Expert Systems with Applications 40 (2013) 7122–7131

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue

Zenón Hernández-Figueroa, Francisco J. Carreras-Riudavets *, Gustavo Rodríguez-Rodríguez

Departamento de Informática y Sistemas, Edificio de Informática y Matemáticas, Campus Universitario de Tafira, Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas, Spain

ARTICLE INFO

Keywords:
Syllabification
Lemmatization
Derivation
Prefix

ABSTRACT

The syllabification of Spanish's words follows a few basic rules, but the syllabification of some words deviates from the general rules according to a number of factors described in this paper. Prefixes are major cause of variations on syllabification. Since, in Spanish, prefixes tend to do not integrate into other syllables when they are prominent, the syllabification of words can vary depending on the prominence of the prefixes. This paper shows that, in many cases, the prominence of a prefix can be inferred by means of some morphological and lexical knowledge. This paper proposes a syllabification algorithm that implements the basic syllabification rules and combines them with morphological and lexical information obtained from three sources: a lemmatizer, a derivation database, and the *Corpus de Referencia del Español Actual* (CREA) of Royal Spanish Academy. Using this additional information, this paper attempts to provide a solution to the problem of taken into account the prefixes according to its prominence for a correct syllabification.

1. Introduction

Syllabification is the process of splitting words into syllables (Real Academia Española, 2009). A syllable is defined as 'articulated sound, or sounds, that make up a single phonic core between two successive depressions of the voice output' (Real Academia Española, 2012). It is a single unit of speech, either a whole word or one of the parts into which a word can be separated (Cambridge Universtity Press, 2012). The division of the words in syllables is an essential requirement to be able to apply correctly the Spanish rules of graphic accentuation (Real Academia Española, 2010).

Availability of correct automatic syllabification tools could be of aid for several fields. Syllables have long been regarded as robust units of speech perception and recognition (Anusuya & Katti, 2009; Janakiraman, 2010), as far as "automatic segmentation and labeling of speech at the phonetic level is not very accurate while syllable boundaries are more precise and well defined" (Ganapathiraju, Hamaker, Picone, Ordowski, & Doddington, 2001). Text-to-speech systems using concatenative waveform synthesis are based on the maintenance of a waveform repository of basic speech units; syllables could be appropriate units for that purpose (Lopez-Gonzalo & Rodríguez-García, 1996; Rao, Samuel Thomas, & Murthy, 2005). Integration in language-learning tools, both for kids and for foreigners is another field where automatic syllabification could be useful.

The syllabification of Spanish words follows a few basic rules which are briefly discussed at Section 2, but the syllabification of some words deviates from the general rules according to a number of factors. By example, some sequences of vowels could be interpreted either as diphthongs or as hiatuses; such is the case of *fluir* ('to flow') that could be syllabified *flu.ir* (disyllabic) or *fluir* (monosyllabic). In a similar way, the same sequence of weak vowel, strong vowel, and weak vowel could be a triphthong, as in *cambiéis* (*cam.biéis*), or a hiatus + diphthong, as in *confiéis* (*con.fi.éis*), at least in Spain and some other geographical areas (Real Academia Española, 2005). Some of the variations respond to geographical-dialectal criteria, such is the case of the group 'tl', as in *atlántico* ('Atlantic'), which is considered non-separable in most of Hispano-America, the Canary Islands and some areas of the rest of Spain (*a.tlán.ti.co*) but it is splitted in Puerto Rico and some other areas of Spain (*at.lán.ti.co*).

Prefixes are another cause of variations on syllabication. Although, they tend to do not integrate into other syllables, they can vary the syllabification of words, depending on their prominence, as we can read on the *Nueva Gramática de la Lengua Española* (Real Academia Española, 2009).

When developing automatic tools for syllabification, geographical criteria for variations must be accepted as they are. Criteria to apply can be selected, but they cannot be inferred from the words themself. The prominence of prefixes is another question. This article shows that, in many cases, the prominence of a prefix can be inferred by means of some morphological and lexical knowledge. Cuayahuitl (2004) describes a syllabification algorithm for Spanish

* Corresponding author. Tel.: +34 928 45 87 29; fax: +34 928 45 87 11.
E-mail address: fcarreras@dis.ulpgc.es (F.J. Carreras-Riudavets).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.056

that takes into account prefixes, but the Cuayáhuitl's algorithm tries to always separate the prefixes, without any consideration about their prominence. Moreover, the Cuayáhuitl's algorithm fails because it does not use any morphological or lexical information, but a list of prefixes. So, given a word, the Cuayáhuitl's algorithm is capable of to identify if the heading string of the word matches a prefix, but it cannot knows if that string is really a prefix, or only a coincidence; for example, both *comida* ('lunch') and *coautor* ('co-author') begin with the string *co-*, but only in *coautor* it is a prefix.

There are some works that attempt to manage the complexity of syllabification of Spanish words using machine-learning systems to capture syllabification rules by means of training examples (Goddard & René MacKinney-Romero, 2006), obtaining similar accuracies. Adsett compares five data-driven syllabification algorithms on nine European languages, including Spanish (Adsett & Marchand, 2009). Data-driven algorithms are attractive because they are not language specific. They work just based on the symbol combination and cannot effectively taking into account exceptions due to external factors as prefixes prominence. Most of available syllabifiers only implement the basic syllabification rules, without taken into account any exception.

In this article we propose a syllabification algorithm that implements the basic syllabification rules and combines them with morphological and lexical information obtained from three sources: a lemmatizer, a derivation database, and the *Corpus de Referencia del Español Actual* (Real Academia Española, 2011). Using this additional information we attempt to provide a solution to the problem of taken into account the prefixes according to their prominence.

We present the prefix prominence issue at Section 3. Sections 4 and 5 describe the morphological and lexical knowledge and tools proposed to locate prefixes. Section 6 describes how to apply knowledge about prefixes and words use to determine prefix's prominence. Section 7 analyses the effects of the proposed algorithms when applied to the wordlist of the *Corpus de Referencia del Español Actual* and, finally, Section 8 presents our conclusions.

2. Basic syllabification rules for Spanish words

A syllable is composed of three parts (Grammar 1): a mandatory nucleus, and two optional margins, one preceding the nucleus (onset) and another following the nucleus (coda). The nucleus is the most sonorous, visible and open element of those which make up the syllable.

