts to improve accuracy of taggers for French are still carried on: MeLT (Denis and Sagot, 2010) is a state-of-art tagger for French using a large lexicon while Febril (Sedah et al, 2010) is a new version of Brill's POS tagger (Brill, 1995). Several research projects aim to improve the quality of tagging so as to improve dependency parsers results (Bohnet, 2010), (Favre et al, 2010), (Crabbé et Candito, 2008). Along with existing taggers for French such as TreeTagger (Schmidt, 1994), these programs achieve state of the art accuracy (around 97% correct word annotations) but with the major drawback that the tagsets they use are very poor when it comes to the morpho-syntactic information they encode (the embedded information consists only of POS and a few morpho-syntactic properties at best – like mode and tense for verbs). That said, there is no publicly available tool that is able to provide detailed POS tagging at a level where the morpho-syntactic information contained in a POS label is satisfactory for advanced NLP applications.

In this paper we present results on building of French resources for the TTL POS tagger (Ion, 2007). TTL is a collection of interconnected text preprocessing modules (sentence splitter, tokenizer, tagger, lemmatizer and chunker) with resources for Romanian and English but with no resources available for French. The tagset that TTL utilizes is MULTEXT-East compliant (Erjavec, 2010), a tagset that is best known by its extensive, language independent morpho-syntactic description (MSD) of each POS label. For the POS tagging module, TTL needs an MSD annotated training corpus and the methodology adopted to obtain such a corpus for French was to employ TreeTagger. But, as mentioned, the tagset of TreeTagger lacks an important part of the morpho-syntactic information and thus, we complemented its output with Flemm (Namer, 2005) which gives, for each POS label of TreeTagger, a set of detailed morpho-syntactic

properties. The combined output of TreeTagger/Flemm was manually corrected and disambiguated to obtain MSDs because even if Flemm proposes detailed information, some features are underspecified (the category of the adjectives and the gender of the nouns for instance).

Following a short description of TTL and its usage as a web service, we present the method we used to build a French language model for TTL, the corpus (with its annotation guide) used for training, the most frequent tagging and lemmatization errors, a first evaluation on the French data and a comparison with other existing POS taggers.

2. TOKENIZING, TAGGING AND LEMMATIZING (TTL) FREE RUNNING TEXTS

TTL (Ion, 2007) is a language independent, text preprocessing module developed in Perl. Its functions are: Named Entity Recognition (NER), sentence splitting, tokenization, POS tagging, lemmatization and chunking. The NER function is included as a preprocessing stage to sentence splitting because end of sentence markers may constitute parts of an NE string (i.e. a period may be a part of an abbreviation). POS tagging is achieved through the Hidden Markov Modeling (HMM) tagging technology. The POS tagger of TTL follows the description of HMM tagger given in (Brants, 2000) but it extends it in several ways allowing for tiered tagging (Tufiş, 1999; Ceauşu, 2006), for a more accurate processing of unknown words and also for tagging of named entities (which are practically labeled by the NER module before actual POS tagging). The TTL's tag-set is the MSD[1] with its smaller superset CTAG (TTL tagging methodology follows the tiered tagging approach where MSDs are recovered from an initial CTAG annotation). Lemmatization is achieved after POS tagging by lexicon lookup (in general, a word form and its POS tag uniquely identify the lemma). In the case of out-of-lexicon word forms the lemmatization is performed by a statistical module which automatically learns normalization rules from the existing lexical stock (for details see (Ion, 2007)). Finally, chunking is implemented with regular expressions over sequences of POS tags. It is not recursive and it does not perform attachments (PPs to NPs for instance).

In the most recent web service implementation of TTL, we adopted the Apache[2] web server as host for the web service interface to TTL's functions. The main advantage of using Apache and the Fast CGI[3] module is that the language resources needed at runtime (the language model for POS tagging, the lemmatization model for out of vocabulary words, etc.) are loaded *only once* (an operation with a high time penalty cost) for each TTL instance. Another advantage of this setup is that the Fast CGI module performs automatic load balancing on the server supported by a complex management of the web service instances.

