Learning Morphology of Romance, Germanic and Slavic languages with the tool *Linguistica*

Helena Blancafort[1,2]

[1]Syllabs
15, rue Jean Baptiste Berlier, 75013 Paris, France
blancafort@syllabs.com
[2]Universitat Pompeu Fabra
Roc Boronat,138, 08018 Barcelona, Spain

Abstract

In this paper we present preliminary work conducted on semi-automatic induction of inflectional paradigms from non annotated corpora using the open-source tool *Linguistica* (Goldsmith 2001) that can be utilized without any prior knowledge of the language. The aim is to induce morphology information from corpora such as to compare languages and foresee the difficulty to develop morphosyntactic lexica. We report on a series of corpus-based experiments run with *Linguistica* in Romance languages (Catalan, French, Italian, Portuguese, and Spanish), Germanic languages (Dutch, English and German), and Slavic language Polish. For each language we obtained interesting clusters of stems sharing the same suffixes. They can be seen as mini inflectional paradigms that include productive derivative suffixes. We ranked results depending on the size of the paradigms (maximum number of suffixes per stem) per language. Results show that it is useful to get a first idea of the role and complexity of inflection and derivation in a language, to compare results with other languages, and that it could be useful to build lexicographic resources from scratch. Still, special post-processing is needed to face the two principal drawbacks of the tool: no clear distinction between inflection and derivation, and not taking allomorphy into account.

1. Introduction

The development of morphosyntactic lexica is a labour-intensive task and the time needed for building this type of resource is difficult to evaluate.

For this reason, work has been carried out for inducing supervised and unsupervised induction of morphological rules using non annotated corpora and as little supervision as possible. This area is of special interest for our research, since in the long run we expect to define a roadmap to predict the difficulty of a language for morphosyntactic processing and to evaluate the difficulty of building the necessary resources. In Blancafort and Loupy (2009) we outline some clues as a result of several experiments run only on corpora and further ones run on corpora by using morphosyntactic information from lexica. In this paper, we want to explore whether large parallel corpora as well as a tool for inducing morphology without any other knowledge can already provide some information about a language, at least for Romance and Germanic languages. The tool used is *Linguistica*[1] (Goldsmith 2001, 2006), open-source software for inducing morphology automatically.

The present article is organized as follows: section 2 briefly reviews the state of the art; section 3 describes the tool Linguistica and discusses some problems. Next, we report on a series of corpus-based experiments run with Linguistica in Romance languages (Catalan, French, Italian, Portuguese, and Spanish), Germanic languages (Dutch, English and German), and Slavic language Polish. Finally, we draw some conclusions and discuss further work.

[1] The tool can be downloaded at the following URL: http://linguistica.uchicago.edu/downloads.html, we used the version 3 for Windows.

2. State of the Art

Recently, some work has been done on the induction of morphology from large corpora using machine-learning approaches and as little supervision as possible. The general goal is to induce morphological information from raw data. The expected output varies from author to author: the obtained morphological information can be limited to a simple list of affixes or may be more sophisticated as a cluster of stems associated to a cluster of affixes. First work in this area concentrates on obtaining affix inventories, mainly applying minimum description length (MDL) (Brent et al. 1995; Kazakov, 1997). MDL is a model introduced by Rissanen (1978) used in information theory and statistical NLP. It can be used for calculating the compression of the data and considers the best hypothesis the one with the largest compression of the data and with the smallest model length. Another strategy to identify the end of a stem (Déjean, 1998) is based on work carried out by Harris work.

More recent work (Goldsmith, 2001; Nakov et al.; 2003; Oliver, 2005; Goldsmith, 2006, Monson et al.; 2007; Loupy et al.; 2009) is more ambitious and aims at finding clusters of stems with their corresponding affixes or even at suggesting inflection paradigms (lemma candidates with all inflected forms and possible morphosyntactic tags).

