Enriching Morphological Lexica through Unsupervised Derivational Rule Acquisition

Géraldine Walther[1], **Lionel Nicolas**[2]

1. Univ. Paris Diderot, Sorbonne Paris Cité & CNRS, LLF, UMR 7110 & INRIA, Alpage, UMR-I 001
175 rue du Chevaleret, 75013 Paris, France
2. Equipe RL, Lab. I3S, Université Nice Sophia-Antipolis & CNRS
2000 route des Lucioles, BP 121, 06903 Sophia Antipolis, France
geraldine.walther@linguist.jussieu.fr, lnicolas@i3s.unice.fr

Abstract

In a morphological lexicon, each entry combines a lemma with a specific inflection class, often defined by a set of inflection rules. Therefore, such lexica usually give a satisfying account of inflectional operations. Derivational information, however, is usually badly covered. In this paper we introduce a novel approach for enriching morphological lexica with derivational links between entries and with new entries derived from existing ones and attested in large-scale corpora, without relying on prior knowledge of possible derivational processes. To achieve this goal, we adapt the unsupervised morphological rule acquisition tool MorphAcq (Nicolas et al., 2010) in a way allowing it to take into account an existing morphological lexicon developed in the Alexina framework (Sagot, 2010), such as the Le*fff* for French and the Le*ff*e for Spanish. We apply this tool on large corpora, thus uncovering morphological rules that model derivational operations in these two lexica. We use these rules for generating derivation links between existing entries, as well as for deriving new entries from existing ones and adding those which are best attested in a large corpus. In addition to lexicon development and NLP applications that benefit from rich lexical data, such derivational information will be particularly valuable to linguists who rely on vast amounts of data to describe and analyse these specific morphological phenomena.

1 Introduction

Among existing lexical resources, morphological resources accounting for an language's inflectional properties are very common. Resources specifying derivation phenomena and derivation links between individual lexical entries, however, appear to be less complete — even for major languages such as French and Spanish. This is not a surprising fact, since, if we look at descriptive grammars, we also notice that the potentially missing parts of a language's morphological description usually concerns derivation, while inflection is thoroughly documented.

In this paper, we use an unsupervised morphological rule acquisition tool to uncover derivation rules for French and Spanish and acquire new lexical information, namely derivation links between existing lexical entries as well as new derived lexical entries, that is missing in two of the major lexical resources existing for these two languages: the Le*fff* (Sagot, 2010), a large-scale morphosyntactic lexicon for French, and the Le*ff*e (Molinero et al., 2009), a large-scale morphological lexicon for Spanish. In order to uncover these derivation rules missing in these two lexica, we adapt the unsupervised morphological rule learning technique MorphAcq (Nicolas et al., 2010) enabling it to take into account lexical data and complete the set of derivation rules in the Le*fff* and the Le*ff*e.

In the following sections, we will first sketch an overview of existing (semi-) automatic morphological rule acquisition techniques and lexical data acquisition techniques (section 2). In section 3, we describe the lexical framework Alexina (Sagot, 2010) and the Alexina lexica we used in our experiments. Then, in section 4, we describe morphological rule acquisition using MorphAcq, the acquisition tool itself, its adaptation to account for lexical data, the input corpora and the obtained raw results. In section 5, we show that using morphological rule acquisition techniques helps enriching existing lexical resources. We finally conclude in section 6.

2 Related Work

Unsupervised methods for morphological rule acquisition can be divided into roughly two types: those that aim at building morphological analysers through the optimisation of a specific set of metrics, and those that concentrate on the explicit uncovering of morphological information.

Among the first type, the most cited are *Linguistica* (Goldsmith, 2001; Goldsmith, 2006) and *Morfessor* (Creutz and Lagus, 2005). *Linguistica* constitutes the first real attempt to use the concept of *MDL* (*Minimum Description Length*) for encoding a complete corpus w.r.t. morphemes using as few bits as possible, thus trying to achieve the best possible affix and stem recognition. In (Creutz and Lagus, 2005), the authors also use the MDL approach without restricting the analysis of a word into only one facultative prefix, only one stem and only one suffix as is the case in (Goldsmith, 2001). Morfessor has later been extended for treating allomorphisms (Kohonen et al., 2009). Later, in (Golenia et al., 2009), MDL is used to pre-select possible stems for given forms; the stems are separated from the rest and the remaining strings considered possible affixes. These possible affixes are then first broken into substrings and then re-assembled according to a metric relying on the number of these substrings' occurrences. Spiegler *et al.* (2010), Bernhard (2008) and Keshava (2006) describe methods inspired by the work of Harris (1955) and extensions thereof (Hafer and Weiss, 1974; Déjean, 1998). These approaches focus on *transition probabilities* and *letter successor variety*, i.e., the distribution of letters following a given sequence of characters. They detect morpheme boundaries using entropy measures. The method described in (Demberg, 2007) also follows the algorithm in (Keshava, 2006), but corrects important drawbacks, in particular by handling with the fact that, for languages such as English, numerous forms are characterised by the absence of any kind of suffix. Dasgupta and Ng (2007) further extend the (Keshava, 2006) methods to the treatment of multiple suffixes.