The nucleus of a syllable may be composed by a single vowel, a diphthong or a triphthong (Grammar 2). In Spanish, a diphthong is composed of two different weak vowels ('iu', 'ui' or 'uy' at end of word) or by a strong vowel and a weak vowel. Diphthongs can be decreasing, when the strong vowel is the first one ('ai', 'au', 'ei'/'ey', 'eu', 'oi'/'oy', 'ou'), or increasing, when the strong vowel is the last one ('ia', 'ie', 'io', 'ua', 'ue', 'uo'). When a diphthong is composed by two weak vowels, it is named a homogeneous diphthong.

```
syllable ::= [onset] nucleus [coda]
```

Grammar 1. Syllable.

```
nucleus      ::= vowel | diphthong | triphthong
vowel        ::= weak-vowel | strong-vowel
strong-vowel ::= 'a' | 'e' | 'o'
weak-vowel   ::= 'i' | 'u'
diphthong    ::=
  weak-vowel weak vowel |
  weak-vowel strong-vowel |
  strong-vowel weak-vowel
triphthong   ::= weak-vowel strong-vowel weak-vowel
```

Grammar 2. Nucleus.

A triphthong is composed by a strong vowel with a weak vowel at each side. Presence of the 'h' does not break a diphthong or triphthong because that character has not sound in Spanish.

The character 'y', at the end of a word or when followed by a consonant (Bezos, 2006), represents the vocalic sound /i/, so it can be part of a diphthong or of a triphthong.

The onset of a Spanish syllable, if any, may be composed by a single consonant or by two consonants (Grammar 3). In this case, the first consonant must be an occlusive 'p', 'b', 't', 'd', 'k', 'g' or fricative-labiodental consonant 'f', and the second one must be 'r' or 'l', taking into account that 'l' cannot be preceded by 'd' or, in the Spanish used at some geographical areas, by 't' (the group 'tl' is not native of Spanish, but it is present in many words from native american languages which have been incorporated to Spanish).

There are three pairs of consonants that represent a single sound: 'ch' (voiceless palato-alveolar affricate), 'll' (palatal lateral approximant /ʎ/), and 'rr' (alveolar trill /r/) inside a word. There are also a number of other pairs that can start a syllable 'pt', 'ct', 'cn', 'ps', 'mn', 'gn', 'ft', 'pn', 'cz', 'tz', 'ts',... (Bezos, 2006). They are typical of cult or foreign words, and they are not cover by the definition of onset presented above, but they can be considered as a single consonant; in fact, the first character tends, in vulgar speech, to be soundless.

The coda inside a word may be composed by any consonant, or by 'm' or 'n' followed by 's' (Grammar 4). The coda at the end of a word can only be composed by an alveolar consonant 'd', 'r', 'l', 'n', or 's'.

An important consideration to syllabification is that the onset is preeminent when splitting a word. When, inside a word, a piece could be part of the coda of the left-side syllable or part of the onset of the right-side syllable, producing valid syllables in both cases, it must be incorporate to the onset (Ualde, 1989).

3. The prefix's prominence issue

The syllabification of Spanish words varies by morphological issues like the presence of prefixes. Words with prefixes will be separated by their components (prefixal syllabification), or the prefix will be merged with the root of the word for syllabification (non-prefixal syllabification) depending on its prominence. By example, the syllabification of *sublunar* ('of under the moon') is *sub.lu.nar* while the syllabification of *sublime* ('eminent') is *su.bli.me* and not **sub.li.me*, because the prefix *sub-* ('under') is identified in the first case (added to the word *lunar*, meaning 'of the moon') and not in the second one, although, as we can read in the *New Grammar of the Spanish Language* (Real Academia Española, 2009), *sublime* contains the prefix *sub-* from the etymological point of view.

```
onset ::= consonant | osc1a osc2 | 'd' 'r'
osc1  ::= 'p' | 'b' | 't' | 'k' | 'g' | 'f'
osc2  ::= 'l' | 'r'
```

Grammar 3. Onset.

```
coda          ::= inside-coda | end-coda
inside-coda   ::= consonant   | ('m' | 'n') 's'
end-coda      ::= alveolar-consonant
alveolar-consonant ::= 'd' | 'r' | 'l' | 'n' | 's'
```

Grammar 4. Coda.

The end of a non-prominent prefix must be grouped into the same syllable with the start of a root in the following cases:

- When the prefix ends with a vowel and the root of the word begins with a weak vowel, it is possible to build a diphthong by joining the end of the prefix with the beginning of the word's root (Real Academia Española, 2010, page 197). For example, *reunir* ('to meet') has the prefix *re-* and its syllabification is *reu.nir*.
- When the prefix ends with a weak vowel and the root of the word begins with a vowel, it is possible to build a diphthong by joining the end of the prefix with the beginning of the word's root (Real Academia Española, 2010, page 197). For example, *diencéfalo* ('betweenbrain') has the prefix *di-* and its syllabification is *dien.cé.fa.lo*.
- When the prefix ends with an increasing diphthong and the root of the word begins with a weak vowel, it is possible to build a triphthong by joining the end of the prefix with the beginning of the word's root (Real Academia Española, 2010, page 198). There are 9 prefixes currently in use ending with an increasing diphthong (*audio-*, *bio-*, *cardio-*, *crio-*, *dia-*, *medio-*, *quimio-*, *radio-*, *socio-*), but only *dia-* is non-prominent, and there are no words with this prefix fitting this rule.
- When the prefix ends with a weak vowel and the root of the word begins with a decreasing diphthong, it is possible to build a triphthong by joining the end of the prefix with the beginning of the word's root (Real Academia Española, 2010, page 198). All the found examples have a prominent prefix (*anti-autovía*, *anti-europeo*, *bi-auricular*, *semi-automático*).
- When the prefix ends with a consonant and the root of the word begins with a vowel, it is possible to join the consonant ending the prefix with the vowel starting the root of the word into the same syllable (Real Academia Española, 2010, page 403). For example, *subalterno* ('subordinate') has the prefix *sub-* and its syllabification is *su.bal.ter.no*.
- When the prefix ends with a occlusive or fricative consonant and the root of the word begins with 'l' or 'r', it is possible to join the consonant ending the prefix with the consonant starting the root of the word into the same syllable (Real Academia Española, 2010, page 403). For example, *subrayar* ('noble') has the prefix *sub-* and its syllabification is *su.bra.yar*.

For all the cases listed above, the presence of an 'h' at the beginning of the root does not alter the rule because the soundless feature of the 'h'.

Prefix's prominence issue arises two questions when we try to develop an automatic syllabification tool:

- How can we know if a word contains prefixes?
- How can we know the prominence of such prefixes?

The answer to both questions is that the identification of prefixes requires morphological knowledge. In addition, the answer to the second question requires lexical knowledge. The methods and tools proposed to look for prefixes in a word and to evaluate its prominence are described in next sections.

3.1. Prominence as visibility and meaning

Bezos (2006) says that "words composed by a prefix and a root whose meaning be that of the separate components, will be splitted by its components if the forms of those components are not changed as a result of their union". This rule implies that prominence of prefixes depends on both, the semantic contribution of the prefix to the meaning of the word, and the morphological visibility of the two components.

The visibility of a prefix is affected by how it is attached to the root. In many cases, some orthographic or grammatical rules must be applied, producing changes to the form of the prefix, the root or both. These addition rules are discussed at Section 4.

