Acquisition of morphological families and derivational series from a machine readable dictionary

Nabil Hathout

Acquisition of morphological families and derivational series from a machine readable dictionary

Nabil Hathout
Université de Toulouse

The question we address in this paper is: how to perform morphological analysis in the framework of word-based morphology, that is without resorting to the notions of morpheme, affix, morphological exponent or any representation of these concepts? We do not present here a fully fledged answer, but we describe a general framework for doing so and a method for computing a large part of the intended analysis. The paper is divided into five parts. In section 1, we outline the objectives of the research and the method. We then detail the measure of morphological similarity (section 2), and, the formal analogy we use to filter the morphological neighborhoods (section 3). We then present some preliminary results (section 4) and a short conclusion (section 5).

1. Toward a computational word-based morphology

1.1. Word-based vs morpheme-based morphology

In standard morpheme-based morphology, words are made up of morphemes. The morphemes are combined by rules of inflection, derivation and composition. They have structures which are usually represented as trees like the ones in figure 1. Morpheme-based morphology is both elegant and easy to use, but it suffers from many drawbacks (Anderson, 1992; Aronoff, 1994); there is no need to enumerate them here.

Figure 1: Word structure of the French noun *dérivabilité* ‘derivability’ and the French adjective *dérivationnel* ‘derivational’ in morpheme-based morphology.

In word-based morphology (Aronoff, 1976; Bybee, 1988, 1995; Neuvel & Singh, 2001; Burzio, 2002; Blevins, 2006), the minimal units are the words. Therefore, they do not have any structure. Morphological structure then becomes a level of organization of the lexicon, made up of the morphological relations that hold between the words. Some of them play a special role, namely the relations between the words that belong to the same lexeme, to the same inflectional series, to the same morphological family and to the same derivational series. These four types of aggregates can be illustrated by the lexeme and the inflectional series of the French verb form *dérivons* ‘derive’ and by the morphological family and the derivational series of the deverbal noun *dérivation*:

- the **lexeme** of *dérivons* contains the inflected forms of the verb *dériver* ‘derive’: *dériver*, *dérive*, *dériverez*, *dérivaient*, *dérivées*, *dérivions*, etc.;
- *dérivons* belongs to an **inflectional series** of first person plural, present indicative verb forms which also contains *acclimatons* ‘acclimate’, *compilons* ‘compile’, *éduquons* ‘educate’, *localisons* ‘localize’, *varions* ‘vary’, etc.;
- the **morphological family** of *dérivation* contains words such as *dériver*, *dérivable*, *dérivatif* ‘derivative’, *dérivationnel*, *dérivabilité*, *dérive* ‘drift’, *dériveur* ‘sailing dinghy’, etc.;

Cascadilla Proceedings Project
Completed May 11, 2009

- *dérivation* belongs to a **derivational series** of deverbal nouns in *-ion* such as *acclimatation* 'acclimation', *compilation*, *éducation*, *localisation*, *variation*, etc.

In the rest of the paper, we concentrate only on the derivational part of the morphological structure.

Notice that morphology does not reduce to this lexical structure. For instance, *anti universal healthcare* in (1) is a morphological construct[1] that is not likely to enter the lexicon nor have a place in the structure.

(1) *All those anti feminist, anti Democrat, anti giving everyone the right to vote,* ***anti universal healthcare****, anti all kinds of things I thought no one was anti.*
`www.talkleft.com/story/2008/4/18/204142/362`

Actually, lexicon and morphology must be clearly separated: the main function of the lexicon is to memorize and organize the words that a speaker knows; the main function of the morphology is to produce and analyze words. The constructs produced by the morphology are designed to enter the lexicon and extend the lexical structure. In return, the lexicon provides the morphology with the information it needs to create and analyze morphological constructs. However, the distribution of morphological information between lexicon and morphology varies according to morphological theories. In morpheme-based morphology, each word has a separate structure, the lexicon is just a bag of morphemes (and possibly of fully analyzed words) and morphological rules encode the bulk of the morphological information. In word-based morphology, the distribution of the information is more even. The lexicon contains both the words and the morphological relations that hold between them. Morphology is made up of processes that extend the existing lexical structure with new words. These processes can also be used to create constructs such as (1) that have no place in the lexicon. In this paper, we are concerned only with the lexicon structure.

The morpheme-based *vs* word-based distinction shows up on the computational level. In the morpheme-based conception, the morphological analysis of a word aims at segmenting it into a sequence of morphemes (Déjean, 1998; Gaussier, 1999; Schone & Jurafsky, 2000; Goldsmith, 2001; Creutz & Lagus, 2002; Bernhard, 2006). For instance, *dérivation* is analyzed as made up of two segments *dérivat* and *-ion*, the first being identified as the root morpheme *dériv* and the second as the suffix *-ion*. In a word-based approach, the aim of a morphological analysis is to discover the relations between the word and the other lexical items and to identify its morphological family and its derivational series. For instance, an analysis of the French word *dérivation* is satisfactory if it connects *dérivation* with enough members of its morphological family (*dériver*, *dérivationnel*, *dérivable*, etc.) and its derivational series (*formation* 'education', *séduction*, *émission*, *vision*, etc.).

The morphological relations are organized into analogical series. For instance, the relation between *dérivation* and *dérivable* is the basis of analogies such as *dérivation*:*dérivable* :: *variation*:*variable*,[2] *dérivation*:*dérivable* :: *modification*:*modifiable*, *dérivation*:*dérivable* :: *adaptation*:*adaptable*, *dérivation*:*dérivable* :: *observation*:*observable*, etc. Similarly, the relation between *dérivation* and *variation* gives rise to a series of analogies such as *dérivation*:*variation* :: *dériver*:*varier*, *dérivation*:*variation* :: *dérivationel*:*variationnel*, *dérivation*:*variation* :: *dérivabilité*:*variabilité*, *dérivation*:*variation* :: *dérivable*:*variable*, etc. These examples show how morphological analogies connect the morphological families and the derivational series.

1.2. Combining morphological relatedness and formal analogy

In the remainder of the paper, we present a computational model that makes the morphological derivational structure of the lexicon emerge from the semantic and the formal regularities of the words it contains. A first experiment is currently underway on the lexicon of French using the *Trésor de la Langue Française informatisé* machine readable dictionary (or *TLFi* for short; `atilf.atilf.fr/tlf.htm`). Our aim is to create a lexicon that provides the morphological family and the derivational series of the words it contains. This morphological lexicon owes its strength to the global description of a significant

[1] *Construct* is used in this paper as a generic term to designate any linguistic object produced by the morphology.