The TTL web service offers the following remote procedures (these are the actual names from the WSDL file which is located at http://ws.racai.ro/ttlws.wsdl):

1. `SentenceSplitter` which takes as parameters the language of the text to process (currently either "en", "ro" of "fr") and text and returns another string which is a list of sentences separated by carriage return/line feed sequence ("\r\n");
2. `Tokenizer` which has as parameters the language code and a sentence and returns a list of tokens separated by "\r\n" each token possibly carrying its NE tag (added to the token with the tab character "\t") given by the NER module of the SentenceSplitter in the case the token is a NE (i.e. a real or integer number, a roman number, percents, abbreviations, dates, clock times, etc.);
3. `Tagger` which takes the language code and a tokenized sentence from Tokenizer and returns a MSD POS tagged sentence which is a string with triples of token, "\t", MSD separated by "\r\n";
4. `Lemmatizer` uses the POS tagged sentence along with the language code and returns a lemmatized sentence which resembles the one from the Tagger's output except that the token annotation is enriched with its lemma which is separated again by a "\t" from the MSD tag;
5. `Chunker` is the final operation of TTL and, besides the language code, it takes a lemmatized sentence and returns the same sentence with chunk information added after the lemma annotation;

[1] http://nl.ijs.si/ME/V4/msd/
[2] http://httpd.apache.org/
[3] FCGI, http://www.fastcgi.com/

6. `XCES` is a helper function which calls all the previously mentioned operations and returns an XML representation of the result.

In principle, TTL operations are to be pipe-lined from 1 to 5, `SentenceSplitter` which takes the actual text as parameter being the first function call, `Tokenizer` the second function call, and so on till the `Chunker` operation. Since TTL operates with SGML entities and not UTF-8 representation of the text, the user is required to transform the input text from UTF-8 to SGML by calling `UTF8toSGML` helper function of the TTL web service and convert the response back to UTF-8 with the reverse function `SGMLtoUTF8`. The conversion cannot be automatically made because the web service cannot know how many calls are stacked and thus, when to convert back to the UTF-8 encoding.

3. BUILDING FRENCH LANGUAGE RESOURCES FOR TTL

In order to build a TTL model for French texts processing, we take the following steps:

a) select a large corpus (about 1 millions of tokens). The corpus contains newspaper articles and text from European legislation;
b) tag and lemmatize this corpus with existing tools (TreeTagger for a first tagging step and lemmatization and Flemm to complete the tagset with MULTEXT-East MSDs and to modify the lemma if necessary);
c) manually check the tagged and lemmatized corpus to eliminate ambiguities (verb forms, incorrect lemmas, etc.);
d) identify systematic tagging errors and define rules to correct these errors;
e) train TTL with the existing corpus;
f) check the output of TTL and restart the procedure at step c).

In the next subsections we present the corpora used for training, the systematic errors and some rules we identified to correct these errors.

3.1. The Training Corpus

Our goal is to use the POS tagger as a preprocessing step for lexical alignment of parallel corpora and a prerequisite to that is to have comparable morpho-syntactic information annotation throughout the corpora. Our parallel corpora used for training a French-Romanian Machine Translation system, are extracted from the "Acquis Communautaire" corpus available in 22 languages of the European countries (Steinberger et al, 2006). For this reason and due to the fact that POS tagger's performances are dependent on the training corpus, we selected texts from both the "Acquis Communautaire" corpus and from a freely available corpus of newspapers ("L'Est Républicain"). In Table 1 we give the composition of our training corpus. "Acquis Communautaire" contains law texts adopted by European member states since 1950. Its style is specific to administrative, official texts. Each document is structured in articles, paragraphs and contains a lot of enumerations. "L'Est Républicain" is a newspaper journal, freely available for research projects (CNRTL – Centre National de Ressources Textuelles et Lexicales[4]) which contains general language words. L'Est Républicain contains news, local events and obituary or marriage announcements. We selected mainly news, local events and some announcements, from the available articles from 2003.

Table 1

The training corpus

Source	Number of words
Acquis Communautaire	498 889
L'Est Républicain (2003)	387 674

[4] http://www.cnrtl.fr/

We ran TreeTagger on each of these corpora but due to some tokenization errors (dots are considered as parts of the previous words if there is no space separating them), we had to modify the tokenizer module of TreeTagger to improve the tokenization. TreeTagger uses a set of 33 POS tags describing the lexical category and some morpho-syntactic properties (i.e. tense or mood for verbs). After tagging and lemmatization, we applied Flemm to obtain detailed morpho-syntactic descriptions and to correct lemmas. Flemm proposes several rules for lemmatization based on the initial POS guessed by TreeTagger: a set of rules to guess the lemmas for regular cases and some lemma exceptions.

MULTEXT-East offers standardized annotation guidelines for corpus linguistic annotation. Thus, for each POS, we have a standardized manner of representing the various morphological features (called a Morpho-Syntactic Description or MSD):

```
Ncfs: N – noun; c – common; f – feminine; s - singular
Vmns: V – verb; m – main; n – infinitive; s – singular
```

POS tagging with this detailed tagset implies a decrease of the labeling accuracy of the tagger. Indeed, a detailed tagging incurs an important number of ambiguities (for example a verb form at present tense is similar for conjunctive or for indicative mood) that complicate the task of the tagger. In order to obtain a good MSD-annotated training corpus, one should make sure that difficult ambiguous cases are statistically well-represented in the corpus.