Circumfixation is the simultaneous addition of both a prefix and a suffix to a root for form a new word. When we remove the prefix only of the new word, it is not recognizable, because it does not exist in that form. For examples, the word *embarcar* is formed by *en-* + *barco* + *-ar*, so if we remove the prefix *en-* only, we get the word **barcar* that does not exist.

> **Prominence rule #0**: Prefixal syllabification cannot be applied when the prefix or the root to what it is attached changes its form as consequence of the prefixing process, nor when the word is formed by circumfixation

4. Lemmatization as tool to looking for prefixes

As a first step in the looking for prefixes process, we propose the lemmatization of the word form; for that goal we use a lexical database of 196,597 canonical forms, producing 4,980,387 inflected forms (Carreras-Riudavets, Rodríguez-del-Pino, Hernández-Figueroa, & Rodríguez-Rodríguez, 2012). The 196,597 canonical forms correspond to words having consolidation in the language and included in the main dictionaries of Spanish. The 4,980,387 inflected forms include conjugated forms of verbs and, for other categories:

- Gender and number for the substantives, adjectives, pronouns and articles.
- Heteronomy by change of sex in the substantives, superlative for adjectives and adverbs.
- Adverbialization for the superlative.
- Diminutive, augmentative and pejorative for substantives, adjectives and adverbs.
- Graphic variants in all grammatical categories.
- Invariant forms such as prepositions, conjunctions, exclamations.
- Some neologisms originated from words of other languages and several expressions or phrases.

As prefixes can be added to any inflected form, first of all we must to enunciate the following general rule for syllabification:

> **Prominence rule #1**: For the syllabification of any inflected form, prefixes must be managed in the same way as for the corresponding canonical form

The lemmatization process is performed in two stages. The first stage is to search for the word form in the inflected forms database. If the word form is found, we get information about its inflection, its frequency of apparition in the CREA, and the id of its canonical form, from which we can obtain the canonical form itself and its grammatical category.

4.1. Second stage of lemmatization: looking for prefixes

When the first stage of lemmatization is unable to find a match for the supplied word form in the inflected forms database, it is the time to consider the presence of prefixes. Second stage of lemmatization searches for possible prefixes and tries to identify the remaining word.

The word form is examined from right to left trying to find the longest-heading-substring matching for a prefix. The remaining word has to be of length equal or greater than two, because, for Spanish, prefixes cannot be attached to words having a length lower than two. A list of valid prefixes is used for this purpose.

When a possible prefix is identified, it is removed, and the remaining word is lemmatized, again applying the two stages process. This recursive solution is capable to identify multi-prefixed words, like *postcontrarreforma*, composed by *post-* + *contra-* + *reforma*.

As far as the remaining word must be recognized to confirm the prefix, neologisms formed by circumfixation are refused.

The prefix separation process could be configured to get multiple responses or only one. The "only one" option searches for a response composed for the longest prefixes found by the process described above. The "multiple" option provides all combinations of prefixes that the process can find. For example, Fig. 1, shows that *biogenético* could be decomposed as *bio-genético* ('generator of life') or *bi-o-genético* ('twice by generator'), but the "only one" option provides only the first one response. For syllabification we prefer the "only one" option, because a simpler combination of few-longer prefixes seems to be more prominent than a more complex one composed by many-shorter prefixes.

Identifying a prefix is not as simple as splitting a heading substring from a word form and search for it in a list. The basic rule is that a prefix is attached to a root by simple concatenation, for example, the union of *ciber* + *café* produces *cibercafé* ('cybercafe'), but there are some exceptions due to grammatical or orthographic reasons that produce a number of changes in the prefix, the root, or both. These changes must be taken into account when looking for prefixes in a word form.

To take into account the changes, the list of valid prefixes must be extended to include all possible variations for each prefix (linked to the canonical form of the prefix), but this is not enough; some additional actions must be done to revert the changes and obtain the original prefix and root. These actions include: character and marks removing rules, character addition rules, character changing, rules, and rules for special prefixes.

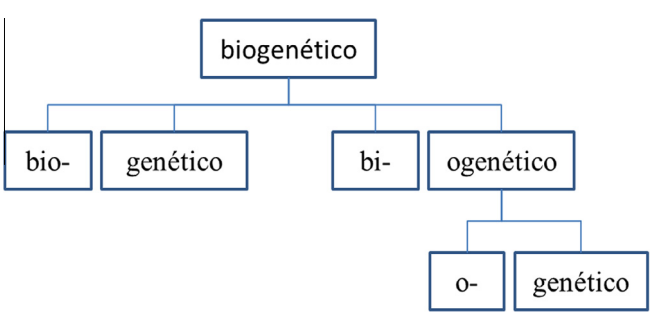

Fig. 1. Prefix separation.

4.1.1. Character and marks removing rules

There are two classes of removing rules: rules for remove accent marks and rules for remove characters.

When the separation of a prefix produces a root that is a monosyllabic word having an accent mark, the accent mark must be removed. The general rule is that monosyllabic words do not carry written accent mark, although there are some exceptions due to the fact that accent marks are used in Spanish not only to o indicate where the stress or emphasis falls on a word when it is pronounced but also to differentiate between identically spelled words; examples of this are: *tu* ('your') and *tú* ('you', subject), *si* ('music note') and *sí* ('yes'), or *té* ('tea') and *te* ('you', indirect object). When a prefix is added to a monosyllabic word, the result is a polysyllabic oxytone word that must carry a written accent mark if ending with 'n', 's' or vowel; so, when doing the opposite process, the accent mark must be removed.

When the separation of a prefix ending with a strong vowel produces a root that is a disyllabic word, ending with 'n', 's', or a vowel, and beginning with a weak having an accent mark, the accent mark must be removed. When adding a prefix ending with a strong vowel to a disyllabic paroxetine word starting with a weak vowel and ending with 'n', 's', or any vowel, an accent mark must be added to the weak vowel to avoid the formation of a diphthong; for example *pre-* + *uva* ('grape') produce *preúva*. When doing the opposite process, the accent mark must be removed.

When the separation of a prefix ending with a vowel produces a root beginning with a double 'rr', one 'r' must be removed. When adding a prefix ending with a vowel to a root starting with a 'r' a second 'r' must be added in order to preserve the alveolar trill sound /r/ that in Spanish is represented as 'rr' inside of words to difference it from the alveolar tap sound /ɾ/ represented by a single 'r'. When doing the opposite process, the additional 'r' must be removed, as in *contrarreloj*, composed by *contra-* and *reloj* ('clock').