[2] The notation $a : b :: c : d$ is used as a shorthand for the statement that (a, b, c, d) forms an analogical quadruplet, or in other words that a is to b as c is to d.

part of the French lexicon. We are building it from a lexicographical resource because we need semantic descriptions for a large number of words. We are fully aware of the limitations of the lexicographical descriptions but the benefits of using dictionaries far exceed them. Dictionaries provide definitions and graphemic / phonological representations for a significant part of the lexicon. Besides, lexicographic descriptions are more easy to use than data extracted from corpora since they only present the sub-senses and the definitions of the most representative usages of the words.

Our method relies on a measure of morphological relatedness that brings the members of morphological families and derivational series closer. This measure takes into account both the formal and the semantic similarities between the words. The method also relies on the discovery of formal analogies among morphological neighbors. The use of analogy is quite common in computational morphology (Skousen, 1989; Lepage, 1998; Van den Bosch & Daelemans, 1999; Pirrelli & Yvon, 1999; Hathout, 2005; Stroppa & Yvon, 2005). The main novelty of the method is to combine it with a measure of morphological relatedness. First, lexical similarity is used in order to select quadruplets of words that are related to each other. The candidates are then checked by means of analogy. The two techniques are complementary. Morphological similarity can be computed for large numbers of words, but it is too coarse-grained to discriminate between the words that are actually morphologically related and the ones that are not. Formal analogy is then used to perform fine grained filtering but is costly to calculate.

More generally, our approach is original in that:

1. The computational model is purely word-based. The discovery of morphological relations between words do not involve the notions of morpheme, affix, morphological exponent, etc. or any representation of these concepts.

2. Membership in families and series is gradient. It accounts, for instance, for the fact that *dériveur* is morphologically and semantically closer to *dérive* than to *dérivationnellement* 'derivationally', even if the three words belong to the same family. The model connects the words that share semantic and / or formal features. The more features they share and the more specific these features are, the closer the words are.

3. It implements the theoretical proposals of Bybee (1988, 1995) and Burzio (2002) in a straightforward manner.

4. It is efficient enough to be used to build a large morphological resource semi-automatically.

Besides, the model integrates semantic and formal information in a uniform manner. All kinds of semantic information (lexicographic definitions, synonyms, synsets, etc.) and formal information (phonological, graphemic, syllabic, etc.) can be used. These specifications can be cumulated easily in spite of differences in nature and origin. The model takes advantage of the redundancy of the features and is fairly insensitive to variation and exceptions. It is robust and language independent.

Technically, the model joins:

1. the representation of the lexicon as a graph and its exploration through random walks, along the lines of (Gaume et al., 2002, 2005; Muller et al., 2006), and

2. formal analogies on words (Lepage, 1998, 2003; Stroppa & Yvon, 2005; Langlais & Patry, 2007). This approach does not make use of morphemes. Correspondence between words is calculated directly on their graphemic representations.

1.3. Network lexicon

The morphological lexicon we intend to build is a network of words with connections mainly defined by the morphological families and the derivational series. This primary structure is completed with a set of analogies between pairs of morphologically related lexemes and with a morphological distance. The resulting lexicon is remarkably flexible and can adequately represent various morphological phenomena. One of them is allomorphy, which corresponds to locations in the network where there is a mismatch between the formal analogies and the organization into families and series. For instance, the French deverbal noun *dénivellation* 'unevenness' can be identified as an allomorphic form because (*i*) it

belongs to a series of words ending in *-ion*, (*ii*) it is a member of the family of the verb *déniveler* 'make uneven' and, (*iii*) it is morphologically the closest noun to this verb. Nouns in *-ion* and more specifically in *-ation* are normally involved in analogies with their closest verbs such as *dérivation*:*dériver* :: *compilation*:*compiler*. The absence of such analogies for *dénivellation* appears as a gap in the analogical grid. This gap is the sign of an allormorphy. Another cue is the near identity of *dénivellation* with the string *dénivelation* which would have allowed *déniveler* to enter the main set of analogies involving the nouns ending in *-ion* (i.e. the set of analogies with the strongest morphological density).

The lexicon also accounts for the similarity and difference between *curieux* 'curious' and *furieux* 'furious' in the same way (Jackendoff, 1975). On the one hand, *furieux* can be analyzed as an adjective derived from the noun *furie* 'fury' but we cannot do so for *curieux* since it is no longer semantically related to *cure* 'care'. On the other hand, both adjectives have the formal and the semantic features of *-eux* derivatives. In the lexicon we propose, both adjectives belong to the same derivational series. On the other hand, *furieux* and *furie* participate in a series of analogies with *mélodieux*:*mélodie* 'melodious':'melody', *harmonieux*:*harmonie* 'harmonious':'harmony', *facétieux*:*facétie* 'facetious':'joke', etc. while *curieux* does not. This example shows the flexibility of our model and the higher descriptive precision we obtain from the derivational series and the morphological analogies. By contrast, the similarity of *curieux* with *furieux* cannot be described in a morphematic model or in any model lacking derivational series. Note that the term "derivational series" is a little misleading since series include both derived and non derived lexemes. Lexemes belong to the series on the basis of their form and meaning only.

Similarly, the representation of words that include interfixes such as *tartelette* 'little tart', *gouttelette* 'droplet', or *vedettariat* 'stardom' (Plénat & Roché, 2003; Plénat, 2005) does not pose any difficulty. These words are full members of their respective families and series. In these series, each of them is the nearest neighbor of its base *tarte* 'tart', *goutte* 'drop' and *vedette* 'star'. The interfixes reinforce the formal integration of these lexemes in their series.

With respect to applications, the lexicon we propose adequately fulfills the main requirements for morphological knowledge in computational linguistics and information retrieval. Morphological resources have several uses in these domains, such as prepositional phrase attachment disambiguation (Bourigault, 2007) or query expansion (Xu & Croft, 1998; Jing & Tzoukerman, 1999; Moreau et al., 2007). The morphological relations used by a syntactic parser such as Syntex (Bourigault, 2007) associate nouns and verbs from the same family with strong morphological similarities. Our lexicon will provide all these relations and even allow the users to select them with more precision. In information retrieval, the retrieval performance can be improved by expanding the queries by adding to them morphologically related words. These words are all members of the morphological families of the words of the seed queries. Besides, the morphological distance we propose can be used to tune the expansions more finely. Our lexicon can also be used in the design of psycholinguistic experimental material. The derived *vs* non derived nature of the words can be determined from their derivational series and their morphological analogies. Among the other features taken into account for the conception of experimental material, let us cite formal likeness andmembership in the same family. All this information is explicitly available in the lexicon we propose. Finally, let us stress that the relational organization of the lexicon does not pose any difficulty as proved by the number of the applications which use WordNet (Miller et al., 1990).