3.2. Manual Validation

After automatic tagging and lemmatization we manually check the tags and the lemmas proposed for each word form. Each corpus has been validated by two human annotators. As a common annotation guide, we instructed them to be as consistent as possible in the annotation process: the same label is to be assigned to similar words in similar contexts.

We identified several difficult POS tagging cases such as participial adjectives and past participles, identification of proper noun categories (e.g. deciding if a named entity is an abbreviation or an organization), identification of borders of named entities, etc. For instance, the MSD of proper nouns does not provide specific proper noun categories such as places, persons or time periods. We decided to consider as named entity some organizations that were very frequent in the corpus such as "Agence pour l'énergie atomique" or the "Parlement européen", but we do not annotate other entities such as book titles or product names. In addition, for tokenization reasons, we build a list of named entities to be annotated as a single unit by the tokenizer.

In addition, we propose some new tags. A difficult case is the case of aggregates: *du (de+le), des (de+les)* which are partitive determiners or simple prepositions followed by a determiner. For aggregates, we propose a new POS tag (Dg) and the lemma is the sequence *de+le*. We proceed in a similar manner for interrogative pronouns aggregated with determiners (*lesquels, auxquelles*). This choice has been made also for Europarl corpus (the CorpusEye on-line interface).

We identified several annotation problems due to tokenization errors. Several titles were written in upper case with no diacritics, and thus, the tags and the lemmas were wrong. Titles are not always separated from the body of the article, so these segmentation errors induce tagging and lemmatization errors. These errors were manually corrected.

The systematic errors produced by Flemm are given in Table 2.

We start the validation procedure with the nouns. To complete gender, we use contextual information: determiners and adjectives. Also, we use the determiner or the adjective for specifying the gender of the noun forms which are similar for plural and singular. Another systematic error is the tagging of each participial adjective as past participle verb. We propose rules to replace the past participle tag with the participial adjective tag if the verb modifies a noun and there is no auxiliary verb in the neighborhood of these elements. In these cases, the lemma might be wrong and the tag of the past participle should be changed into ADJ : A—fp :

```
                    Les          DET(ART):Da3-p      le
                  règles        NOM:Ncfp            règle
                 communes      ADJ:A--fp           commun
visées       VER(pper):Vmps-pf   viser  => visées          ADJ:A—fp   visé
```

Some of the personal pronouns tags contain information about the syntactic function of the pronoun. The syntactic information is influenced by the syntactic context. The lemma might also be wrong. We decide to not specify information about the syntactic function of the pronoun. We correct the lemma as follows:

```
laquelle     PRO(REL):Pr3fs--    laquell  =>  laquelle PRO(REL):Pr-fs-- lequel
```

Table 2

Systematic errors produced by Flemm, ordered by POS.

Noun	gender	No information about gender
	number	When plural and singular form are identical
Adjective	type	No type for qualifiers
	participial	Tagged as past participle tags
Pronoun	Syntactic function	Wrong function
Determiner	gender	For plural determiners or aggregates
	type	No information about the type
Verb	ambiguities	Several forms various tenses and mood
	type	Main vs. auxiliary

Adjective categories are completed with a list of indefinite adjectives (“certains“, “même“, “aucun“), and they are always positioned before the modified nouns. For the qualificative adjectives, this information is completed if the adjective is not an indefinite article and the gender is completed with the help of the modified noun.

```
 autres       ADJ:A---p          autre  => autres         ADJ:Ai-mp  autre
membres      NOM:Ncmp          membre => membres        NOM:Ncmp   membre
```

Determiners might be also wrongly annotated as other POS. For instance, demonstrative determiners are defined systematically as demonstrative pronouns. This error could be recovered if the determiner is followed by a noun:

```
ce            PRO(DEM):Pd3msn     cet   => ce              DET(DEM):Dd-ms cet
                     cas          NOM:Ncms           cas
```

Verb forms are ambiguous: one form could have up to 8 different tags. The choice is manually done and there is no systematic way of recovering this information. Actually, even if we manually corrected these information, confusion between mood and tense are still quite frequent. Also, a frequent error is to consider the verb avoir/'to have' and être/'to be' as main verbs, whereas these verbs were auxiliaries. We propose some contextual rules to correct this error as well.