4.1.2. Addition rules

When the separated prefix ends with a vowel and the separated root is not recognized, the vowel ending the prefix must be replicate at the beginning of the root and the recognition must be tested again. When attaching a prefix that ends with a vowel to a root starting with the same vowel, it is possible, but not mandatory, remove one of the repeated vowels, for example: *contra-* + *almirante* can produce *contralmirante* or *contraalmirante* ('rear admiral'), *anti-* + *imperialismo* can produce *antimperialismo* or *antiimperialismo* ('anti-imperialism'), *micro-* + *organismo* can produce *microrganismo* or *microorganismo* ('microorganism'). Removing is usually banned when it produces a form that is equal to an existing word which has a different meaning, and when the vowel starting the root is itself a prefix, for example: *re-* + *emitir* produce *reemitir* ('reissue'), not *remitir* ('refer'), and *semi-* + *ilegal* produce *semiilegal* ('near to be out of law'), not *semilegal* ('near to be in law') because *ilegal* is composed of the prefix *i-*, a variation of *in-* meaning negation, and the root *legal*, meaning 'in law'.

When the separation of a prefix ending with 'l' or 's' produces a root beginning with a vowel, there are two possible word forms for the remaining word: the root as is, and the same root with the character at the end of the prefix replicated at its start. The group 'ss' is not used in Spanish and the group 'll' represents a different sound than 'l'; so when a prefix ending with 's' is attached to a word beginning with 's', or a prefix ending with 'l' is attached to a word beginning with 'l', the double consonant is reduced to a single one. For example, *mal-* + *laboral* produces *malaboral*, not *mallaboral*, or *trans-* + *sexual* produces *transexual*, not *transsexual*. The reduction is not applicable for other groups of double consonants that are valid for Spanish, as 'nn', 'bb' or 'rr'. For example, *con-* + *notar* produces *connotar*, *sub-* + *bloque* produces *subbloque*, and *hiper-* + *rebelde* produces *hiperrebelde*. The prefixes ending

with 's' or 'l' are: *cis-*, *des-*, *dis-*, *es-*, *plus-*, *pos-*, *res-*, *trans-*, *tras-*, *social-* and *mal-*.

4.1.3. Change rules

When the separation of a prefix ending with 'm' produces a root beginning with 'b' or 'p', the 'm' ending the prefix must be changed by 'n'. When adding a prefix ending with 'n' to a word starting *with* 'b' or 'p', the 'n' must be changed by 'm', following orthographic considerations. When doing the opposite process, the 'm' must be reverted to 'n', as in *biempensar*, composed by *bien-* and *pensar*. This rule only affects to eight prefixes: *bien-*, *circun-*, *con-*, *en-*, *in-*, *pan-*, *pen-*, and *sin-*.

4.1.4. Rules for special prefixes

The prefixes *a-* and *ana-* only can be attached to words beginning with consonant. If *the* word begins with vowel, the variant *an-*, common for both prefixes must be used instead; so is the separation of one of these prefixes results in a root beginning with a vowel, the separation is not valid (the separated prefix is not really present in the word).

The prefixes *arc-* and *arz-*, which are variant forms of the prefix *archi-* must be followed by a root beginning with a vowel.

The prefix *in-* can be followed by a word beginning by any letter but not 'l' or 'r'; is this is the case, the variant *i-* must be used, like in *ilegal* composed by *in + legal*.

4.2. Effects of the irregularities over the prominence of the prefixes

The irregularities listed above do decrease the visibility of both the prefix and the root; moreover, when the union of a prefix with a root produces changes on the form of the prefix, the root, or both, it is not possible to syllabify the prefix separately because it is strongly linked to the root, and the separation could produce incorrect results. For example, the word *contralmirante* ('rear admiral') is composed by the prefix *contra-* (which is usually prominent) and the word *almirante* ('admiral') but, if for syllabification, we separate the prefix *contra-* we obtain the root **lmirante*, that is not a Spanish word, and the syllabification **con.tra.lmi.ran.te*, with an illegal onset 'lm'. In opposition, if we separate the word *almirante*, we obtain the prefixal form **contr-*, that is not a recognizable prefix, and the syllabification *contr.al.mi.ran.te*, with an illegal coda 'ntr'; so *contraalmirante* must be syllabified as *con.tral.mi.ran.te*, merging the end of the prefix and the start of the root. A similar example is *microspora* ('microspore') that if composed of *micro-* (prefix meaning 'very small') and *espora* ('spore') but cannot be syllabified as **mi.cro.spo.ra* because 'sp' is an illegal onset and also **spora* is not a Spanish word. Finally there are cases like *inepto* ('inept') that was composed by the prefix *in-* (meaning 'negation') and the word *apto* ('suitable'), but has suffered the change of the 'a' of *apto* by a 'e', producing **epto*, that is not a reckonable word, although the separation does not produce any illegal coda or onset.

> **Prominence rule #2**: All new words formed by a prefixing process must be syllabified by their components, i.e. applying prefixal syllabification, unless it cannot by applying prominence rule #0

5. Derivation families as tool to in-deep looking for prefixes

Prefixes detected by the process described on the previous section are part of words that are not consolidated in the language. When a word originally composed by a prefix and a root is consolidated in the language, the first stage of lemmatization identifies it as a unit and the presence of a prefix is not detected. A prefix embedded in a consolidated word probably has less prominence that an externally added prefix, but, in many cases, the remaining prominence can suffice to justify a prefixal syllabification. Taking into account those embedded prefixes requires information on how words were formed. We obtain this information from a derivation database composed of primitive-derivative pairs.

These derivation relationships are established between pairs of canonical forms and affect to all inflected forms from them. A derivation relationship between a pair of canonical forms implies that one of them can be derived from the other by adding a prefix, a suffix, or both (circumfixation). The meaning of the derived word must be composed from the meaning of the original word and the semantics of the affixes. A derivation relationship may imply a change of grammatical category, like between *amortizar* ('amortize', verb) and *amortización* ('amortization', noun), more likely when the relationship is established by the addition of a suffix.

A derivation relationship is synchronic. It reflects that a word could be derived from another, independently from what was the case when the word was formed at first. For example, the Spanish words *culpar* ('culpate') and *inculpar* ('inculpate') came from the corresponding latin words *culpare* and *inculpare*, but we have established a relationship between them because, formally, functionally and semantically, *inculpar* could be derived from *culpar* by adding the prefix *in-*. A speaker can identify the prefix *in-* independently of if he knows how the word was formed, so in order to get a correct syllabification, the prefix *in-* must be separated.

A derivation family is a set of words that are related by derivation relationships: two words, w1 and w2, are in the same family if there is a derivation relationship between them, or there is a word, w3, such as there is a derivation relationship between w1 and w3 and between w2 and w3.

We used a derivation database of 120,435 suffixal relationships, 13,896 prefixal relationships and 3872 relationships by circumfixation (Carreras-Riudavets, 2002). This database has currently 62,203 families. The relationships have been established using 195 suffixes, 237 prefixes and 316 pairs of prefix-suffix for the circumfixation.