1.4. Related works

In this research, we adopt a global approach to the lexicon which differs from other efforts such as the MorTAL project aiming at creating a morphological database for French (Dal et al., 1999; Hathout et al., 2002). In this project, the database is made up by analyzing a selection of French affixes, one at a time, by means of the Dérif analyzer (Namer, 2005). By contrast, our objective is to create an entire lexicon at once.

Many works in the field of computational morphology aim to recover relations between lexical units. All of them rely primarily on finding similarities between the word graphemic forms. These relations are mainly prefixal or suffixal with two exceptions, (Yarowsky & Wicentowski, 2000) and (Baroni et al., 2002), who use string edit distances to estimate formal similarity. As far as we know, all the others perform some sort of segmentation even when the goal is not to find morphemes, as in (Hathout, 2000) or (Neuvel & Fulop, 2002). The model we propose differs from these approaches in that the graphemic

similarities are determined solely on the basis of the sharing of graphemic features. This is the main contribution of this paper.

This model is also related to approaches that combine graphemic and semantic cues in order to identify morphemes or morphological relations between words. Usually, this semantic information is automatically acquired from corpora by means of various techniques such as latent semantic analysis (Schone & Jurafsky, 2000), mutual information (Baroni et al., 2002) or co-occurrence in n-word windows (Xu & Croft, 1998; Zweigenbaum & Grabar, 2003). In the experiment presented here, semantic information is extracted from a machine readable dictionary and semantic similarity is calculated through random walks in a lexical graph. The approach presented here can also be compared with (Hathout, 2002, 2003), where morphological knowledge is acquired by using semantic information extracted from dictionaries of synonyms and from WordNet.

2. Morphological relatedness

We assume here a minimalist definition of morphological relatedness: two words are morphologically related if they share phonological and semantic properties. In the experiment, graphemic properties have been used instead of phonological ones because the TLFi does not provide the pronunciation of all the headwords. The morphological relatedness is estimated by means of a bipartite graph like the one presented in figure 2, with one subset of vertices representing lexemes and the other representing the formal and the semantic features of these lexemes. Lexeme vertices are identified by the lemma and the grammatical category.

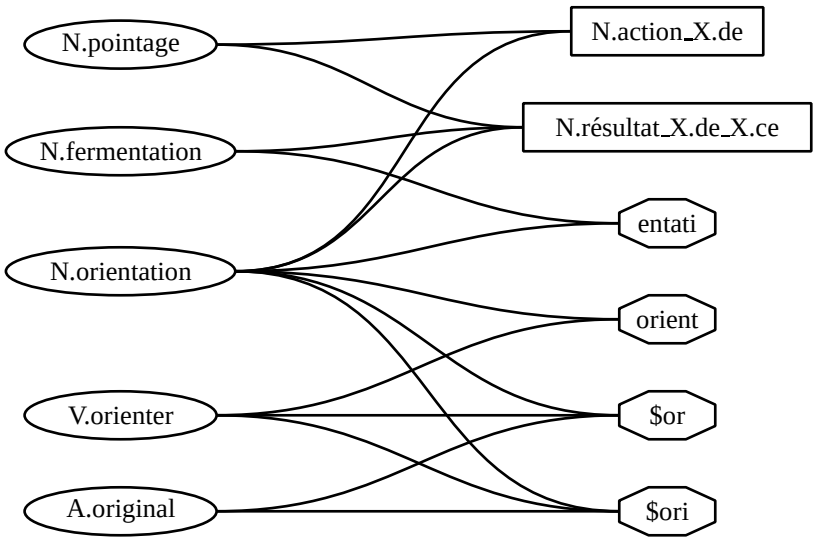

Figure 2: Excerpt of the bipartite graph which represents the lexicon. Words are displayed in ovals, semantic features in rectangles and formal features in octagons. The graph is symmetric.

2.1. Formal and semantic features

The formal properties associated with a lexeme are the n-grams of letters that occur in its lemma. The beginning and the end of the lemma are marked by the character $. We impose a minimum size on the n-grams ($n \geq 3$). For instance, the formal features associated with the French noun *orientation* are the n-grams of figure 3, with n ranging from 13 down to 3.

Figure 3 shows that the set of features associated with a given word is quite redundant. An interesting property of this description is that it does not confer a special status to any of the individual n-grams which characterize the lexemes. All n-grams play the same role and therefore none has the status of morpheme. These features are only used to bring together the words that share the same sounds.

Alternatively, one could have used the n-grams that occur in the inflected forms of the lexemes as formal features. Such an extended characterization is more faithful to word-based morphology and

```
$orientation$
$orientation orientation$
$orientatio orientation rientation$
$orientati orientatio rientation ientation$
...
$ori orie rien ient enta ntat tati atio tion ion$
$or ori rie ien ent nta tat ati tio ion on$
```

Figure 3: Excerpt of the formal features associated with the noun *orientation*.

makes the inflectional allomorphies available at the derivational level. However, we did not retain this option because inflectional endings reduce the homogeneity of the formal representations. For instance, with a threshold $n \geq 3$, the verb *malaxer* 'knead' would become connected to all the words that contain `xie` (*an**xie**ux*, *le**xie***, *orthodo**xie***, etc.) because of its inflected form *mala**xie**z* (second person plural, imperfect indicative and present subjunctive). In order to avoid giving too much importance to these very specific features, it is necessary to weight the contribution of each inflected form with an estimation of its frequency, computed for instance from a large text corpus. A form like *malaxiez* is likely to be very rare or even missing from most corpora. In this way, the unwanted connections will be demoted or eliminated.

The semantic features associated with a lexeme are the n-grams of words that occur in its definitions. The n-grams that contain punctuation marks, not counting apostrophes, are eliminated. In other words, we only use n-grams of words that occur between two punctuation marks. The words in the definitions are POS tagged and lemmatized. The tags are A for adjectives, N for nouns, R for adverbs, V for verbs and X for all other categories. For instance, the semantic features induced by the definition *Action d'orienter, de s'orienter ; résultat de cette action* 'act of directing, of finding one's way; result of this action' of the noun *orientation* are presented in figure 4. Notice that the semantic features are heavily redundant, just as the formal features are.

```
N.action_X.de_V.orienter N.action_X.de X.de_V.orienter
N.action X.de V.orienter X.de_V.s'orienter V.s'orienter
N.résultat_X.de_X.ce_N.action N.résultat_X.de_X.ce X.de_X.ce_N.action
N.résultat_X.de X.de_X.ce X.ce_N.action N.résultat X.ce
```

Figure 4: Semantic features induced by the definition *Action d'orienter, de s'orienter ; résultat de cette action* of the noun *orientation*.