Concerning the input data, some authors report work on raw data without linguistic knowledge (Jacquemin, 1997; Schone and Jurafsky, 2001), others include linguistic knowledge to improve results as explained later in this section. The amount and kind of linguistic data is different depending on the authors and the expected output.

Unsupervised induction without linguistic knowledge

Schone and Jurafsky (2001) suggest an algorithm for inducing inflection rules in German, English and Dutch from a corpus without any human intervention nor linguistic knowledge. Their algorithm combines different clues to induce morphology: a Latent Semantic Analysis approach to calculate the semantic relatedness of the affixed forms, affix frequency, syntactic distribution and orthography. As far as we know, they obtained the best results for a knowledge-free algorithm, an F-score of 88,1% on the identification of words corresponding to a same cluster of inflectional and derivational affixes calculated on the hand-labeled CELEX lexicon (Baayen et al. 1993).

Unsupervised Induction using linguistic knowledge

More recent approaches utilize previous morphological knowledge to improve results.

Nakov et al. (2003) use a German lexicon to learn automatically all possible endings of a word. Then they apply the Maximum Likelihood Estimation (Mikheev, 1997) to generate all possible stems of an unknown word and its morphological class. The same experiment was carried out on Bulgarian, but results were less encouraging due to the difficulty to identify unknown nouns from raw data. This shows that it is more difficult to learn morphology from a language with very rich inflectional morphology with numerous ambiguous endings.

Clément et al. (2004) present work carried out to build a French lexicon from a big corpus using morphological information. They apply a verbal inflection engine developed manually following the inflection patterns for open classes described in French grammars. The basic idea behind is that a hypothetical lemma can be guessed when several words found in the corpus are best interpreted as morphological variants of this lemma. First, they extracted verbs and adjectives from a corpus of 25 million words. Results are very satisfying, because they also cover specific terms not encoded in a general lexicon. However, they are confronted to the problem of incomplete representation of a lemma and corresponding inflection forms in a corpus: as we know, a corpus does not necessary contain all inflection forms of a word, which is a drawback for lexical acquisition. Thus, the generated lexicon contains incomplete paradigms. This problem is addressed in Oliver (2005) who presents work on Croat and Russian. First inflection paradigms are learned automatically using a morphosyntactic lexicon. Then, new words identified in a corpus are associated to those paradigms to enrich the lexicon. To solve the problem of incomplete paradigms, they use the internet to find missing forms of a stem.

Zanchetta and Baroni (2005) generate a morphosyntactic lexicon for Italian called Morph-it using a corpus and a part of speech tagger. First the corpus is parsed with TreeTagger (Schmid, 1994) to obtain the part of speech tag and corresponding lemma of a form. Then the obtained lemmas are inflected using human validated inflection rules. Another interesting approach is presented by Forsbert et al. (2006) who developed a tool to extract pairs of lemma-paradigms from non annotated corpora. The tool requires linguistic input, more precisely hand-written inflection paradigms and a list of function tools. They report on a positive experiment run on a French corpus with only 43 inflection paradigms written with regular expressions containing variables and combined with propositional logic to identify lemmas and assign the corresponding inflection paradigm. Loupy et al. (2008) present work carried out for lexical acquisition in French in order to help the linguist to add new words to the lexicon by suggesting one or more lemma candidates with their corresponding inflection rules as well as morphosyntactic tags. The implemented probabilistic model ranks the candidates such as to reduce the number of rules to validate.

Other authors as Gaussier (1999), Dal and Namer (2000), Namer (1999), Hathout (2005), Hathout and Tanguy (2005) work on learning derivation rather than inflectional morphology. Zweigenbaum et al. (2003) conduct research on morphology induction for terminology purposes in the medical domain.

Goldsmith (2001, 2006) suggests an unsupervised learning of the morphological segmentation of a language that with the exception of capitalization removal and tokenization rules is knowledge-free. Specific tokenisation rules can be defined in the preferences of the tool. Goldsmith applied the same algorithm to various languages but just evaluated results obtained for English and French. In the next section we will present his tool Linguistica for the unsupervised induction of morphology.