Fig. 2 shows an example of a derivation family, composed by six words. The family root is a word without any affixes, *amortizar* ('amortize'). The words *amortización* ('amortization') and *amortizable* ('amortizable') can be derived from *amortizar* by adding a suffix, while the word *desmortizar* can be built by adding a prefix. The word *desamortización* could be built by adding the prefix *des-* to the word *amortización*, or the suffix *-ación* to the word *desamortizar*, and the word *desamortizable* could be built by adding the prefix *des-* to the word *amortizable*, or the suffix *-able* to the word *desamortizar*. Both *desamortizable* and *desamortización* could be parasynthetic derivatives from *amortizar* in case of no existence of the intermediate words *amortización*, *desamortizar*, nor *amortizable*, but they are not because these words exist.

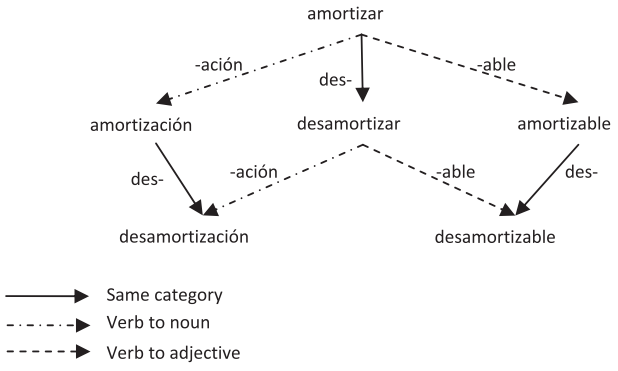

Fig. 2. Derivation family.

Derivation relationships allow finding "hidden prefixes", prefixes that the lemmatizer does not recognize because they form part of consolidated words that the lemmatizer recognizes as a whole. While prefixes used to create neologisms are intentionally used, and so prominent, prefixes used in the past to create a word that nowadays has been consolidated may be or not be prominent. More information about words and prefixes must be taken into account to decide; next section treats about that information and how use it, but a rule about prefix's prominence can be obtained from derivation families:

Prominence rule #3: The prominence of a prefix for words in the same derivation family is transitive

That is, if a word includes a prefix, and there are others words that can be built from that word by adding one or more suffixes, the derived words conserve the prominence of the prefix in the original word. For example, the lemmatization of the word *desordenadamente* ('untidily') says us that it is an adverb, without any information about the presence of a prefix. The derivation database is useful to find the presence of a prefix in consolidated words. The derivation database says us that *desordenadamente* is composed by *desordenado* ('untidy') + *-mente*, and the word *desordenado* ('untidy') is composed by *desordenar* + *-ado*. The word *desordenar* is a consolidated word that is composed by the prominent prefix *des-* and the root *ordenar*; so the prefix *des-* must be considered prominent in the words *desordenado* and *desordenadamente*.

6. Decisions about prominence of prefixes

As we said above, the prominence of a prefix depends on both, its influence over the meaning of the word, and how much visible it is. In previous sections, we explain how first separating the more visible, unconsolidated prefixes, considering that all of them are prominent as far as they conserve their forms and the roots to what they are attached conserve their forms too. Then we describe the use of derivation information to find less-visible prefixes that are embedded into consolidated words. Now we must decide about the prominence of these prefixes. In order to define general criteria, we have examined large samples of words carrying prefixes; for this examination we take into account several factors as: morphology, semantics, etymology, use and productivity.

6.1. Productivity of prefixes in the current Spanish

The productivity of a prefix is a measure of its contribution to the formation of new words. The idea is that if a prefix is frequently used to create new words, it is an "alive prefix", so more prominent than a prefix that is never used to create new words.

We have examined 15,706 new words, 3449 from the Dictionary of neologisms on-line (Freixa et al., 2011), and 12,257 from New Dictionary of voices of the current use (Alvar, 2003), and we found that 6210 of them are formed by de addition of a prefix. These new words were formed using 138 different prefixes, which are listed at Table 1 sorted by their frequencies of use. As we can see the top ten prefixes at Table 1 sums near of the 50% of the productivity.

Table 1
Frequencies of the used prefixes to create new words.

Pref.	Freq.	Pref.	Freq.	Pref.	Freq.	Pref.	Freq.
auto-	752	anti-	530	super-	344	tele-	252
re-	252	des-	248	mini-	176	post-	167
pre-	155	micro-	147	ciber-	142	semi-	138
multi-	134	seudo-	129	bio-	102	hiper-	101
neo-	101	in-	100	macro-	97	euro-	93
inter-	89	sobre-	86	sub-	84	contra-	82
ultra-	78	a-	75	pro-	75	con-	67
extra-	59	video-	57	mono-	49	foto-	47
eco-	47	mega-	46	en-	40	bi-	39
radio-	38	trans-	35	psico-	34	neuro-	30
meta-	28	electro-	28	nano-	28	intra-	25
tecno-	25	poli-	24	hidro-	23	aero-	23
narco-	22	agro-	22	archi-	22	vice-	21
pluri-	21	para-	21	paleo-	21	pan-	19
uni-	19	geo-	19	retro-	19	info-	17
proto-	15	tri-	15	socio-	15	termo-	14
de-	14	moto-	13	filo-	13	maxi-	13
infra-	12	franco-	12	supra-	12	mal-	11
afro-	11	ciclo-	11	hispano-	10	turbo-	10
cuasi-	9	hipo-	9	endo-	9	porno-	9
medio-	9	anglo-	8	audio-	8	tardo-	8
tetra-	8	centro-	8	social-	7	crio-	7
giga-	7	cardio-	6	ex-	6	penta-	6
etno-	5	alter-	5	hetero-	5	fito-	5
germano-	5	cuadri-	5	homo-	5	bien-	4
requete-	4	cito-	4	hemi-	4	di-	4
ana-	4	italo-	4	zoo-	3	quimio-	3
mili-	3	ante-	3	andro-	3	anarco-	3
xeno-	3	crono-	3	tera-	3	epi-	2
sin-	2	omni-	2	xero-	2	peta-	2
magneto-	2	exo-	2	hepta-	2	ecto-	1
es-	1	dis-	1	peri-	1	sota-	1
entre-	1	iso-	1	plus-	1	baro-	1
meso-	1	deca-	1	cuarto-	1	kilo-	1
cromo-	1	apo-	1				

6.2. Classification of the prefixes

As a result of our analysis we found five groups of prefixes:

(1) Prefixal-compositive elements.
(2) Prefixes that always require prefixal syllabification.
(3) Prefixes whose syllabification depends on their meaning.
(4) Prefixes whose syllabification depends on the use of the words into where they are attached.
(5) Prefixes that usually require non-prefixal syllabification.