The second type of unsupervised morphological rule acquistion methods concerns ways to identify morphological information *per se*. Thus, Lavallée and Langlais (2010) succeed in identifying word-formation using analogical processes such as *live vs. lively* and *cordial vs. cordially*. In this approach, every analogical process is weighted according to its productivity, i.e. the number of attested forms w.r.t. to the potential applicability of the analogical process. A similar approach is described in (Lignos et al., 2009). In this latter approach, however, productivity is measured according to the number of shared stems and the length of the attached affixes for each given form pair. In (Bernhard, 2010), the similarity of two forms is measured either by an edit distance or, when it is too small, by automatically extracted morphological and analogical rules. This similarity measure is then used in a *clustering* algorithm used to group possible forms for a given lemma. In (Can and Manandhar, 2009), the authors start with grouping forms according to similarity and then try to identify analogical processes between the forms of distinct groups. The productivity of the analogical processes is measured according to the number of shared stems. Finally, in (Monson et al., 2008) the morphological affixation rules applying to a given position class (in the sense of (Stump, 2001)) are directly identified, without prior identification of concrete possible affixes. This task uses a series of heuristics that control the output of the morphological rule detection method.

Concerning the acquisition of lexical data, several algorithms have been designed to extract new lemmas from a limited amount of information. They have been applied to several languages such as Russian (Oliver et al., 2003), French verbs (Clément et al., 2004), German nouns (Perera and Witte, 2005), Slovak (Sagot, 2005), Italian (Zanchetta and Baroni, 2005), French verbs, nouns and adjectives (Forsberg et al., 2006) and Polish (Sagot, 2007). These techniques differ from one another in various aspects, such as the soundness of the underlying probabilistic model and/or heuristics, the completeness of the manually described linguistic information that are exploited (e.g.,constraints on possible stems for each inflectional class, derivation patterns, etc.), the use of Google for checking the “existence” of a form, or the use of (probabilized since uncertain) part-of-speech information when it becomes available.

The acquisition of derivational links and derived lexical entries has also been studied. Systems like GéDériF (Dal and Namer, 2000) and its successors Walim (Namer, 2003) and Webaffix (Hathout, 2002) are for instance able to acquire new derived lemmas whenever their base lemma and their derivation rules are known.

In this work, we focus on the acquisition of new derived lemmas and derivation links in cases where the derivation rules have yet to be found. We propose an approach using the uncovering of these derivation rules through unsupervised morphological rule acquisition.

3 Presentation of the Input Lexica

3.1 The Alexina Framework

In our experiments, we used our morphological rule learning tool MorphAcq (described in section 4.1) jointly with lexical resources developed within the Alexina framework (Sagot, 2010). The lexica developed within the Alexina framework have the advantage of being all freely available[1] for quite a reasonable number of morphologically relatively diverse languages.

In this section, we thus briefly describe the Alexina framework underlying the two lexica we conducted our experiments on.

Although the Alexina framework covers both the morphological and the syntactic level, we only exploit the morphological level of the developed resources. Alexina allows for representing lexical information in a complete, efficient and readable way, that is meant to be independent of the language and of any grammatical formalism. It is compatible with the LMF standard[2] (Francopoulo et al., 2006). Numerous resources are being developed within this framework, such as the Le*fff*, a large-coverage morphological and syntactic lexicon for French (Sagot, 2010), the Le*ff*e for Spanish (Molinero et al., 2009), and also the Le*ff*ga for Galician, PolLex for Polish (Sagot, 2007), SkLex for Slovak (Sagot, 2005), PerLex for Persian (Sagot and Walther, 2010), SoraLex for Sorani Kurdish (Walther and Sagot, 2010) and KurLex for Kurmanji Kurdish (Walther et al., 2010).

The Alexina model is based on a two-level representation that separates the description of a lexicon from its use:

- The intensional lexicon factorises the lexical information by associating each lemma with a morphological class (defined in a formalised morphological description) and deep syntactic information; it is used for lexical resource development;
- The extensional lexicon, which is generated automatically by *compiling* the intensional lexicon, associates each inflected form with a detailed structure that represents all its morphological and syntactic information; it is directly used by NLP tools such as parsers.