Another error which is difficult to correct and which represents a source of ambiguities is the case when a verb form could be a form of several lemmas:

```
convient     VER(pres):Vmip3p--1 convier ||
convient     VER(pres):Vmip3s--3 convenir ||
convient      VER(pres):Vmsp3p--1 convier
```

In this case, it is difficult for to select the correct lemma (*convier* or *convenir*). For these cases, we propose a list of frequencies of lemmas, after checking the frequency of the lemmas on some existing corpora (Frantext, Europarl). The proposed lemma is then most frequent lemma found in the training corpus. For example, *faut*/must is the 3rd person, singular form of *falloir* and *faillir*, but the most frequent occurrences are the forms of the verb *falloir*.

Some errors are caused by some sequence of ambiguous words (some determiners and personal pronouns might have the same form but various POS). If in a sequence of words, all of them are ambiguous, then the tags might be wrongly assigned. A noun modified by two adjectives might be tagged as a noun modified by two adjectives preceding the noun, if these nouns are ambiguous as in the following example:

```
la           DET(ART):Da-fs--d   le
politique    ADJ:Af-fs--         politique
agricole     ADJ:Af-fs--         agricole
commune      NOM:Ncfs--          commune
```

The word `politique` should be tagged as noun while `commune` should be tagged as adjective. These errors were automatically identified and, after selecting all the sequences DET ADJ ADJ NOM, we manually correct these tags.

Other errors concern wrong identification of prepositions. For some ambiguous forms (`entre 'enter'`- might be a verb but also a preposition, A at the beginning of the sentence might be a preposition or a verb), the verb is marked as preposition

```
entre        PRP          entre
en           PRP          en
vigueur      NOM:Ncfs--   vigueur
```

We defined rules to automatically change the POS of `entre` to verb instead of PRP.

For some of these errors, we define rules to automate their correction. For most of the cases, a manual validation is available.

3.3. Evaluation

Few corpora or tools are available for French providing annotations in MULTEXT-East format. To the best of our knowledge, only Flemm provides MULTEXT tags but introduces many ambiguous cases (for verb forms) and only the JOC corpus (Véronis and Langlais, 1997) is annotated in MULTEXT format (approx. 200K words for French).

Firstly we compare the output of TTL with the hand-annotated corpus used for training. For both law text corpus and newspaper corpus the precision is 97.92% and respectively 98,10%. The most frequent errors are the confusion between past participles and participial adjectives, the lack of disambiguation of determiner gender for “les”, “l’”, partitive articles and determiners, wrong lexical tagging (pas is an adverb or a noun), lemmas errors for ambiguous forms.

Secondly we compare the output of TTL with that of TreeTagger which is one of the most popular POS taggers available for Modern and Ancient French. For our evaluation, we decided to train TreeTagger on a small part of our training corpus (approx. 270K tokens), the same part used to train the language model for TTL. We evaluate the two POS taggers on a small corpus of approx. 15K tokens composed of an excerpt of “Acquis Communautaire”, an excerpt from the newspapers corpus, an excerpt from a novel by Alexandre Dumas and an excerpt of a computer science corpus (extracted from the Web, containing more specialized language). The results are displayed Table 3 and we noticed that they are comparable with existing state-of-art French taggers such as MeLT with the important difference that TTL runs with the MULTEXT-East tagset. When training on Acquis and L'Est Républicain, we obtain good lemmatization precision, while TreeTagger has worst results for tagging and especially for lemmatization. Some frequent errors of TreeTagger consist of not recognizing correctly proper nouns (if these nouns are lexical entries), neither their lemma. TTL does not propose the <unknown> tag. Meanwhile, for a specialized corpus, from the computer science area, TreeTagger obtains worse results than TTL. These results are explained by the number of unknown words found in the texts, but also due to some segmentation errors (no clear distinction between the title and the beginning of the paragraph).

Table 3

A comparison of TTL and TreeTagger tagging and lemmatization

	TTL tagging precision	**TTL lemmatization precision**	**Tree-tagger precision**	**TreeTagger lemmatization**
Acquis	97.31%	98.74%	95.60%	96.86%
Dumas	97.01%	98.01%	95.81%	96.60%
L’Est	97.22%	98.00%	96.12%	97.00%
Computer Science	97.00%	97.45%	94.85%	93.00%

The errors are complementary: TTL fails to correctly identify determiners such as *les, l'*, because it is not able to decide about the gender of the determiner (for *les*, the tag is *Da-mp, Da-fp*), while TreeTagger

correctly disambiguates the determiner. Also, TTL wrongly tags auxiliary verbs as main verbs, while TreeTagger works better. For verb forms, several TTL errors concern the use of subjunctive form, while the correct mood is the indicative. A small set of errors are common: both taggers fail to identify either the POS or the lemma for participial adjectives, determiners and aggregates.