3. *Linguistica*: How it works

First, Linguistica computes a set of heuristics to produce rapidly a probabilistic morphological grammar. Then, it uses minimum length description (MDL), the expectation-maximization algorithm (EM) and other triage procedures to help eliminate inappropriate analysis for every word in the corpus.

Signatures

Linguistica uses *signatures* to regroup bases with common affixes. They can be seen as morphological patterns, with a list of affixes that occur with a particular stem in a corpus. Goldsmith (2006) also defines them as sort of miniparadigms. One of their final functions is to help in building constructively a satisfactory morphological grammar.

The algorithm first splits some words in two and treats the first piece as a stem and the second as a suffix. For each stem, it builds a cluster of suffixes. Then, it associates to each cluster of suffixes a number of stems that appear with that cluster of suffixes. Common signatures in English look as follows:

- NULL.s → primarily for nouns
- NULL.ed.ing.s → for verbs
- NULL.er.est.ly → for adjectives

Figure 1: Examples of Linguisticas' signatures for English

The length of the candidate stems is restricted to a size of three letters. If the suffix is just one letter long, as in the signature NULL.s, it is only accepted as candidate if it occurs with a sufficient number of examples, otherwise, these types of signatures would be too noisy. In addition to this, any signature with more than 25 stems is permitted, while those with fewer stems have to include at least two affixes of at least two characters. These parameters can be modified in the preferences file of the tool.

Moreover, a function called *check signature* using MDL is applied to examine each signature and further heuristics follow until deciding the final signatures. One of the main hurdles that is still to be resolved is allomorphy. For the time being, it constitutes a limitation, as the program is not capable to associate allomorphs as being a variation from another stem, as in Spanish the stems *colg-* and *cuelg-* from the verb *colgar*. In the version used, *Linguistica* only knows putting together allomorphs showing the deletion of word final *–e* in English and spelling changes as final *–y* turning to *–i*, as in the inflected form *studied* from the verb *study*.

4. Experiments and Analysis of the Results

In this section we present results obtained by using the open-source tool Linguistica described in the previous subsection. Our aim is to evaluate which kind of information can be extracted from non annotated corpora with the aid of such a tool and to evaluate if the morphological information induced was coherent to the one obtained by using the lexicons in a previous corpus and lexicons based study to compare inflection paradigms for those languages (Blancafort and Loupy, 2009). In other words, we wanted to evaluate if we could induce morphological information using a tool instead of using morphosyntactic lexicons, as such type of resources are not always freely available and can be difficult to find for under-resourced languages. Results are based on the Bible corpus (Resnik et al. 1999), one of the corpus used in our previous study cited above. The study was conducted for five languages for which a multilingual parallel corpus as well as morphosyntactic corpora were available: English, French, German, Italian and Spanish. In the present study we included other languages: two Romance languages, Catalan and Portuguese, one Slavic language, Polish, and another Germanic language, Dutch. For those languages we only present analysis based on corpora, as we do not have any lexicographic data.

With Linguistica we can obtain different information from raw data: number of suffixes, number of prefixes, number of compounds as well as signatures, a sort of miniparadigms that put together inflections and derivations belonging to a same base. The most relevant and coherent information we obtained using the tool was the maximal length of signatures, i.e., the number of forms associated to a single stem, which can be interpreted as the maximum number of inflections that may contain a paradigm in the lexicon.

We found out that Polish followed by Romance languages had more forms in a signature than German, and that English was the language with the shortest signatures, which can be interpreted as the language with the lowest inflection number per paradigm. We got similar results when analyzing the lexicon. The rank of languages according to the maximum number of forms per signature provided by Linguistica is given in table 1.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Language	pl	Es	cat	it	pt	fr	de	nl	en
max. nb of forms per signature	39	31	29	28	26	24	14	13	9

Table 1: Linguistica's language rank according to the number of suffixes per signature

The next tables show results induced using MulText[2] (Ide and Véronis, 1994), and using FreeLing (Atserias et al., 2006) and Lefff (Sagot et al., 2006) for the languages for which we had the necessary lexicon data. We can see that the rank of number of forms per paradigms based on the lexicons or Linguistica is the same. Thus, without previous resources we can evaluate which language might have more inflections per paradigm.