This is a very coarse semantic representation inspired from the repeated segments (Lebart et al., 1998). It offers several advantages:

1. being heavily redundant, it can capture various levels of similarity between the definitions;
2. it integrates information of a syntagmatic nature without a deep syntactic analysis of the definitions;
3. it slightly reduces the strong variations in the lexicographical treatment of the headwords, especially in the division into sub-senses and in the definitions.

2.2. Connecting the lexemes through their features

The semantic and formal features are used in the same graph. The bipartite graph is built up by connecting each headword to its semantic and formal features symmetrically. For instance, the noun *orientation* is connected with the formal features `$or`, `$ori`, `$orie`, `$orien`, etc. which are in turn connected with the words *orienter*, *orientable*, *orientement* 'orientation', *orienteur* 'orientator', etc. Likewise, *orientation* is connected with the semantic features `N.action X.de`, `N.résultat X.de X.ce N.action`, etc. which are themselves connected with the nouns *orientement*, *harmonisation*, *pointage* 'checking', etc. The general schema is illustrated in figure 2. It shows that the semantic

and formal properties are used in the same manner. This representation corresponds precisely to the Network Model of Bybee (1988, 1995).

Actually, the bipartite structure is not essential. All we need is to be able to compute a morphological distance between the words. We use a bipartite graph mainly because it allows us to spread an activation simultaneously into the formal and the semantic subparts of the graph. The graph is also interesting because it contains representations of properties that are useful for morphological studies. They could for instance be used to describe the semantics of the *-able* suffixation or to find the characteristic endings of boat names in French (*voilier, pétrolier, bananier, thonier, sardinier*…; *patrouilleur, torpilleur, caboteur, dériveur, dragueur*…).

2.3. Estimating the morphological similarity between words

The morphological similarity between a word and its neighbors is estimated by simulating the spreading of an activation initiated at the vertex that represents that word. Since the graph is bipartite, the activation has to be propagated an even number of times. The graph being heavily redundant, two steps of propagation are sufficient to obtain the intended proximity estimations.

For instance, if we want to determine what the closest neighbors of *orientation* are, we initiate an activation at the vertex that represents *orientation*. Then, this activation is uniformly spread toward the formal and semantic features of *orientation*. In the next step, the activation located on the feature vertices is spread toward the lexeme vertices. The greater the number of features shared by a lexeme with *orientation* and the more specific these features are, the stronger the activation it receives. The assumption is that the strength of the activation is an estimation of the degree of morphological relatedness.

Technically, the spreading is simulated as a random walk in the graph (Gaume et al., 2002, 2005; Muller et al., 2006). It is classically computed as a multiplication of the stochastic adjacency matrix of the graph. More precisely, let $G = (V, E)$ be a graph consisting of a set of vertices $V = \{v_1, \ldots, v_n\}$ and a set of edges $E \subset V \times V$. Let A be the adjacency matrix of G, that is a $n \times n$ matrix such that $A_{ij} = 1$ if $(v_i, v_j) \in E$ and $A_{ij} = 0$ if $(v_i, v_j) \notin E$. We normalize the rows of A in order to get a stochastic matrix M:

$$\forall i \in [1, n], \forall j \in [1, n], M_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik}}$$

Then $(M^n)_{ij}$ is the probability of reaching vertex v_j from the vertex v_i through a walk of n steps. This probability can also be regarded as an activation level of node v_j following an n-step spreading initiated at node v_i.

In the experiment presented in this paper, one half of the activation is spread toward the semantic features and the other half toward the formal features. The edges of the bipartite graph can be divided into three parts $E = J \cup K \cup L$ where J contains the edges that connect a headword to a formal feature, K the edges that connect a headword to a semantic feature and L the edges that connect a formal or semantic feature to a headword. The actual values of M are defined as follows:

$$\text{if } e_{ij} = (v_i, v_j) \in J, \quad M_{ij} = \begin{cases} 0.5 \dfrac{A_{ij}}{\sum_{e_{ih} \in J} A_{ih}} & \text{if } v_i \text{ is connected to a semantic feature} \\ \dfrac{A_{ij}}{\sum_{e_{ik} \in J} A_{ik}} & \text{otherwise} \end{cases}$$

$$\text{if } e_{ik} = (v_i, v_k) \in K, \quad M_{ik} = \begin{cases} 0.5 \dfrac{A_{ik}}{\sum_{e_{ih} \in K} A_{ih}} & \text{if } v_i \text{ is connected to a formal feature} \\ \dfrac{A_{ik}}{\sum_{e_{ih} \in K} A_{ih}} & \text{otherwise} \end{cases}$$

$$\text{if } e_{il} = (v_i, v_l) \in L, \quad M_{il} = \frac{A_{il}}{\sum_{e_{ih} \in L} A_{ih}}$$

2.4. Morphological neighbors

The graph used in the experiment was built from the headwords and the definitions of the TLFi. We only removed the definitions of non standard uses (old, slang, regionalism, etc.). The extraction and cleaning-up of the definitions were carried out in collaboration with Bruno Gaume and Philippe Muller. The bipartite graph was created from 225 529 definitions describing 75 024 headwords (lexemes). They induced about 9 million features, 90% of them being associated with only one headword. These features were removed because they do not contribute to the connections of different headwords. Table 1 shows that this reduction is stronger for the semantic features (93%) than it is for the formal ones (69%). Indeed, semantic descriptions show greater variability than formal ones.

features	complete	reduced	hapax
formal	1 306 497	400 915	69%
semantic	7 650 490	548 641	93%
total	8 956 987	949 556	90%

Table 1: Numbers of semantic and formal features.