6.2.1. Prefixal-compositive elements

The *New Grammar of the Spanish Language – NGSL* (Real Academia Española, 2009) defines prefixal-compositive element as "Greco-Latin lexical base that has an intermediate status between bound and free forms" while define prefix as "morpheme that precedes the lexical base". Prefixal-compositive elements are prefixes as far as they are morphological units that are attached to a lexical base in a preceding position. The main difference seems to be that the *NGSL* considers that prefixal-compositive elements build words by composition, while the "normal" prefixes built words by derivation. The border between prefixes and prefixal-compositive elements is fuzzy, but from the syllabificatioń s viewpoint, prefixal-compositive elements could be considered simply as prefixes with an especial prominence. Examples of prefixal-compositive elements are: *acro-*, *aero-*, *afro-*, *agro-*, *alter-*, *andro-*, *anfi-*, *anglo-*, *aniso-*, *anisó-*, *ante-*, *archi-*, *arque-*, *arqui-*, *audio-*, *auto-*, *baro-*, *bi-*, *bien-*, *bio-*, *cardio-*, *centi-*, *centro-*, *ciber-*, *ciclo-*, *circun-*, *cito-*, *cromo-*, *crono-*, *cuadri-*, *cuarto-*, *cuasi-*, *cuatri-*, *deca-*, *deci*, *di-*, *eco-*, *ecto-*, *electro-*, *endo-*, *enea-*, *equi-*, *etno-*, *euro-*, *filo-*, *foto-*, *franco-*, *geo-*, *germano-*, *giga-*, *hecto-*, *hemi-*, *hepta-*, *hetero-*, *hexa-*, *hidro-*, *hiper-*, *hipo-*, *hispano-*, *homo-*, *info-*, *infra-*, *iso-*, *kilo-*, *macro-*, *mal-*, *maxi-*, *medio-*, *mega-*, *megalo-*, *meso-*, *meta-*, *micro-*, *mili-*, *mini-*, *mono-*, *moto-*, *multi-*, *nano-*, *narco-*, *neo-*, *neuro-*, *omni-*, *paleo-*, *pan-*, *penta-*, *peta-*, *pluri-*, *plus-*, *poli-*, *polí-*, *proto-*, *pseudo-*, *psico-*, *quimio-*, *radio-*, *retro-*, *semi-*, *seudo-*, *sobre-*, *social-*, *socio-*, *sota-*, *soto-*, *super-*, *supra-*, *tardo-*, *tecno-*, *tele-*, *tera-*, *termo-*, *tetra-*, *tri-*, *turbo-*, *ultra-*, *uni-*, *vice-*, *video-*.

Prominence rule #4: Words formed by the addition of a prefixal-compositive element to a root must be syllabified applying prefixal syllabification, unless it cannot by applying prominence rule #0

The list of productive prefixes showed at Table 1 is mostly (over 78%) composed by prefixal-compositive elements.

6.2.2. Prefixes that always require prefixal syllabification

This group comprises prefixes having a high prominence: *anti-*, *co-*, *con-*, *des-*, *extra-*, *in-*, *intra-*, *post-*/*pos-*, *pre-*, *pro-*, *trans-*/*tras-*, *contra-*, *de-*, *entre-*, *para-*, and *en-*. Five of these prefixes (*contra-*, *de-*, *entre-*, *para-*, *en-*) are homographs of prepositions with the same meaning, what increments their prominence, but the prefix *en-* often creates words by circumfixation; in these cases it is considered non prominent by applying prominence rule #0.

All the prefixes in this group appear at Table 1, except *pos-* that it is a variant of *post-*, and *tras-* that it is a variant of *trans-*.

Prominence rule #5: Words formed by the addition to a root of any of the prefixes: anti-, co-, con-, des-, extra-, in-, intra-, post-/pos-, pre-, pro-, trans-/tras-, contra-, de-, entre-, para-, or en-, must be syllabified applying prefixal syllabification, unless it cannot by applying prominence rule #0

6.2.3. Prefixes whose syllabification depends on its meaning

Most of prefixes has various meanings and for some prefixes one of their meanings is much known while the rest are usually unknown, so such a prefix is prominent only when is used with the known meaning.

This group comprises the prefixes: *sin-*, *di-*, and *ex-*. The prefix *sin-* may mean 'lack' or 'union'. It is a homograph of the preposition *sin*, that means 'lack'; so, when *sin-* means 'lack', its coincidence with the preposition reinforces its prominence, as in *sin.hue.so* ('without bone' = 'tongue'), but not when it means 'union', as in *si.nal.gia*. Moreover, the meaning 'union' is usually associated to words from Latin or Greek which have been consolidated in Spanish for a long time.

The prefix *di-* is a prefixal-compositive element when it means 'two', as in *di.a.tó.mi.co* (with two atoms), and has a low prominence for any other meaning, usually corresponding to consolidated words.

The prefix *ex-* requires prefixal syllabification when it means 'out of' or 'beyond' and normal syllabification for all other cases. For example, *exoftalmia* ('exophthalmia') is syllabified *ex.of.tal.mia*, but *exornar* ('decorate') is syllabified *e.xor.nar* because, in this case, the prefix *ex-* does not carry a meaning.

Prominence rule #6: Words formed by the addition to a root of any of the prefixes: sin-, di-, or ex-, must be syllabified applying prefixal syllabification only in the case that they have the proper meaning and unless it cannot by applying prominence rule #0

6.2.4. Prefixes whose syllabification depends on the use of the words to where they are attached

In some cases prominence of a prefix may compete against the prominence of the whole word. If a word is very used and contains a prefix that is not as frequent, the presence of the prefix may be obscured. We can say that when the word is very used loses its prefixal morphology, while this morphology remains visible when the knowledge of the word is lower. The prefixes in this group are: sub-, re-, and inter-. We propose to separate them depending on the frequency of use of the word to what they are attached. We use the frequencies of the CREA wordlist for that purpose.

Prominence rule #7: Words formed by the addition to a root of any of the prefixes: sub-, re-, or inter-, must be syllabified applying prefixal syllabification only in the case that the frequency of use of the word to what they are attached is higher than a prefixed threshold, and unless it cannot by applying prominence rule #0

6.2.5. Prefixes that usually require non-prefixal syllabification

This group is composed by the little used prefixes: *a-*/*an-*, *ad-*, *ana-*, *cata-*, *dia-*, *dis-*, *e-*, *epi-*, *es-*, *peri-*, *res-*, *ab-*, *al-*, *ambi-*, *apo-*,