Rank	**1**	**2**	**3**	**4**	**5**
Language	it	fr	es	de	en
max. nb of forms per paradigm	63	62	55	29	14

Table 2: MulText's language rank according to the maximum number of forms per paradigm

Rank	**1**	**2**	**3**	**4**
Language	it	fr	es	en
max. nb of forms per paradigm	68	62	56	12

Table 3: FreeLing and Lefff language rank according to the maximum number of forms per paradigm

It is obvious that we cannot expect to get the same results nor the same quality from the knowledge-free tool than from a lexicon, as Linguistica does not use linguistic information. One of the problems of Linguistica is that it cannot separate derivation from inflection; this is why signatures are different from paradigms and include both inflected forms and derivatives. A further obvious limitation is due to the nature of corpora: a corpus is

[2] ELRA catalogue (http://catalog.elra.info), MULTEXT lexicons, reference: ELRA-L0010.

incomplete for extracting inflection, as it is very unlikely that all possible inflection forms of a word occur in a single corpus, especially verbal inflections. Moreover, Linguistica cannot regroup irregular forms. However, irregular forms irregular forms constitute a much smaller class than regular forms. In English for instance Quirck et al. (1985) estimate 250 existing irregular verb forms, which means that they can be encoded manually. This is why signatures might be incomplete paradigms, especially signatures regrouping verbal inflections.
The next figure illustrates the size of the paradigms ranked by the number of different suffixes included in a paradigm.

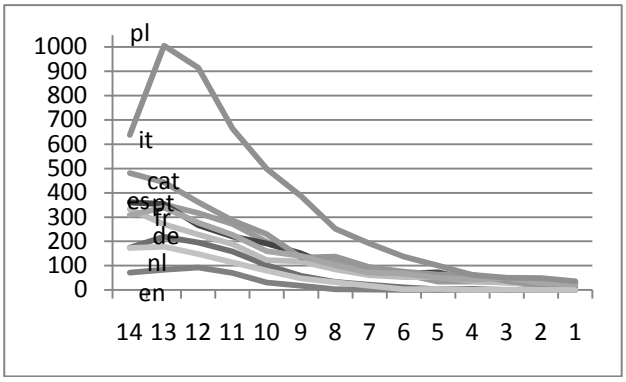

Figure 2: Number of paradigms and number of suffixes included in a paradigm

We can see that Polish has not only the longest paradigm, but that there are more paradigms and that they are longer than in other languages, which means that the number of affixes shared by a stem is higher. Paradigms found for Romance languages are larger than the ones in Germanic languages, while English and Dutch seem to have less productive and varied suffixation.
Table 4 presents the longest signatures found for each language. If we have a closer look to the suffixes in the signatures, we can observe that in Romance languages most of the suffixes correspond to verbal suffixes. The signatures of French, Catalan and Spanish for instance correspond to the inflection of a regular verb (French *répondre*, Catalan *posar* and Spanish *anunciar*), even if same verbal forms are ambiguous and are a noun as well, as *anuncio* in Spanish or *posada* in Catalan. In Italian not all affixes correspond to the same verb, which means that the cluster does regroup suffixes for two different verbs, *menare* and *mentire*, which could pollute a lexicon if we used Linguistica to enrich the lexicon automatically. In Polish we also find affixes for two different verbs *dawać* and *dać*, but they are morphologically related anyway (equivalent to the verb *to give*). However, the paradigm for Polish also includes a noun *dach* that does not have any linguistically motivated relation to the verbal stem *da*. As Linguistica does not make the difference between affixes for inflection, derivation and cliticization, we already expected to have affixes corresponding to different part of speechs in the same paradigm, as Portuguese noun *habitante* in the verbal paradigm for *habitare* and Italian *menalo* consisting of a verb and clitic *lo*. Nevertheless, we observed that Romance and Polish long paradigms correspond to verbal paradigms with some nominalisations or adjectives, whereas in Germanic languages categories are completely mixed. In English for instance, we can observe verbal forms of the verb *light* as well as nominalizations as *lightness* and the adjective *lightly*.