3.2 The Input Lexica

The Le*fff* is the first lexical resource developed within the Alexina formalism (Clément et al., 2004) and has been continuously manually and automatically completed since then. The Le*ff*e (2009) is more recent. Still, the Le*ff*e contains a complete morphological description.[3]

In Alexina lexica, morphological information is encoded in a separate morphological description file that encodes the

[1] Under LGPL-LR licences, downloadable at the following address: `http://alexina.gforge.inria.fr`.

[2] Lexical Markup Framework, the ISO/TC37 standard for NLP lexica.

[3] The difference in scale between the Le*fff* and the Le*ff*e mainly lies in the syntactic level.

operations necessary to create the different forms for each given lemma according to a specific inflection table it belongs to. An example of inflection rules (`<form .../>`) and derivation rules (`<derivation .../>`) is given below: the inflection rule adds the suffix *es* to the stem of verbs in *–er*, indicated by the name of the table the rule belongs to. It thus creates an inflected form with the morphological tag `PS2s` (present indicative or subjunctive, second person singular). The derivation rule indicates that a derived lemma can be created from a nominal base in *-ion* by adding *ner* to the base stem.[4]

```
<table name="v-er" canonical_tag="W"
 rads="...*">
<form suffix="es" tag="PS2s"/>

<derivation suffix="ner" table="v-er"/>
```

In Alexina lexica, the relevant inflection class is specified for each lemma in the column immediately following the citation form. The lemmas are listed in a POS specific file containing the intensional lexical entries. Lemmas and inflection tables from the Le*fff*'s verbal entries are represented as below[5].

```
agacer v-er:std
agir   v-ir2
```

Adding new derivation rules requires encoding the rule in the Alexina language. Adding new derived lemmas hence entails indicating their newly associated inflection table.

4 Morphological Rule Acquisition from Raw Corpora

4.1 The MorphAcq System

MorphAcq (Nicolas et al., 2010) is a tool that takes as an input raw corpus data in a given language, that is supposed concatenative,[6] and automatically computes a data-representative description of the language's morphology.

Eventhough MorphAcq is still in a preliminary state of development, it has already proven its ability to compete with the state of the art, in particular by its first participation to the MorphoChallenge (Kurimo et al., 2009) competition.

MorphAcq can be thought of as a set of filters that sequentially refines a list of (candidate) affixes and a list of sets of related affixes, which are meant to belong to the same inflectional or derivational paradigm: such sets are called *morphological families*. The combination of an affix from a morphological family and a stem associated with this morphological family is expressed as a *morphological rule*. For MorphAcq, a morphological rule, be it derivational or inflectional, consists in adding one (possibly empty) affix (prefix or suffix) to a given stem with no character deletion or substitution whithin the stem or derivational base. Linguistic phenomena that might modify the stem and/or the affix thus lead to various different morphological rules.[7]

The overall MorphAcq algorithm can be decomposed into five steps:

1. Generate an over-covering and "naive" list of candidate affixes, i.e., substrings that may be affixes. In other words, each form found in the corpus is split into a large number of stem+affix combinations (among which most are incorrect).

2. Detect candidate affix pairs that seem to be related (see discussion of step 2 below for details). For example, if affixes a, b and c belong to the same morphological family (e.g., to the same inflection class), then this step should detect pairs $\{a, b\}$, $\{b, c\}$ and $\{a, c\}$.

3. Build morphological families according to sets of pairs that share a common stem. For instance, if affixes a, b and c have all been seen on the same stem, and if the pairs $\{a, b\}$, $\{b, c\}$ and $\{a, c\}$ have been detected as "related" in step 2, the morphological family $\{a, b, c\}$ is built.

4. Split compound affixes. For example, split the English suffixes *-ingly* into *-ing* and *-ly*.

5. Detect which substrings can connect stems and split compound stems. For instance, detect that the hyphen ("-") can connect English stems and split the form "brother-in-law" into "brother + in + law".

All these steps are based mostly on simple computations with no or few free parameters. Therefore, MorphAcq can be used on virtually any concatenative language with almost no expert work.

We focus here on steps 2 and 3, which needed adaptation for this work in order to take into account external lexical data. The first step was left unchanged, and the fourth and fifth steps provide data that is not relevant here.

Step 2 exploits the following crucial observation about form- vs. lemma frequency: the frequency of a lemma's inflected forms tends to vary consistently with the the lemma's overall frequency. For example, in general texts, all inflected forms of the lemma *to talk* are more frequent than their corresponding forms from the lemma *to orate*. Moreover, this observation is not limited to the inflected forms of a lemma, but applies also derived lemmas and forms. For example, let us consider a set of forms found in the input corpus and that can be split into a stem and one of the two affixes a_1 or a_2. The goal of step 2 is to decide whether a_1 and a_2 belong to the same morphological family, i.e., whether they belong to the same inflectional or

[4] Examples are from the Le*fff*'s morphological description.