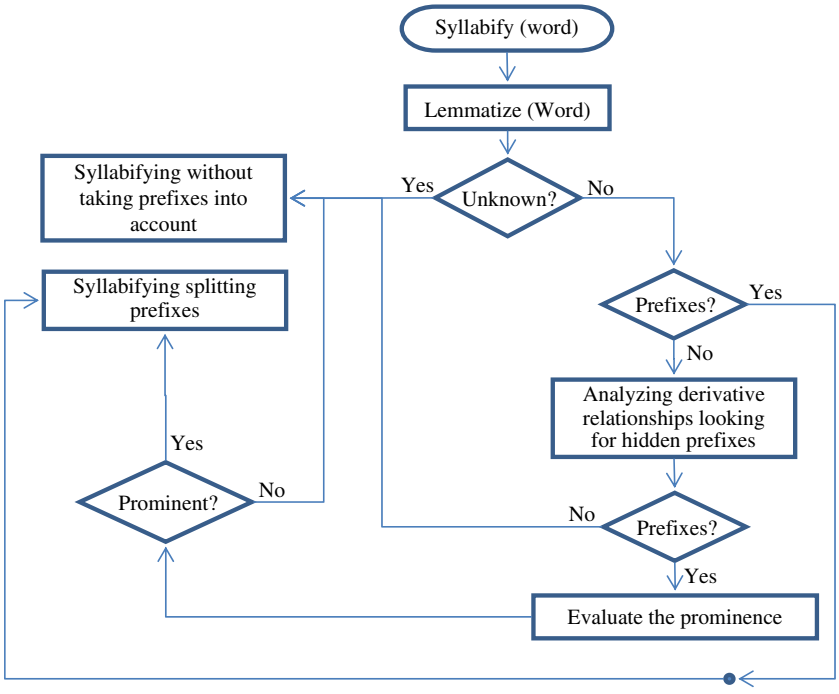

Fig. 3. Prefix prominence decision algorithm.

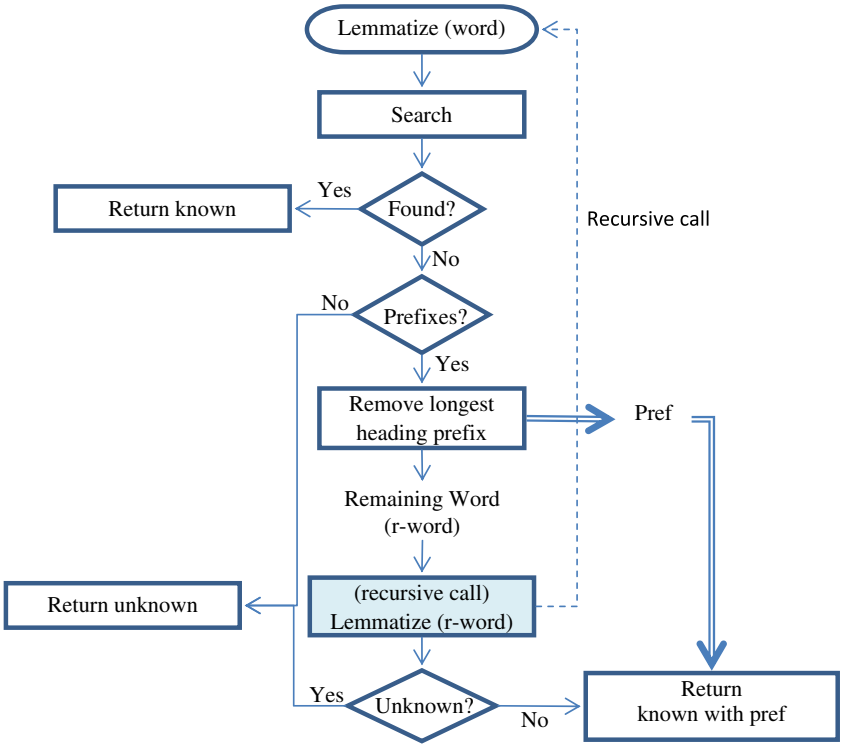

Fig. 4. Lemmatization algorithm.

Table 2
Attached prefixes to word of CREA that can be syllabified in two different ways depending on if prefixes are taken into account or not.

Pref.	Freq.	Pref.	Freq.	Pref.	Freq.	Pref.	Freq.
des-	115,103	in-	53,134	re-	16,417	co-	13,872
inter-	6199	sub-	5852	bien-	5038	mal-	4162
anti-	2451	pre-	2262	pan-	1553	hiper-	806
super-	770	trans-	746	di-	699	semi-	583
ciber-	488	post-	438	intra-	341	pos-	326
contra-	307	sobre-	296	tras-	272	bi-	271
ex-	270	poli-	260	agro-	195	auto-	189
pluri-	182	infra-	174	para-	152	electro-	131
im-	130	ad-	103	bio-	101	multi-	97
centro-	93	porta-	85	pro-	72	mega-	71
extra-	65	ultra-	62	mini-	42	en-	39
i-	37	an-	34	proto-	33	micro-	27
kilo-	25	su-	23	neuro-	22	psico-	20
guarda-	12	uni-	9	tele-	9	entre-	9
anglo-	8	mili-	8	op-	8	euro-	7
hidro-	7	foto-	6	anfi-	5	ante-	4
archi-	4	aero-	4	radio-	3	homo-	3
macro-	3	mono-	3	a-	3	endo-	2
centi-	1	em-	1	con-	1	cuasi-	1
equi-	1	magneto-	1	pseudo-	1	sin-	1
eco-	1						

bati-, *cachi-*, *circa-*, *cis-*, *citra-*, *eu-*, *exo-*, *ob-*, *pen-*, *per-*, *preter-*, *yuxta-*, and *za-*. Most of these prefixes do not appear on Table 1, meaning that they are not currently used to create new words. Only the prefix *a-* occupies a significant position on Table 1, but it is a prefix composed by a single character that is the character most frequently used at the beginning of Spanish words, so it is difficult to be recognized as a prefix.

7. Algorithm implementation

Fig. 3 shows a diagram of the proposed prefix prominence decision algorithm. The first step is to lemmatize the word. The lemmatization is a recursive process (Fig. 4) that begins searching for the word into lexical database. If the word is found, its lemma is returned to the main process.

When the word is not found into lexical database, the lemmatizer tries to find and remove a prefix in the word and recursively lemmatizing the remaining word. If a prefix is not found or the remaining word is unknown, the initial word is returned as unknown too; the other hand, both are returned: the prefix and the remaining word.

If the lemmatizer returns a word as unknown, it is syllabified without taking prefixes into account. If the lemmatizer returns a word as known with one or more prefixes, it is syllabified splitting

the prefixes. If the lemmatize returns a word as just known, the main process analyses derivative relationships looking for hidden prefixes. If there are hidden prefixes, their prominences will be evaluated using the decision factors as described above and the word will be syllabified accordingly.

8. Effects of the prominence rules on the syllabification of the CREA wordlist

The *Corpus del Español Actual* (Corpus of Current Spanish, CREA) is a corpus built by the Royal Academy of Spanish Language. It contains 737,799 different words, producing a total of 152,558,294 occurrences. 356,185 of these words are recognized by our lemmatizer, and 381,614 are not. The unrecognized words are more than the 50% of the different words; but they represent less than the 2% of the occurrences. Moreover, 212,562 of the non-recognized words appear only once in the CREA, this number is over the 55% of the non-recognized words. The non-recognized words include: names and family names, foreign words, orthographic errors, onomatopoeias, and neologisms.