The use of the graph is illustrated in figure 5. It shows the 40 nearest neighbors of the verb *fructifier* 'bear fruit' for three propagation configurations. The first row (*form*) presents the neighbors of *fructifier* in a graph that only contains formal features. It shows that the members of the morphological family tend to appear as the closest neighbors and that the members of the derivational series (i.e. the verbs ending in *-ifier*) are more distant. The members in the second row (*sem*) have been computed in a graph that only contains the semantic features and the ones in the third row (*form + sem*) in the full graph.

form	**V.fructifier N.fructification A.fructificateur A.fructifiant A.fructifère V.sanctifier V.rectifier** *A.rectifier V.fructidoriser N.fructidorien N.fructidor* **N.fructuosité R.fructueusement A.fructueux** *N.rectifieur A.obstructif A.instructif A.destructif A.constructif* **N.infructuosité R.infructueusement A.infructueux V.transsubstantifier V.substantifier V.stratifier V.schistifier V.savantifier V.refortifier V.ratifier V.quantifier V.présentifier V.pontifier V.plastifier V.notifier V.nettifier V.mystifier V.mortifier V.justifier V.idiotifier V.identifier**
sem	**V.fructifier** *V.trouver N.missionnaire N.mission A.missionnaire N.saisie N.police N.hangar N.dîme N.ban V.affruiter N.melon N.saisonnement N.azédarach A.fruitier A.bifère V.saisonner N.roman N.troubadour V.contaminer N.conductibilité N.alevinage V.profiter* **A.fructifiant** *N.pouvoir V.agir N.opération V.placer N.rentabilité N.jouissance N.avocat N.report* **A.fructueux** *V.tourner V.chiper N.économat N.visa N.société N.réserve N.récréance*
form + sem	**V.fructifier A.fructifiant N.fructification A.fructificateur** *V.trouver* **A.fructifère V.rectifier V.sanctifier** *A.rectifier V.fructidoriser N.fructidor N.fructidorien N.missionnaire N.mission A.missionnaire* **A.fructueux R.fructueusement N.fructuosité** *N.rectifieur N.saisie N.police N.hangar N.dîme N.ban A.fruitier V.affruiter A.instructif A.obstructif A.destructif A.constructif N.conductibilité V.saisonner N.melon N.saisonnement N.azédarach A.bifère V.contaminer N.roman N.troubadour N.alevinage*

Figure 5: The 40 nearest neighbors of the verb *fructifier* when the activation is spread only toward the formal features in the first row, only toward the semantic ones in the second row and toward both the semantic and formal features in the third. Words that belong to the family or series of *fructifier* are in boldface; the others are in italic.

The first two rows show clearly that formal features are the more predictive ones while semantic features are the less reliable ones. These examples provide an insight into some of the features of the morpholog-

ical families and the derivational series that could be used in order to separate them: families are small sets; series are larger sets; families have a strong semantic and formal cohesion; members of a series have looser semantic and formal connections. The last two features explain why the members of the morphological families tend to show up before the members of the derivational series. The examples also show that the morphological similarity is not selective enough and that the list of neighbors cannot be used as is. We need to further filter them and we propose to do so with formal analogy.

3. Analogy

3.1. Familial and serial analogies

The members of the series and families are massively involved in analogies which structure the lexicon. For instance, *fructifier* and *fructification* which belong to the same family form analogies with large numbers of pairs of members of other families (*rectifier* 'correct', *rectification*), (*certifier* 'assure', *certification* 'attestation'), (*plastifier* 'coat with plastic', *plastification* 'lamination of document'), (*sanctifier*, *sanctification*), (*vitrifier*, *vitrification*), etc. Besides, the first elements of each of these pairs belong to the series of *fructifier* and the second ones to the series of *fructification*. In a dual manner, *fructifier* and *sanctifier* which belong to the same series form analogies with the members of other series (*fructificateur* 'fructifier', *sanctificateur* 'sanctifier'), (*fructification*, *sanctification*) or (*fructifiant* 'fructifying', *sanctifiant* 'sanctifying'). These pairs are respectively made of members of the families of *fructifier* and *sanctifier*.

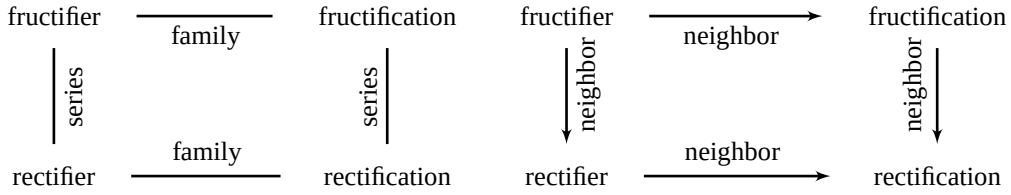

Figure 6: Morphological relations and neighborhood relations between the members of the *fructifier*:*fructification* :: *rectifier*:*rectification* analogy.

Formal analogies can be used in order to filter the morphological neighbors of a word. Actually, we are interested in analogies such as *fructifier*:*fructification* :: *rectifier*:*rectification*. Since *fructification* belongs to the family of *fructifier* and *rectifier* to its series, both are morphological neighbors of *fructifier*. Similarly, *rectification* belongs to the series of *fructification* and to the family of *rectifier*. Therefore, it is a morphological neighbor of both *fructification* and *rectifier*. These relations are illustrated in figure 6. Conversely, if we consider that the morphological neighbors of a word are likely to be morphologically related to that word, then we can use them to look for quadruplets that could form analogies. These quadruplets could be found as follows:

For a given word a,
  look for two of its neighbors b and c, then
  for every d that is a neighbor of both b and c,
    the quadruplet $a : b :: c : d$ is likely to be an analogy.

More generally, if b is a correct morphological neighbor of a, then it is either a member of the family of a or a member of its series. Therefore, there exists another neighbor c of a (c belongs to the family of a if b belongs to the series of a or vice versa) such that there exists a neighbor d of b and of c such that $a : b :: c : d$. We then have only two configurations:

1. if $b \in F_a$, then $\exists c \in S_a, \exists d \in S_b \cap F_c, a : b :: c : d$
2. if $b \in S_a$, then $\exists c \in F_a, \exists d \in F_b \cap S_c, a : b :: c : d$

where F_x is the morphological family of x and S_x the derivational series of x.

3.2. Formal analogy

A formal or graphemic analogy is a relation $a : b :: c : d$ that holds between four strings such that the graphemic differences between a and b are the same as the ones between c and d. This is the case for *fructifier*:*fructification* :: *rectifier*:*rectification* (see figure 7). Naturally, more than one difference can appear in the pair as with the four Arabic words *kataba*:*maktoubon* :: *fa3ala*:*maf3oulon* which respectively are transcriptions of the verb 'write', the noun 'document', the verb 'do' and the noun 'effect.'[3] The differences between the first two words and between the two last ones can be described as in figure 7. They are identical for the two pairs of words. This example shows that even analogies in a templatic language like Arabic can be checked in this way.