4. CONCLUSIONS

We have presented a general methodology to build training data with a fine-grained tagset, containing detailed morpho-lexical information, essential for a multitude of NLP applications. TTL is a language independent text pre-processing toolkit written in Perl that is able to work with large (fine-grained) tagsets and is available as a web service (WSDL file is located at http://ws.racai.ro/ttlws.wsdl). The methodology presupposes that an MSD annotated, lemmatized and human-validated corpus is available from which the POS tagging language models are to be learnt. In addition to training corpora, a wordform-lemma-MSD lexicon is also needed for the lemmatization module which can initially be derived just from the training corpus. The multilingual processing platform which integrates TTL has been developed for several years and it is currently used in several multilingual projects such as machine translation or cross-lingual question answering. Adding new languages to the public web-services platform (www.racai.ro/webservices) is a constant preoccupation at RACAI (we plan to include German as well) in accordance with the international research priorities.

The French language integration into TTL has been successful as TTL is now able to POS tag (with more than 97% accuracy), lemmatize and chunk French texts (besides English and Romanian). We will continue to improve the French resources (adding new entries in the French wordform-lemma-MSD lexicon, extend the training corpus, etc.), in order to have TTL perform even better on this language, on par with Romanian and English.

ACKNOWLEDGEMENTS

The research reported here was partially supported (for the second author) by the STAR project, financed by the Ministry of Education, Research and Innovation (UEFISCDI) under the grant no. 742/19.01.2009. We also gratefully acknowledge the support of University of Strasbourg for the rest of the authors.

REFERENCES

1. BRANTS, T., *TnT – A Statistical Part-Of-Speech Tagger*, Proceedings of the 6th Applied NLP Conference ANLP-2000. Seattle, WA, pp. 224–231.
2. BOHNET, B., *Top Accuracy and Fast Dependency Parsing is not a Contradiction*, Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.
3. BRILL, E., *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging*, Computational Linguistics, **21**, pp. 543–565, 1995.
4. CRABBE B., CANDITO M.-H., *Expériences d'analyse syntaxiques statistiques du français*, Proceedings of TALN 2008, Avignon, 2008.
5. CEAUŞU, A., *Maximum Entropy Tiered Tagging*, in Janneke Huitink & Sophia Katrenko (editors), Proceedings of the Eleventh ESSLLI Student Session, ESSLLI 2006, June 2006, Malaga, Spain, pp. 173–179.
6. ERJAVEC, T., *MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*, Proc. of the LREC 2010, Malta, 19-21 May, 2010.
7. FAVRE, B., BOHNET, B. HAKKANI-TÜR, D., *Evaluation of Semantic Role Labeling and Dependency Parsing of Automatic Speech Recognition Output*, Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP), Dallas (USA), 2010.
8. DENIS, P., SAGOT, B., *Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français*, Actes de Traitement Automatique des Langues Naturelles (TALN 2010), Montreal, Canada, 2010.
9. ION, R., *Word Sense Disambiguation Methods Applied to English and Romanian* (in Romanian), PhD Thesis, Romanian Academy, Bucharest, 2007.

10. NAMMER, F, *La morphologie constructionnelle du français et les propriétés sémantiques du lexique : traitement automatique et modélisation*, Université Nancy 2: Mémoire d'Habilitation à diriger des recherches, 2005.
11. SCHMIDT, H., *Probabilistic Part-of-Speech Tagging Using Decision Trees*, International Conference on New Methods Language Processing, Manchester, UK, 1994, pp. 44–49.
12. SEDDAH, D., CHRUPAŁA, G., CETINOGLU, O., GENABITH J., CANDITO M., *Lemmatization and Lexicalized Statistical Parsing of Morphologically-Rich Languages: the Case of French,* Proceedings of the NAACL-HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2010), Los Angeles, June, 2010.
13. STEINBERGER, R., POULIQUEN, B., WIDIGER, A., IGNAT, C., ERJAVEC, T., TUFIŞ, D., *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*, Proceedings of the 5th LREC Conference, Genoa, Italy, 22–28 May, 2006, pp. 2142–2147.
14. TUFIŞ, D., *Tiered Tagging and Combined Classifiers*, in F. Jelinek, E. Nth (Eds.), *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*. Springer, 1999, pp. 28–33.
15. VERONIS J. and LANGLAIS, P. ARCADE, *Evaluation de systèmes d'alignement de textes multilingues*, ARCADE report, 1997.

Received February 10, 2011