We have applied our syllabification method to the recognized words in the CREA, and we have found that 12,571 of them could be syllabified in two different ways depending on if prefixes are taken into account or not. This 12,571 words produce a total number of 365,562 occurrences in the CREA, this numbers are about a 1.7% of the words, but only a 0.23% of the occurrences. Table 2 shows the prefixes attached to those words, we can see that most of them are in the top half of Table 1, corresponding to the more productive prefixes.

9. Conclusions

Although the syllabification of Spanish words seems to obey a few simple rules, the presence of prefixes introduces a complexity factor that is not easy to treat. Words containing prefixes could vary their syllabification with respect to that of the standard rules in order to syllabify the prefixes as separated units, but this separated syllabification applies only when the prefix is considered prominent into the word. A basic rule to consider that a prefix can be prominent is that both, the prefix and the unprefixed word to what it is attached must be recognizable, that is, they must not suffer any significant change during the formation of the prefixed word.

Automatic determination of prefixes prominence is a difficult issue. In some cases prominence depends on cultural or geographical aspects which may cause confusion or ambiguity, but in many others it may be inferred if enough knowledge about the word is taken into account.

However, to take prefixes into account, the first requisite is to be able of recognize them when they are attached to a word. The presence at the beginning of a word of a characters sequence that is equal to a known prefix is not enough. It is needed that such a sequence is functioning really as a prefix, that is: when we remove it, we obtain a real word whose meaning is properly modified when the prefix is attached.

Prefixes recognition requires morphological knowledge about the words and the relationships between them. A lemmatizer with the capability to recognize neologisms formed by the addition of a prefix to a consolidated word can identify the prefixes used in such a case. Prefixes used to build neologisms must be considered as having a high prominence because they are intentionally used to create new words.

A normal lemmatizer does not recognize prefixes when the prefixed word has been consolidated in the language. The information provided by the derivation database is useful in this case because it makes manifest the hidden prefixes originally used to create that word.

Once the prefixes have been identified, additional information is needed to calculate their prominence. In this work we use information about prefixes' current productivity, according to various dictionaries of neologisms, and frequency of use of the words in the Corpus of Current Spanish (CREA) published by the Spanish Royal Academia. This information allows assign prominences to the prefixes and takes it into account for syllabification.

Future research would compile more information about prefixes, words, and about the context where they are used. As far as prefix identification is related to semantic issues, taking into account the word context in the discourse could be used to select the proper syllabification solution when more of one is possible. Other corpus could be used to refine the knowledge about of use for the words. A collection of Spanish texts from Internet using a web crawler could be an alternative to get information about the current use of the words. Adding the option to select the geographical context could be and interesting feature for the syllabifier. A web-based syllabifier could automatically get this information from the web-browser setup. All this information could be very useful to improve syllabification, although we have to remember that the prefix prominence issue seems to affect to less than the 2% of the words currently used in Spanish, so the effects of future improvements could be marginal.

Another research line to board in the future is to study the case of composite words and their influence on the syllabification. The effects of our proposal over other linguistic issues must be study too, as the phonology and the hyphenation.

References

Adsett, C. R., & Marchand, Y. (2009). A comparison of data-driven automatic syllabification methods. *String processing and information retrieval* (pp. 174–181). Berlin Heidelberg: Springer.

Alvar, Manuel (2003). *Nuevo diccionario de voces de uso actual*. Madrid: Arco Libros.

Anusuya, M., & Katti, S. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security, 6*(3), 181–205.

Bezos, Javier (2006). División de palabras con guiones. Available on-line at: http://www.let.rug.nl/alfa/tex/tetex30/generic/spanish/division.pdf. Accessed 2.05.2012.

Cambridge Universtity Press. *Cambridge Dictionaries Online*. Available on-line at: http://dictionary.cambridge.org. Accessed April, 2012.

Carreras-Riudavets, Francisco. (2002). Sistema Computacional de Gestión Morfológica del Español (SCOGEME). Doctoral thesis. Procesamiento del Lenguaje Natural, n° 28, pp. 105–106.

Carreras-Riudavets, Francisco, Rodríguez-del-Pino, Juan. C., Hernández-Figueroa, Zenón., & Rodríguez-Rodríguez, Gustavo. (2012). A morphological analyzer using hash tables in main memory (MAHT) and a lexical knowledge base. *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, 7181*, 80–91.

Cuayahuitl, Heriberto. (2004). A syllabification algorithm for Spanish. Computational linguistics and intelligent text processing. *Lecture Notes in Computer Science, 2945*, 412–415.

Freixa, Judit, Raúl Araya, Jenny Azarian, Bernat Bardagil, Mariona Barrera, Marc Colell, Alba Coll, Sabela Fernández-Silva, Marta Folia, Cristina Mayoral, Albert Morales, Magalí Sirera, and Jesús Carrasco. *Diccionario de neologismos on line*. Available on-line at: http://obneo.iula.upf.edu/spes. Accessed 10.01.2011.

Ganapathiraju, Aravind, Hamaker, Jonathan, Picone, Joseph, Ordowski, Mark, & Doddington, George R. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing, 9*, 358–366.

Goddard, John, René MacKinney-Romero (2006). Finding Spanish syllabification rules with decision trees. In *Proceedings of the 5th international conference on Advances in Natural Language Processing* (pp. 333–340).

Janakiraman, Rajesh, Chaitanya J. Kumar, Hema A. Murthy (2010). Robust syllable segmentation and its application to syllable-centric continuous speech recognition. In *National Conference on Communications* (pp. 1–5).

Lopez-Gonzalo, Eduardo José M. Rodríguez-García. (1996). Statistical methods in data-driven modelling of Spanish prosody for text speech. In *The Fourth International Conference on Spoken Language Processing (ICSLP 96)* (pp. 1377–1380).

Rao, Nageshwara, M., Samuel Thomas, Nagarajan, T., Hema A. Murthy. (2005). Text-to-speech synthesis using syllable-like units. In *Proceedings of National Conference on Communications ITT, India* (pp. 277–280).

Real Academia Española. (2005). *Diccionario panhispánico de dudas*. Real Academia Española, Madrid.
Real Academia Española. (2009). *Nueva gramática de la lengua española*. Real Academia Española y Asociación de Academias de la Lengua Española, Madrid.
Real Academia Española. (2010). Ortografía de la lengua española. Real Academia Española y Asociación de Academias de la Lengua Española, Madrid.
Real Academia Española. Banco de datos (CREA). *Corpus de referencia del español actual*. Available on-line at: http://www.rae.es. Accessed Feb, 2011.
Real Academia Española. *Diccionario de la lengua española*. Available on-line at: http://www.rae.es. Accessed April, 2012.
Ualde, José Ignacio (1989). Silabeo y estructura morfémica en español. *Hispania, 72*(4), 821–831.