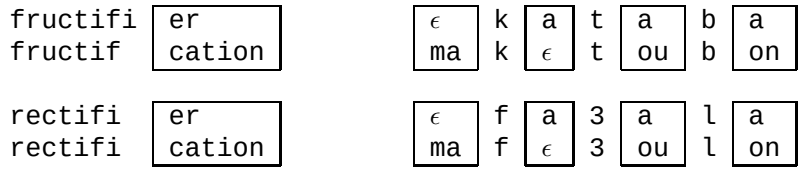

Figure 7: Formal analogies *fructifier*:*fructification* :: *rectifier*:*rectification* and *kataba*:*maktoubon* :: *fa3ala*:*maf3oulon*. The differences are located in the frame boxes. ϵ represents the empty string.

More generally, formal analogies can be defined in terms of factorization (Stroppa & Yvon, 2005). Let L be an alphabet and $a \in L^\star$ a string over L. A factorization of a is a sequence $f = (f_1, \cdots, f_n) \in L^{\star n}$ such that $a = f_1 \oplus \cdots \oplus f_n$ where $\oplus$ denotes concatenation. For instance, (ma, k, ϵ, t, ou, b, on) is a factorization of length 7 of maktoubon. Morphological analogies can be defined as follows. Let $(a, b, c, d) \in L^{\star 4}$ be four strings. $a : b :: c : d$ is a formal analogy iff there exists $n \in \mathbb{N}$ and four factorizations of length n of the four strings $(f(a), f(b), f(c), f(d)) \in (L^{\star n})^4$ such that, $\forall i \in [1, n], (f_i(b), f_i(c)) \in \{(f_i(a), f_i(d)), (f_i(d), f_i(a))\}$. For the analogy *kataba*:*maktoubon* :: *fa3ala*:*maf3oulon*, the property holds for $n = 7$ (see figure 7).

3.3. Implementation

Formal analogies are checked at the graphemic level. The differences between the first and second pairs of strings are calculated from the sequence of string edit operations that transform the first form of each pair into the second one. Both sequences must minimize Levenshtein edit distance (i.e. have the least cost). Each sequence corresponds to a path in the edit lattices of the pair of words. The lattices are represented by a matrix computed using the standard string edit algorithm (Jurafsky & Martin, 2000). The path which describes the sequence of string edit operations starts at the last cell of the matrix and climbs to the first one. It is made up as follows: for each cell, select the neighboring one with the least cost ; in case of equal costs, prefer the cell to the left (insertion), then the one upward (deletion) and otherwise the one in the upper left diagonal direction (substitution). Figure 8 presents the path that is selected in the string edit matrix of *fructueux* 'fruitful' and *infructueusement* 'fruitlessly' and figure 9, the sequence of edit operations for this pair.

Sequences of edit operations can be simplified by merging the series of identical character matchings. The sequence in figure 9 then becomes (2). This simplified sequence is identical to the one for the pair *soucieux*:*insoucieusement* 'worried':'unworriedly' except for the matching operation (3).

(2) ((I,ϵ,i), (I,ϵ,n), (M,fructueu,fructueu), (S,x,s), (I,ϵ,e), (I,ϵ,m), (I,ϵ,e), (I,ϵ,n), (I,ϵ,t))

(3) ((I,ϵ,i), (I,ϵ,n), (M,soucieu,soucieu), (S,x,s), (I,ϵ,e), (I,ϵ,m), (I,ϵ,e), (I,ϵ,n), (I,ϵ,t))

The two sequences can be made identical if the matching sub-strings are not specified (i.e. replaced by a wildcard character @). The resulting sequence can then be assigned to both pairs as their edit signatures (σ). The formal analogy *fructueux*:*infructueusement* :: *soucieux*:*insoucieusement* can be stated in terms of identity of the edit signatures of the two pairs (4).

[3] This example is adapted from examples in Lepage (1998, 2003).

	◇	i	n	f	r	u	c	t	u	e	u	s	e	m	e	n	t
◇	0	← 1	← 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
f	1	1	2	↖ 2	3	4	5	6	7	8	9	10	11	12	13	14	15
r	2	2	2	3	↖ 2	3	4	5	6	7	8	9	10	11	12	13	14
u	3	3	3	3	3	↖ 2	3	4	5	6	7	8	9	10	11	12	13
c	4	4	4	4	4	3	↖ 2	3	4	5	6	7	8	9	10	11	12
t	5	5	5	5	5	4	3	↖ 2	3	4	5	6	7	8	9	10	11
u	6	6	6	6	6	5	4	3	↖ 2	3	4	5	6	7	8	9	10
e	7	7	7	7	7	6	5	4	3	↖ 2	3	4	5	6	7	8	9
u	8	8	8	8	8	7	6	5	4	3	↖ 2	3	4	5	6	7	8
x	9	9	9	9	9	8	7	6	5	4	3	↖ 3	← 4	← 5	← 6	← 7	← 8

Figure 8: Least cost path describing a sequence of string edit operations that transforms *fructueux* into *infructueusement*. ◇ represents the beginning of the string. Cell (i, j) in the matrix indicates the Levenshtein distance between the substring consisting of the first i characters of *fructueux* and the one consisting of the first j characters of *infructueusement*.

I	I	M	M	M	M	M	M	M	M	S	I	I	I	I	I
ϵ	ϵ	f	r	u	c	t	u	e	u	x	ϵ	ϵ	ϵ	ϵ	ϵ
i	n	f	r	u	c	t	u	e	u	s	e	m	e	n	t

Figure 9: Sequence of edit operations that transform *fructueux* into *infructueusement*. The types of the operations are indicated on the first row: D for deletion, I for insertion, M for matching and S for a substitution by a different character.

(4) $\sigma(\texttt{fructueux}, \texttt{infructueusement}) =$
$\sigma(\texttt{soucieux}, \texttt{insoucieusement}) =$
((I,ϵ,i), (I,ϵ,n), (M,@,@), (S,x,s), (I,ϵ,e), (I,ϵ,m), (I,ϵ,e), (I,ϵ,n), (I,ϵ,t))

More generally, four strings $(a, b, c, d) \in L^{\star 4}$ form a formal analogy $a : b :: c : d$ iff $\sigma(a, b) = \sigma(c, d)$.