	Nb of affixes	Stem	signature
pl	39	da	NULL.ch.cie.dzą.j.je.jmy.jmyż.ją.jąc.li.liście.liśmy.m.my.na.ne.nej.ni.nie.niu.no.ny.ną.rze.sz.wa.wał.wszy.ć.ł.ła.łby.łbyś.łem.łeś.ło.ły.ń
es	31	anunci	a.ad.ada.adas.adlo.ado.amos.an.ando.ar.ara.arles.aron.aros.arte.ará.arán.arás.aré.as.ase.asen.e.emos.en.es.o.áis.é.éis.ó
cat	29	pos	a.ada.ades.ant.ar.aren.arà.aràs.aré.at.ava.aven.em.en.es.essin.essis.eu.i.in.is.o.t.ta.ts.à.és.éssim
it	28	men	NULL.a.ai.ali.alo.ano.are.ata.ate.ati.ato.ava.erai.eranno.erebbe.erete.erà.erò.i.ino.o.ta.te.ti.to.tre.zione.ò
pt	26	habita	NULL.da.das.do.i.is.m.mos.ndo.nte.r.ra.ram.rdes.rei.reis.rem.remos.res.ria.rà.rá.rás.rão.s.stes
fr	24	répond	NULL.aient.ait.ant.e.ent.es.ez.ions.irent.is.it.ra.rai.rais.rait.ras.re.rez.ront.s.u.îmes.ît
de	14	heil	NULL.e.en.et.ig.los.lose.loser.sam.same.sames.t.te.ten
nl	13	heilig	NULL.de.den.dom.e.en.er.heden.heid.ing.s.ste.t
en	9	light	NULL.ed.en.er.ing.ly.ness.ning.s

Table 4: Longest signatures suggested by Linguistica for a stem

A further drawback for lexical acquisition from corpora is the fact that a corpus does not necessarily contain all the inflected forms of a lemma and thus, it is impossible to output a complete paradigm. For French, we obtained 24 forms for the verb *répondre*, 15 are missing. For Spanish 27 forms are missing for the verb *anunciar*, for Polish 15 forms are missing out of 48 forms. Missing forms in Polish concern feminine verbal forms, especially plural forms. In all Romance languages there are missing forms for the conditional, subjunctive and imperfect tense. Futhermore, Linguistica is not able to put together in a same signature stems that share suffixes as well as prefixes, as prefixation is treated apart. This means that for German and Dutch, the past participle often built with a prefix *ge* is never included in the signature and thus, will be missing in the inflection paradigm.

Moreover, we also compared results concerning the number of suffixes, but they were completely different as the ones induced from the lexicon. The fact that derivation and inflection are not distinguished might partially explain this difference. Another reason is that the lexicon includes suffixes for irregular forms and for all cases of vowel alteration, which explains why the number of suffixes is considerably higher in the lexicon. The only relevant information was that English had a lower number of suffixes than the other languages, as we can see in the tables above.

Rank	**1**	**2**	**3**	**4**	**5**
Language	De	es	fr	it	en
number of suffixes	1106	641	541	502	86

Table 5: Number of suffixes extracted from Multext lexicon

Rank	**1**	**2**	**3**	**4**	**5**
Language	De	es	fr	it	en
number of suffixes	28.844	735	562	542	77

Table 6: Number of suffixes extracted from FreeLing and Lefff lexicons

Rank	**1**	**2**	**3**	**4**	**5**
Language	pl	it	cat	es	fr
number of suffixes	571	409	385	359	317
Rank	**6**	**7**	**8**	**9**	
Language	pt	de	nl	en	
number of suffixes	233	256	201	101	

Table 7: Number of suffixes generated by Linguistica

Learning prefixes

Moreover, it is worth mentioning that further valuable data induced by Linguistica concerns prefixes, information that is usually not provided in lots of morphosyntactic lexicons.