[5] Syntactic information, including detailed valency information, is included in the Le*fff*, but is not shown here out of clarity reasons, as it is not relevant in this paper.

[6] We define here a concatenative language as a language that uses morphological operations that can all be entirely described through affixation. The rules are applied to graphemic sequences. Sandhi phenomena are treated independently, e.g., through the operation `<fusion .../>` in an Alexina lexicon.

[7] For instance, in French, *chantons* and *mangeons*, inflected forms corresponding to stems *chant-* and *mang-* correspond to two different morphological rules, one that adds the suffix *-ons* and another one involving the suffix *-eons*. The fact that the "real" suffix is *-ons* in both cases and that the extra *-e-* is the consequence of a phonographemic rule is not extracted.

derivational paradigm. If this is the case, which means that the frequency of lemmas and the frequency of their forms are found to vary accordingly, sorting stems s according to the frequency of the forms $s + a_1$ or according to the frequency of the forms $s + a_2$ should lead to similar orderings. Oppositely, if a_1 and a_2 are not related, both orderings should be very different.

Once pairs of related affixes are identified, step 3 builds sets of affixes that constitute morphological families by putting together pairs that have been seen on at least one common stem. It then uses four different heuristic filters for removing incorrect affix sets. Among these filters, the main one relies on the same observation as step 2. Indeed, this form- and lemma-level frequency consistency implies that the more frequent a lemma is, the more of its inflected forms will occur in the corpus. Therefore, less frequent lemmas should be attested in the corpus only by some of the inflected forms generated by their inflection class, whereas more frequent lemmas from the same inflection class are attested by more distinct forms. This means that we should be able to relate a morphological family involving n affixes with morphological families involving only $n -$ 1, $n -$ 2,…, 1 of these affixes, and that these families should be associated with stems of decreasing frequency. Therefore, we use a filter that keeps a morphological family with n affixes only if it at least one of its morphological subfamilies involving $n - 1$ of its affixes is identified as such.

4.2 Adapting MorphAcq

In order for MorphAcq to take into account lexical data, we modified steps 2 and 3 as follows.

First, step 2 uses the lexicon for grouping inflected forms of a same lemma, considered as a combination stem+inflection class. Instead of applying the frequency-based observation described above on two stem+affix sequence pairs, which allows to compare the two corresponding affixes, we now apply this observation on a stem+inflection class sequence and a stem+affix sequence such that the form stem+affix is not generated by the inflection class. Thus, we are able to identify affixes that are "related" to inflection classes, by means of stems they are both associated with (by the lexicon as far as the inflection class is concerned, and by the corpus as far as the affix is concerned).

For each inflection class c_b, step 3 then tries to group into affix sets the "related" affixes found during step 2. These "related" affixes generate forms that do not belong to the known inflectional paradigm of the (base) lemma l_b corresponding to their stems. They might therefore correspond to missing inflectional rules or to missing derivational rules.

We first suppose that all these rules are derivational, i.e., these forms are candidates for being inflected forms of lemmas l_d (with inflection class c_d) that are derived from that base lemma l_b. If, for at least one stem s, one of these candidate derived forms is known to the lexicon, then the lexicon provides us with its lemma l_d and inflection class c_d. This allows for computing a morphological (derivation) rule that transforms l_b into l_d. By removing the longest substring l_b and l_d have in common, we can turn this morphological rule into a generic rule that might apply to any lemma with inflection class c_b.

	LANGUAGE	CORPUS SIZE (IN TOKENS)
Le*fff*	*French*	*∼18 215 000*
Le*ffe*	*Spanish*	*∼540 000*

Table 1: Corpora used as an input to MorphAcq

If this process fails on a given affix, this affix is considered inflectional: we then build the corresponding missing inflection rule. The fact that it is missing explains why the form is unknown to the lexicon although its lemma l_b is known.

Finally, MorphAcq is able to associate a confidence score with each morphological rule it outputs, based on paradigm coverage and form frequency.

4.3 The Input Corpora

As input data to MorphAcq, we used a corpus extracted from the French newspaper *le Monde diplomatique* [8] for French, and the raw data of the *Ancora* corpus (Taulé et al., 2008) for Spanish. We were able to detect several missing derivational rules for both our input lexica. The corresponding figures are given in Table 1.

4.4 Results and Evaluation of the Morphological Rule Acquisition

When we first confronted the output of MorphAcq with the forms generated with the two Alexina lexica, the results showed that both resources seem to reasonably well encode the inflectional system of both languages. The inflectional rules that were suggested as missing rules were the result of isolated typographical errors or English loanwords. Therefore, we simply ignored the few inflection rules that were suggested by MorphAcq.