4. First results

This is work in progress and we only have preliminary results. We have computed the 100 nearest neighbors of the headwords of the TLFi, then collected the formal analogies for 22 headwords belonging to 4 morphological families and checked them manually. An analogy $a : b :: c : d$ is accepted as correct if:

- b belongs to the family of a, c belongs to the series of a, d belongs to series of b and to the family of c, or
- b belongs to the series of a, c belongs to the family of a, d belongs to family of b and to the series of c.

We present some examples of correct analogies in (5) and erroneous ones in (6). We can see that the collected analogies involve words that are derived one from the other (5a), words that are derived from a common base (5b) and words connected through a sequence of derivations (5c).

(5) a. N.fructification:N.identification :: V.fructifier:V.identifier

b. A.fructifiant:A.fructificateur :: A.glorifiant :: A.glorificateur

c. A.fructueux:A.affectueux :: N.infructuosité:N.inaffectuosité

d. A.frugivore:A.végétivore :: R.frugalement:R.végétalement

e. A.fruitarien:A.végétarien :: N.fruitarisme:N.végétarisme

f. A.fruitier:A.laitier :: N.fruiterie:N.laiterie

g. R.fructueusement:R.affectueusement :: N.fructuosité:N.affectuosité

(6) a. A.fruité:N.fruste :: A.truité:N.truste

b. N.fruit:N.frumentaire :: A.instruit:A.instrumentaire

c. N.fruiterie:N.friterie :: V.effruiter:V.effriter

We have tested three configurations (see § 2.4). In the first, we have used neighbors from the graph that contains the formal features only, in the second, the semantic features only, and in the third, both the formal and the semantic features. The results are summarized in table 2. Their quality is quite satisfactory. We observe that the number of analogies depends on the configuration of propagation. The use of the semantic features improves the precision but reduces the total number of analogies that are collected. The best trade-off is a simultaneous propagation toward the semantic and the formal features.

configuration	analogies	correct	errors
formal	169	163	3.6%
semantics	5	5	0.0%
sem + form	130	128	1.5%

Table 2: Number of the analogies collected for a sample of 22 headwords and error rate.

The performance of the method strongly depends on the length of the headwords because the method mainly relies on formal similarity and because formal similarity is stronger for long words. Table 3 show this correlation clearly. It presents the number of analogies and the error rate of 13 samples of 5 words, selected randomly. The analogies have been collected from neighborhoods in the full graph. The words in each group are of the same length. Lengths range from 4 to 16 letters. We can see that the analogies collected for words of 10 letters or more are all correct.

length	analogies	correct	errors
4	29	14	51.7%
5	22	14	36.4%
6	8	7	12.5%
7	10	8	20.0%
8	55	54	1.8%
9	29	27	6.9%
10	30	30	0.0%
11	32	32	0.0%
12	19	19	0.0%
13	11	11	0.0%
14	35	35	0.0%
15	63	63	0.0%
16	39	39	0.0%

Table 3: Numbers of analogies and error rates for headwords of length 4 to 16.

5. Conclusion and directions for further research

We have presented a computational model that makes the morphological structure of the lexicon emerge from the formal and semantic properties of the words it contains. The model is radically word-based. It integrates the semantic and formal properties of the words in a uniform manner and represents them in a bipartite graph. Random walks are used to simulate the spreading of activations in the lexical network. The level of activation obtained after the propagation indicates the lexical relatedness of the words. The members of the morphological family and the derivational series of a word are then identified among its lexical neighbors by means of formal analogies.

Let us stress that this method is promising because it is mainly computational. Almost no theoretical assumptions have been made. The method primarily exploits the memory and the computing power of the processors. Another interesting feature is that the formal and semantic properties of the words are represented separately. Therefore, the method deals with so-called parasynthetic derivatives like any other lexemes.

The next steps of this research are to create an initial network with only long words and then use a bootstrap method. One important task that remains to be done is to separate the members of the families from the ones of the series. We also intend to conduct a similar experiment on the English lexicon and to evaluate our results in a more classical manner by using the CELEX database (Baayen et al., 1995) as gold standard. The evaluation should also be done with respect to well-known systems like *Linguistica* (Goldsmith, 2001) or the morphological analyzer of Bernhard (2006).

Acknowledgments

I would like to thank the ATILF laboratory and Jean-Marie Pierrel for making available to me the TLFi. I am in debt to Bruno Gaume and Philippe Muller for the many discussions and exchanges we have had on the cleaning-up of the TFLi and its exploitation through random walks. I am also grateful to Gilles Boyé, Olivier Haute-Cœur, Ludovic Tanguy, Jesse Tseng and audiences at University of Toulouse, University of Bordeaux, the Textgraphs-3 Workshop and the Decembrettes-6 Conference for their comments, suggestions and feedback on earlier versions of this paper. All errors are mine.

References

Anderson, Stephen R. (1992). *A-Morphous Morphology*. Cambridge University Press, Cambridge.

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. MIT Press, Cambridge, Mass.

Aronoff, Mark (1994). *Morphology by Itself. Stem and Inflectional Classes*. MIT Press, Cambridge, Mass.

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers (1995). The CELEX lexical database (release 2). Linguistic Data Consortium, Penn.

Baroni, Marco, Johannes Matiasek & Harald Trost (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-2002*, ACL, Philadelphia, Penn., 48–57.

Bernhard, Delphine (2006). Automatic acquisition of semantic relationships from morphological relatedness. *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, Springer, vol. 4139 of *Lecture Notes in Computer Science*, 121–132.

Blevins, James P. (2006). Word-based morphology. *Journal of Linguistics* 42:3, p. 531–573.

Van den Bosch, Antal & Walter Daelemans (1999). Memory-based morphological analysis. *Proceedings of the 37th Annual Meeting of the Association of Conputational Linguistics*, University of Maryland, ACL.

Bourigault, Didier (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Habilitation à diriger des recherches, Université Toulouse-Le Mirail, Toulouse.

Burzio, Luigi (2002). Surface-to-surface morphology: when your representations turn into constraints. Boucher, Paul (ed.), *Many Morphologies*, Cascadilla, Somerville, Mass., p. 142–177.

Bybee, Joan L. (1988). Morphology as lexical organization. Hammond, Micheal & Michael Noonan (eds.), *Theoretical Morphology. Approaches in Modern Linguistics*, Academic Press, San Diego, CA, chap. 7, 119–141.

Bybee, Joan L. (1995). Regular morphology and the lexicon. *Language and cognitive processes* 10:5, 425–455.

Creutz, Mathias & Krista Lagus (2002). Unsupervised discovery of morphemes. *Proceedings of the ACL Workshop on Morphological and Phonological Learning*, ACL, Philadelphia, Penn., 21–30.