With Linguistica we generated a list of prefixes for each language with very low occurrences for all languages except German and Dutch, indicating that some prefixes exist, but that they are not as productive as in German and Dutch, where we obtained a list of more than 20 productive prefixes as shown in table 8.

We removed prefixes with only one character and occurring with less than five different stems. As we can see, most of the prefixes are all correct. In German only two are erroneous: **üb*, due to a bad segmentation of the prefix *über* and **nied* corresponding to a bad segmentation of the suffix *nieder*. In Dutch, there are two errors as well, *we and *oo. Prefixes in Dutch seem to be less productive than German, but still we get a total of 18 prefixes occurring with more than four different stems. So we can conclude that Linguistica can discriminate languages with less productive prefixation as Romance languages as well as English, and identify more productive prefixation for the Germanic languages German and Dutch.

GERMAN			**DUTCH**		
Prefix	**Occurrence with different stems**	**Corpus Count**	**Prefix**	**Occurrence with different stems**	**Corpus Count**
ge	40	252	uit	23	186
aus	30	226	af	20	58
ver	21	311	ge	17	97
hin	20	265	aan	17	55
auf	19	224	op	15	157
ab	19	218	toe	13	184
ein	16	243	be	12	61
her	13	261	ver	11	192
un	13	303	weg	9	208
weg	11	318	on	9	148
be	10	229	in	6	124
zu	10	326	na	5	141
*üb	9	3	weder	5	207
an	9	220	neder	4	144
er	8	247	over	4	163
*nied	7	283	*we	4	205
bei	6	230	ont	4	151
heim	6	259	door	4	83
über	5	4	samen	4	170
durch	5	241	*oo	4	154
ent	4	246			
zwei	4	327			
um	4	301			

Table 8: List of prefixes for German generated by Linguistica

5. Conclusions and Further Work

Inducing morphological information directly from corpora without previous resources seems an interesting approach for our tasks of comparing languages and for building lexicographic resources. The main advantage is that it is useful for inducing the size of paradigms in different languages, even if paradigms are incomplete. Suggested signatures could be used for helping the linguist to build lexicographic resources or an inflection engine. Main drawbacks are the fact that derivational suffixes cannot be separated from inflection ones and that no processing is provided for handling allomorphs.

Hence, further research will focus on how these results can be better exploited to build lexicons and use it as an aid for the linguist to build resources from scratch or from minimal knowledge. We expect to use a small dictionary with complete inflectional paradigms for some frequent words with regular inflections and project these complete paradigms on the signatures output by Linguistica in order to correct and complete them. Another possibility is to use use Linguistica's output to write inflection paradigms needed for the tool developed by Forsberg et al. (2006) described in section 2 to extract morphological lexica from raw text data. Furthermore, we could consider using the web to complete further paradigms, as already carried out by Oliver (2005) to find inflections not occurring in the corpus.

References

Atserias J., Casas B., Comelles E., González M., Padró L., Padró M., (2006). "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library". *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, ELRA. Genoa, Italy.

Baayen, R.H., R. Piepenbrock, and H. van Rijn. (1993). *The CELEX lexical database* (CD-ROM), LDC, Univ. of Pennsylvania, Philadelphia, PA

Blancafort, H. and Loupy, C. de (2009). Clues to Compare Languages for Morphosyntactic Analysis. In *Proceedings of LTC'09*, Poznán, Poland.

Brent, M. R., Murthy, S., and Lundberg, A. (1995). Discovering morphemic suffixes: A case study in minimum description length induction. In Fifth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale.Florida.