MorphAcq generated 3,131 derivational rules from our Spanish data, and 36,430 derivational rules from our French data. This huge difference is mostly due to the fact that the French corpus we gave as an input to MorphAcq is over 30 times bigger than the Spanish one. However, many of these rules have to be considered as noise. This is why we applied various filters before using them in practical lexicon enrichment experiments, as explained in the next section.

5 Enriching Lexical Resources through Automatic Acquisition of Morphological Rules

5.1 Evaluation of Acquired Derivation Rules through External Information

Derivation is a morphological process that generates a new lemma from the derivation-base of a first one. The new lemmas are part of the set of lexical entries available in the lexicon of a given language. They have to be associated with the right inflection tables since they are themselves possibly inflectable. Recall that in Alexina

[8] `http://www.monde-diplomatique.fr/`, February 2011.

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
basculement	*nc-2m*	basculer	v-er:std
centreur	*nc-2m*	centrer	v-er:std
crochetage	*nc-2m*	crocheter	v-er:std
déloyalement	*adv*	déloyal	adj-al4
fasciste	*nc-2*	fasciser	v-er:std
gourmand	*nc-2f*	gourmander	v-er:std
insolation	*nc-2f*	insoler	v-er:std
minimaliser	*v-er:std*	minimal	adj-al4
perfectionnement	*nc-2m*	perfectionner	v-er:std
reboisement	*nc-2m*	reboiser	v-er:std
soûler	*v-er:std*	soûl	adj-4
trébuchement	*nc-2m*	trébucher	v-er:std

Table 2: Examples of French derivation links acquired automatically

lexica, the inflection class is specified for each lemma in the column immediately following the citation form (the above example is simplified, since the syntactic — e.g., valency — information is not shown).

```
agacer v-er:std
agir   v-ir2
```

Before adding derivation rules to the morphological descriptions underlying the Le*fff* and the Le*ff*e, we first filtered out from the derivation rules output by MorphAcq those that seemed less likely, in the following way. First, we automatically filtered the output given by MorphAcq using a beam filter: for a given morphological family (including the associated base inflection class), many derivation rules may be suggested by MorphAcq, each affix in the morphological family being covered by more than one of these derivation rules (each derivation rule, in turn, usually covers more than one affix, as it creates a derived lemma that has several inflected forms). For each affix in the considered morphological family, we identify the suggested morphological rule that has the best score among those that cover that affix: it is the affix's best rule. Then, we only keep those morphological rules that are the best rules for at least one of its affixes.

Among the remaining derivation rules, we require that suffixation rules be suggested for at least two distinct morphological families and prefixation rules by 25 morphological families for French and five for Spanish.[9]

Then we automatically added all remaining derivation rules into the Le*fff*'s or the Le*ff*e's morphological description.

We were also able to retrieve the possible *variant* a new lemma belongs to: variants are used in Alexina to differentiate lemmas that show particular morphotactic properties with minor impact on the lemmas inflection.[10] Hence, derivation rules are represented as follows:

```
<derivation suffix="ner" table="v-er"
 var="std" />
```

Converting and filtering MorphAcq's output led to the introduction of 823 derivation rules into the French morphological description. These new rules are scattered over most existing inflection classes. For Spanish, only 132 derivation rules could be identified and added. This difference in scale has again to be imputed to the difference in size the the corpora used as input to MorphAcq.

Once the new derivation rules added into the lexica, we generated all possible derived lexical entries by applying to each existing entry all derivation rules associated with its inflection class. We obtained as large a result as 2.9 million candidate entries for French and 1.0 million candidate entries for Spanish. However, Alexina inflection tables are often associated with constraints on stems: e.g., French adjectives inflecting according to class `adj-n4` in the Le*fff*, such as *parisien(s) / parisienne(s)*, are requested to have a stem ending in *n*. Trying to inflect the new derived lemmas hence allowed us to discard all those new lemmas whose stem was not compatible with the inflection class suggested by MorphAcq.

The remaining derived lemmas were used in two different ways. First, derived lemmas that correspond to existing entries in the Le*fff* or the Le*ff*e were preliminarily validated as correct derived lemmas, i.e., we considered that derivation links between base and derived lemmas could be added. The entries corresponding to derived lemmas thus received a derivation link of the form *derived from X*.[11] At this point 16,646 derivational link candidates were added for French and 10,745 for Spanish.