Dal, Georgette, Nabil Hathout & Fiammetta Namer (1999). Construire un lexique dérivationnel : théorie et réalisation. Amsili, Pascal (ed.), *Actes de la VI^{e} conférence sur le Traitement Automatique des Langues Naturelles (TALN'99)*, ATALA, Cargèse, 115–124.

Déjean, Hervé (1998). Morphemes as necessary concept for structures discovery from untagged corpora. *Proceedings of the Workshop on Paradigms and Grounding in Natural Language Learning*, Adelaide, Australia, 295–299.

Gaume, Bruno, Karine Duvigneau, Olivier Gasquet & Marie-Dominique Gineste (2002). Forms of meaning, meaning of forms. *Journal of Experimental and Theoretical Artificial Intelligence* 14:1, 61–74.

Gaume, Bruno, Fabienne Venant & Bernard Victorri (2005). Hierarchy in lexical organization of natural language. Pumain, D. (ed.), *Hierarchy in natural and social sciences*, Methodos series, Kluwer, 121–143.

Gaussier, Éric (1999). Unsupervised learning of derivational morphology from inflectional lexicons. *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, ACL, University of Maryland.

Goldsmith, John (2001). Unsupervised learning of the morphology of natural language. *Computational Linguistics* 27:2, 153–198.

Hathout, Nabil (2000). Morphological pairing based on the network model. *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, Pyrgos, 35–38.

Hathout, Nabil (2002). From WordNet to CELEX: acquiring morphological links from dictionaries of synonyms. *Proceedings of the Third International Conference on Language Resources and Evaluation*, ELRA, Las Palmas de Gran Canaria, 1478–1484.

Hathout, Nabil (2003). L'analogie, un moyen de croiser les contraintes et les paradigmes. Acquisition de connaissances morphologiques à partir de dictionnaires de synonymes. *Revue d'Intelligence Artificielle* 17:5-6, 923–934.

Hathout, Nabil (2005). Exploiter la structure analogique du lexique construit : une approche computationnelle. *Cahiers de lexicologie* 87:2, 5–28.

Hathout, Nabil (2008). Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. *Proceedings of the Coling workshop Textgraphs-3*, ACL, Manchester, 1–8.

Hathout, Nabil, Fiammetta Namer & Georgette Dal (2002). An Experimental Constructional Database: The MorTAL Project. Boucher, Paul (ed.), *Many Morphologies*, Cascadilla, Somerville, Mass., 178–209.

Jackendoff, Ray (1975). Morphological and semantic regularities in the lexicon. *Language* 51:3, 639–671.

Jing, Hongyan & Evelyne Tzoukerman (1999). Information retrieval based on context distance and morphology. *Proceedings of 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, ACM, Berkeley, CA., 90–96.

Jurafsky, Daniel & James H. Martin (2000). *Speech and language processing*. Prentice-Hall.

Langlais, Philippe & Alexandre Patry (2007). Translating unknown words by analogical learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning, EMNLP-CoNLL 2007*, ACL, Prague, Czech Republic, 877–886.

Lebart, Ludovic, André Salem & Lisette Berry (1998). *Exploring textual data*. Kluwer Academic Publishers, Dordrecht.

Lepage, Yves (1998). Solving analogies on words: an algorithm. *Proceedings of COLING-ACL'98*, Montréal, vol. 2, 728–735.

Lepage, Yves (2003). *De l'analogie rendant compte de la commutation en linguistique*. Habilitation à diriger des recherches, Université Joseph Fourier, Grenoble.

Miller, Georges A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3:4, 335–391.

Moreau, Fabienne, Vincent Claveau & Pascale Sébillot (2007). Automatic morphological query expansion using analogy-based machine learning. *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, Rome, 222–233.

Muller, Philippe, Nabil Hathout & Bruno Gaume (2006). Synonym extraction using a semantic distance on a dictionary. Radev, Dragomir & Rada Mihalcea (eds.), *Proceedings of the HLT/NAACL Workshop Textgraphs*, Association for Computational Linguistics, New York, NY, 65–72.

Namer, Fiammetta (2005). *La Morphologie Constructionnelle du Français et les Propriétés Sémantiques du Lexique : Traitement Automatique et Modélisation*. Habilitation à diriger des recherches, Université de Nancy 2.

Neuvel, Sylvain & Sean A. Fulop (2002). Unsupervised learning of morphology without morphemes. *Proceedings of the ACL Workshop on Morphological and Phonological Learning 2002*, ACL, Philadelphia, Penn.

Neuvel, Sylvain & Rajendra Singh (2001). Vive la différence ! What morphology is about. *Folia linguistica* 35:3-4, 313–320.

Pirrelli, Vito & François Yvon (1999). The hidden dimension: a paradigmatic view of data-driven NLP. *Journal of Experimental & Theoretical Artificial Intelligence* 11:3, 391–408.

Plénat, Marc (2005). *Rosinette, cousinette, putinette, starlinette*. décalage, infixation et épenthèse devant *-ette*. Choï-Jonin, Injoo, Myriam Bras, Anne Dagnac & Magali Rouquier (eds.), *Questions de classification en linguistique : méthodes et descriptions. Mélanges offerts au Professeur Christian Molinier*, Peter Lang, Berne, 275–298.

Plénat, Marc & Michel Roché (2003). Prosodic constraints on suffixation in French. Booij, Geert, Janet DeCesaris, Angela Ralli & Sergio Scalise (eds.), *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting*, Universitat Pompeu Fabra, Barcelona, 285–299.

Schone, Patrick & Daniel S. Jurafsky (2000). Knowledge-free induction of morphology using latent semantic analysis. *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, Lisbon, 67–72.

Skousen, Royal (1989). *Analogical Modeling of Language*. Kluwer Academic Publishers, Dordrecht.

Stroppa, Nicolas & François Yvon (2005). An analogical learner for morphological analysis. *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, ACL, Ann Arbor, Michigan, 120–127.

Xu, Jinxi & W. Bruce Croft (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems* 16:1, 61–81.

Yarowsky, David & Richard Wicentowski (2000). Minimally supervised morphological analysis by multimodal alignment. *Proceedings of the Association of Computational Linguistics (ACL-2000)*, Hong Kong, 207–216.

Zweigenbaum, Pierre & Natalia Grabar (2003). Learning derived words from medical corpora. *9th Conference on Artificial Intelligence in Medicine Europe*, Cyprus, 189–198.