Clément L., Sagot B., Lang B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC'04*, Lisboa, Portugal. pp. 1841-1844.

Dal G., Namer F. (2000). GéDériF: automatic generation and analysis of morphologically constructed lexical resources. In *Proceedings of LREC 2000*, Athens, Greece, pp. 1447-1454.

Déjean, H. (1998). "Morphemes as necessary concepts for structures: Discovery from untagged corpora". *Workshop on paradigms and Grounding in NaturalLanguage Learning*, pp. 295-299.Adelaide, Australia.

Fosberg, M., Hammarström, H., and Ranta, A. (2006). Morphological Lexicon Extraction from Raw Text Data. In *Proceedings of the 5th International Conference on Advances in Natural Language Processing, FinTAL*, Finland.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:2 pp. 153-198.

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(3): 1-19.

Hathout N., Tanguy, L. (2005). Webaffix : une boîte à outils d'acquisition lexicale à partir du Web. In *Revue Québéquoise de Linguistique*. Volume 32, numéro 1.

Hathout N. 2005. Exploiter la structure analogique du lexique construit : une approche computationnelle. In *Cahiers de Lexicologie*. Volume 87, numéro 2.

Harris, Z. (1951). *Structural Linguistics*. University of Chicago Press.

Ide, N., Véronis, J., (1994). "MULTEXT: Multilingual Text Tools and Corpora". *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94*, Kyoto, Japan, pp. 588-92.

Loupy, C. de ; Bagur, M. ; Blancafort, H. (2009). Association automatique de lemmes et de paradigmes de flexion à un mot inconnu. In *Actes de TALN 2009*. Senlis, France.

Mikheev A., (1997). Automatic Rule Induction for Unknown-Word Guessing. In Computational Linguistics vol 23(3), ACL 1997. pp. 405-423.

Nakov P., Bonev Y., angelova G., Gius E., Hahn, W. von, (2003). Guessing Morphological Classes of Unknown German Nouns. In *Proceedings of Recent Advances in Natural Language Processing (RANLP'03)*. pp. 319-326. Borovetz, Bulgarie

Namer F. (1999). Le traitement automatique des mots dérivés : le cas des noms et adjectifs en -et(te). in D. Corbin, G. Dal, B. Fradin, B. Habert., F. Kerleroux, M. Plénat & M. Roché éds, La morphologie des dérivés évaluatifs, Silexicales 2, pp. 169-179. Villeneuve d'Ascq.

Namer F., (1999). Le traitement automatique des mots dérivés : le cas des noms et adjectifs en -et(te). in D. Corbin, G. Dal, B. Fradin, B. Habert., F. Kerleroux, M. Plénat & M. Roché éds, La morphologie des dérivés évaluatifs, Silexicales 2, pp. 169-179. Villeneuve d'Ascq.

Oliver, A. (2005). Adquisició d'informació lexica i morfosintáctica a partir de corpus sense anotar: aplicación al rus i al croat. *Procesamiento del lenguaje natural*, ISSN 1135-5948, Nº. 35, pp. 45-50.

Quirk, R.; Greenbaum, S.; Leech, G.; and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Resnik, P., Broman Olsen, M., Diab, M., (1999). The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues." *Computers and the Humanities* 33, 1–2, pp. 363–379.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, nº 14, pp. 445–471.

Sagot B., Clément L., Villemonte de la Clergerie E., Boullier P., (2006). "The Lefff 2 syntactic lexicon for French: architecture, acquisition, use". *Proceedings of the Language Resources and Evaluation Conference, LREC'06*.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. September 1994.

Schone, P., Jurafsky, D. (2001). Knowledge-Free Induction of Inflectional Morphologies. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*.

Zanchetta, E. and Baroni, M. (2006). Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings of Corpus Linguistics 2005*.

Zweigenbaum, P., Hadouche, F., and Grabar, N. (2003). Apprentissage des relations morphologiques en corpus. In *Actes de TALN*. 2003, pp 285–294. Batz-sur-mer, France.