Among the candidates, the derived lemmas are necessarily correct as lexical entries, since they were found within the lexica. Only the correctness of the derivation links with the base lemma needs to be assessed. To do so, we performed manual evaluation on randomly selected samples containing 100 candidates. For Spanish, all 100 morphosemantic links were correct (see Table 3). For French, we obtained 92 correct links out of 100 (see Table 2), but from a larger set of candidates (errors are shown in Table 4). It also became apparent that the longer the base and/or the derived lemma is, the greater the certainty of the established link's correctess. Indeed, Table 4 shows that most errors involve at least one relatively short lemma.

5.2 Using Newly Acquired Rules for Enriching Large Scale Resources

Once the derrivation links between the lemmas already contained within the Le*fff* and the Le*ff*e had been identified, we developed a procedure for adding new (unknown) derived lemmas (with their corresponding derivational links that initially led to suggesting them). For selecting which derived lemmas had to be added, we used form frequency information extracted from large-scale corpora.

[9] The apparent striking difference in the selectivity imposed on prefixation rules between French and Spanish comes from the different scales of the acquisition corpora fed into MorphAcq. Using the same threshold for both languages would have led to either to much noise in the French data or to few acquirable rules for Spanish.

[10] See for instance French verbs that double their stems last consonants when preceding certain suffixes: infinitive *je***t***er* "throw" vs. P1sg of the present indicative *je je***tt***e* "I throw".

[11] This tag is meant to facilitate future use of the Le*fff* as a resource for studies on derivational relations.

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
calcular	*V2*	calculadamente	R1
conspirador	*N8*	conspirar	V2
desencadenante	*N1*	desencadenar	V2
extremo	*N1*	extremar	V2
horadable	*A2*	horadar	V2
justo	*N4*	justar	V2
modoso	*A1*	modosamente	R1
patrimonialista	*A2*	patrimonial	A3
racional	*A3*	ración	N3
rotulista	*N6*	rotular	A3
temperado	*A1*	temperadamente	R1
zanqueador	*N8*	zanquear	V2

Table 3: Examples of correct Spanish derivation links acquired automatically

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
attiser	*v-er:std*	attiquement	adv
bafouiller	*v-er:std*	bafouer	v-er:std
cotte	*nc-2*	coter	v-er:std
entassement	*nc-2m*	enter	v-er:std
must	*nc-2m*	muser	v-er:std
présentement	*adv*	présenter	v-er:std
salement	*adv*	saler	v-er:std
sire	*nc-2m*	sirex	nc-1m

Table 4: Examples of incorrect French derivation links. Most links involve at least a "short" lemma

For French, we used a part of the *Est Républicain* corpus[12], composed of newspaper articles published in 1999. We tokenized the corpus of the *Est Républicain* into 37.5 million tokens using the "light" version of the shallow processing chain SxPipe which is included in the distribution of the MElt POS-tagger (Denis and Sagot, 2009). For Spanish, we used a cleansed dump of the Spanish Wikipedia[13]. The Spanish Wikipedia was tokenized with the same "light" version of SxPipe. We retained the first 10 million tokens.

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
maltraitance	*nc-2f*	maltraiter	v-er:std
recapitalisation	*nc-2f*	recapitaliser	v-er:std
incinérable	*adj-2*	incinérer	v-er:std
rissolette	*nc-2f*	rissoler	v-er:std
abreuvement	*nc-2m*	abreuver	v-er:std
rétractable	*adj-2*	rétracter	v-er:std
plastification	*nc-2f*	plastifier	v-er:std
tronçonnement	*nc-2m*	tronçonner	v-er:std
grenailleur	*nc-2m*	grenailler	v-er:std
désencadrement	*nc-2m*	désencadrer	v-er:std
regardable	*adj-2*	regarder	v-er:std
grêleux	*nc-x3*	grêler	v-er:std

Table 5: Examples of new French derived lemmas acquired automatically

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
orbitador	*N5*	orbitar	V2
presentacional	*A3*	presentación	N3
correlacional	*A3*	correlación	N3
insercional	*A3*	inserción	N3
confrontante	*N1*	confrontar	V2
agudismo	*N1*	agudizar	V3
multidireccionalidad	*N7*	multidireccional	A3
distintal	*N5*	distinto	A1
letalidad	*N7*	letal	A3
aleteador	*N5*	aletear	V2
aconfesionalidad	*N7*	aconfesional	A3
zanqueador	*N8*	zanquear	V2

Table 6: Examples of new Spanish derived lemmas acquired automatically

We used these corpora as follows. First, we filtered out candidate derived lemmas whose canonical form is not attested in the corpus. This first filtering reduced the number of derivation candidates from respectively 2.9 million and 1.0 million derived lemma candidates to 62,158 for French and 22,814 for Spanish. Then, we inflected all these candidates, generating 191,000 possible new inflected entries for French and 94,000 for Spanish. We associated those with two basic sources of information: whether each inflected form is known to the lexicon or not, and its number of occurrences, if any, in the corpus.

Then, we ranked the remaining candidates in the following iterative way: at each step, we computed a score for each derived lemma candidate by adding contributions for every one of its inflected forms; these contributions were computed as their number of occurrences, taken positively if the form is unknown to the lexicon and negatively if it is known to the lexicon. The idea of this ranking is to suggest only those new lemmas that have the best coverage of corpus forms still unknown to the lexicon and do not at the same time cover forms already known to the lexicon. After having ranked all candidates, we output the best one. All its inflected forms were now considered as known to the lexicon. This means we needed to re-compute the scores and iterate the process[14]. Each iteration outputs one candidate. We stopped when the best candidate had a score smaller or equal to 1. As a result, we obtained 1,511 new derived lemmas for French and 563 new derived lemmas for Spanish. We added these new lemmas to the Le*fff* and the Le*ff*e respectively, specifying the corresponding derivation tag for each {base lemma, derived lemma} pair.

After adding the new lemmas we performed a small manual evaluation on 100 randomly chosen new lemmas and their derivation tags for both languages. Examples of correct new derived lemmas are shown in Tables 5 and 6, whereas the quantitative results of this evaluation are given in Table 7.

[12] http://www.cnrtl.fr/corpus/estrepublicain/

[13] http://download.wikimedia.org/eswiki/latest/eswiki-latest-pages-articles.xml.bz2, dump from Feb 3, 2011.

[14] Of course, we did not recompute all scores, but only updated those which had been affected by the last output.

FRENCH	• 42 correct lemmas & derivation links, • 1 correct lemma with false derivation link, • 14 correct canonical forms with incorrect inflection tables, • 10 incorrect lemmas due to the presence of English words in the corpus, • 28 incorrect lemmas due to typographical errors in the corpus, • 5 other incorrect candidates.
SPANISH	• 40 correct lemmas & derivation links, • 7 correct canonical forms with incorrect inflection tables, • 39 incorrect lemmas due to the presence of English words in the corpus, • 9 incorrect lemmas due to typographical errors in the corpus, • 5 other incorrect candidates.

Table 7: Derived Lemma Evaluation

6 Conclusion and Future Work

In this paper, we have presented a novel method for enriching large-scale lexica with concrete derivation links and a straightforward manner to use the acquired explicit derivational information to increase a lexicon's coverage. The new derivation rules have been acquired through a specifically adaptated version of the unsupervised morphological rule acquisition tool MorphAcq.

An obvious interesting side result of this method is that the lexica on which our method has been applied now show an improved quality: derivation links have been specified within the Le*fff* and Le*ffe*, hence allowing to use both resources for theoretical and descriptive linguistic studies on derivation.[15]

A further step in enriching lexical resources (in general, and Alexina lexica in particular) should be to combine the morphological rule acquisition tool MorphAcq with other methods designed for identifying new possible lemmas, as described in (Sagot, 2005). We plan on running the tools developed by Sagot (2005) jointly with MorphAcq. These lemma acquisition methods that rely on information from the morphological description should benefit from the improved description provided by MorphAcq's output. MorphAcq will in return benefit from being combined with resources with greater coverage. In particular the identification of the correct inflection classes for new derived lemmas should be significantly improved. Thus, using morphological rule acquisition and lemma acquisition techniques iteratively seems a promising way for efficient lexical resource enriching. This method should help rapidly developing new lexica with completely automatic methods, hence giving access to new resources for undescribed languages.

Acknowledgements

This work was supported in part by the EDyLex project funded by the French National Research Agency (grant number ANR-09-CORD-008).

7 References

Delphine Bernhard. 2008. Simple morpheme labelling in unsupervised morpheme analysis. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the CLEF (revised selected papers)*, pages 873–880. Springer-Verlag, Berlin, Heidelberg.

Delphine Bernhard. 2010. MorphoNet: Exploring the use of community structure for unsupervised morpheme analysis. In *Multilingual Information Access Evaluation, 10th Workshop of the CLEF (revised selected papers)*, Corfu, Greece. Springer.

Burcu Can and Suresh Manandhar. 2009. Clustering morphological paradigms using syntactic categories. In *CLEF*, pages 641–648.

Lionel Clément, Benoît Sagot, and Bernard Lang. 2004. Morphology Based Automatic Acquisition of Large-coverage Lexica. In *Proceedings of LREC'04*, pages 1841–1844, Lisbon, Portugal.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. In *Helsinki University of Technology*.

Georgette Dal and Fiammetta Namer. 2000. Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations. *TAL*, 41-2:423–446.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *NAACL HLT 2007: Proceedings of the Main Conference*, pages 155–163.

Hervé Déjean. 1998. Morphemes as necessary concept for structures discovery from untagged corpora. In *NeMLaP3/CoNLL '98: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 295–298, Sydney, Australia.

Vera Demberg. 2007. A language-independent unsupervised model for morphological segmentation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 920–927, Prague, Czech Republic, June.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong-Kong, China.

Markus Forsberg, Harald Hammarström, and Aarne Ranta. 2006. Morphological lexicon extraction from raw text data. In *Proceedings of FinTAL 2006, LNAI 4139*, pages 488–499, Turku, Finland. Springer-Verlag.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Genoa, Italy.

John Goldsmith. 2001. Unsupervised learning of the

[15] The results are freely available on `http://www.linguist.univ-paris-diderot.fr/~gwalther/homepage/Publications_(en).html`.

morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

John Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Nat. Lang. Eng.*, 12(4):353–371.

Bruno Golenia, Sebastian Spiegler, and Peter Flach. 2009. Ungrade: Unsupervised graph decomposition. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece*, September.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10(11-12):371–385.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Ludovic Hathout, Nabil et Tanguy. 2002. Webaffix: finding and validating morphological links on the WWW. In *Proceedings of LREC'02*, pages 1799–1804, Las Palmas de Gran Canaria, Spain.

Samarth Keshava. 2006. A simpler, intuitive approach to morpheme induction. In *In PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, pages 31–35.

Oskar Kohonen, Sami Virpioja, and Mikaela Klami. 2009. Allomorfessor: towards unsupervised morpheme analysis. In *Evaluating systems for multilingual and multimodal information access, 9th Workshop of the CLEF*, pages 975–982, Berlin, Heidelberg. Springer-Verlag.

Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of morpho challenge 2009. In *Multilingual Information Access Evaluation, 10th Workshop of the CLEF (revised selected papers)*, CLEF'09, pages 578–597, Berlin, Heidelberg. Springer-Verlag.

Jean-Francois Lavallée and Philippe Langlais. 2010. Unsupervised morphology acquisition by formal analogy. In *Lecture Notes in Computer Science*.

Constantine Lignos, Erwin Chan, Mitchell P. Marcus, and Charles Yang. 2009. A rule-based acquisition model adapted for morphological analysis. In *Evaluating systems for multilingual and multimodal information access, 9th Workshop of the CLEF*, pages 658–665.

Miguel Ángel Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The Le*ffe*. In *Proceedings RANLP 2009*, Borovets, Bulgaria.

Christian Monson, Alon Lavie, Jaime Carbonell, and Lori Levin. 2008. Evaluating an agglutinative segmentation model for paramor. In *SigMorPhon '08: Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 49–58, Morristown, NJ, USA.

Fiammetta Namer. 2003. WaliM : valider les unités morphologiquement complexes par le Web. In B. Fradin et al., editor, *Silexicales 3 : les unités morphologiques*, pages 142–150, Villeneuve d'Ascq, France. Presses Universitaires du Septentrion.

Lionel Nicolas, Jacque Farré, and Miguel A. Molinero. 2010. Unsupervised learning of concatenative morphology based on frequency-related form occurrence. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, Helsinki, Finland, September.

Antoni Oliver, Irene Castellón, and Lluís Màrquez. 2003. Use of Internet for augmenting coverage in a lexical acquisition system from raw corpora: application to Russian. In *Proceedings of the RANLP'03 International Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria.

Praharshana Perera and René Witte. 2005. A self-learning context-aware lemmatizer for German. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 636–643, Vancouver, Canada.

Benoît Sagot and Géraldine Walther. 2010. A Morphological Lexicon for the Persian Language. In *Proceedings of LREC'10)*, Valetta, Malta. ELDA.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658, Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic, September. Springer-Verlag.

Benoît Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of the 3rd Language & Technology Conference*, pages 423–427, Poznań, Poland, October.

Benoît Sagot. 2010. The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of LREC'10*, Valetta, Malta.

Sebastian Spiegler, Bruno Golenia, and Peter Flach. 2010. Unsupervised word decomposition with the promodes algorithm. In *Multilingual Information Access Evaluation, Lecture Notes in Computer Science*, volume I. Springer Verlag, February.

Gregory T. Stump. 2001. *Inflectional Morphology. A Theory of Paradigm Structure*. CUP, Cambridge, UK.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of LREC'08*, Marrakesh, Morroco.

Géraldine Walther and Benoît Sagot. 2010. Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop (LREC'10 Workshop)*, Valetta, Malta.

Géraldine Walther, Benoît Sagot, and Karën Fort. 2010. Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrad, Serbia.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK. University of Birmingham.