Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy

Nabil Hathout
Université de Toulouse
Nabil.Hathout@univ-tlse2.fr

Abstract

The paper presents a computational model aiming at making the morphological structure of the lexicon emerge from the formal and semantic regularities of the words it contains. The model is purely lexeme-based. The proposed morphological structure consists of (1) binary relations that connect each headword with words that are morphologically related, and especially with the members of its morphological family and its derivational series, and of (2) the analogies that hold between the words. The model has been tested on the lexicon of French using the TLFi machine readable dictionary.

1 Lexeme-based morphology

Morphology is traditionally considered to be the field of linguistics that studies the structure of words. In this conception, words are made of morphemes which combine according to rules of inflexion, derivation and composition. If the morpheme-based theoretical framework is both elegant and easy to implement, it suffers many drawbacks pointed out by several authors (Anderson, 1992; Aronoff, 1994). The alternative theoretical models that have been proposed falls within lexeme-based or word-based morphology in which the minimal units are words instead of morphemes. Words then do not have any structure at all and morphology becomes a level of organization of the lexicon based on the sharing of semantic and formal properties.

The morpheme-based / lexeme-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; Goldsmith, 2001; Creutz and Lagus, 2002; Bernhard, 2006). In a lexeme-based approach, it is to discover the relations between the word and the other lexical items. These relations serve to identify the morphological family of the word, its derivational series, and the analogies in which it is involved. For instance, the analysis of the French word *dérivation* may be considered as satisfactory if it connects *dérivation* with enough members of its family (*dériver* 'derivate', *dérivationnel* 'derivational', *dérivable*, *dérive* 'drift', *dériveur* 'sailing dinghy', etc.) and of its derivational series (*formation* 'education', *séduction*, *variation*, *émission*, etc.). Each of these relations is integrated into a large collection of analogies that characterizes it semantically and formally. For instance, the relation between *dérivation* and *dérivable* is part of a series of analogies which includes *dérivation*:*dérivable*::*variation*:*variable*, *dérivation*:*dérivable*::*modification*:*modifiable*, etc. Similarly, *dérivation* and *variation* participates in a series of analogies such as *dérivation*:*variation*::*dériver*:*varier*, *dérivation*:*variation*::*dérivationnel*:*variationnel*, *dérivation*:*variation*::*dérivable*:*variable*.

2 Computational modeling

The paper describes a computational model aiming at making 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 TLFi machine readable dictio-

Coling 2008: Proceedings of 3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing, pages 1–8
Manchester, August 2008

nary.[1] The main novelty of the paper is the combination of lexical proximity with formal analogy. We first use lexical similarity in order to select a set of words that are likely to be morphologically related to each other. Then, these candidates are checked by means of analogy.

The two techniques are complementary. The first one brings closer the words that are morphologically close and especially the ones that are members of the same morphological families and the same derivational series. It is able to deal with large number of words, but it is too coarse-grained to discriminate the words that are actually morphological related from the ones that are not. The second technique, formal analogy, is then used to perform a fine-grained filtering. Technically, our model joins:

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

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

More generally, our approach is original in that:

1. Our computational model is pure lexeme-based. The discovery of morphological relations between words do not involve the notions of morpheme, affix, morphological exponent, etc. nor any representation of these concepts.

2. The membership to the families and series is gradient. It accounts, for instance, for the fact that *dériveur* is morphologically and semantically closer to *dérive* than to *dérivationnellement*, 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 are shared, the closer the words are.

Besides, the model integrates semantic and formal informations in a uniform manner. All kind of semantic informations (lexicographic definitions, synonyms, synsets, etc.) and formal ones (graphemic, phonological, etc.) can be used. They can be cumulated easily in spite of the differences in nature and origin. The model takes advantage of the redundancy of the features and is fairly insensitive to variation and exceptions.

[1] *Trésor de la Langue Française* (http://atilf.atilf.fr/).

3 Related work

Many works in the field of computational morphology aim at the discovery of 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 and Wicentowski, 2000) and (Baroni et al., 2002), who use string edit distances to estimate formal similarity. As far as we know, all the other perform some sort of segmentation even when the goal is not to find morphemes as in (Neuvel and Fulop, 2002). Our model differs from these approaches in that the graphemic similarities are determined solely on the basis of the sharing of graphemic features. It is the main contribution of this paper.

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

4 Lexeme Description

In our model, the lexical units and their properties are represented in a bipartite graph with the vertices representing the lexemes in one sub-set and the vertices representing the formal and semantic features in the other. Lexeme vertices are identified by the lemma and the grammatical category.

In the experiment reported in the paper, the formal properties are the n-grams of letters that occur in the lexemes lemma. Figure 1 shows a sub-set of

```
$or; $ori; $orie; ...
$orientation; ori; orie; ...
orientation; orientation$; ...
tio; tion; tion$; ion; ion$; on$
```

Figure 1: Excerpt of the formal features associated with the noun *orientation*.

```
N.action; N.action X.de; N.action
X.de V.orienter; X.de; X.de
V.orienter; V.orienter; X.de
V.s'orienter; V.s'orienter;
N.résultat; N.résultat X.de;
N.résultat X.de X.ce; N.résultat
X.de X.ce N.action; X.de X.ce;
X.de X.ce N.action; X.ce; X.ce
N.action; N.action
```

Figure 2: Semantic features induced by the definition "Action d'orienter, de s'orienter ; résultat de cette action." of the noun *orientation*

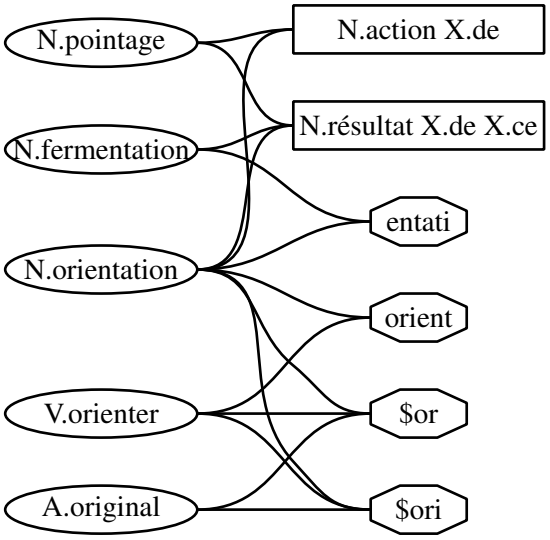

Figure 3: Excerpt of the bipartite graph which represents the lexicon. Words are displayed in ovals, semantic feature in rectangles and formal features in octagons. The graph is symmetric.

the formal features associated with the word *orientation*. 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$).

The model is pure lexeme-based because this decomposition 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 no one has the status of morpheme. These features are only used to bring closer the words that share the same sounds.

The semantic properties we have used are extracted from the TLFi definitions. Each headword is provided with the n-grams of words that occur in its definitions. The n-grams that contain punctuation marks are eliminated. In other words, we only use n-grams of words that occur between two punctuation marks. For instance, the semantic features induced by the definition *Action d'orienter, de s'orienter ; résultat de cette action.* ('act of orienting, of finding one's way; result of this action') of the noun *orientation* are presented in figure 2. 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.

This is a very coarse semantic representation inspired from the repeated segments (Lebart et al., 1998). It offers three advantages: (1) being heavily redundant, it can capture various levels of similarity between the definitions; (2) it integrates informations 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.

The bipartite graph is built up by symmetrically connecting each headword to its semantic and formal features. For instance, the noun *orientation* is connected with the formal feature `$or`, `$ori`, `$orie`, `$orien`, etc. which are in turn connected with the words *orienter*, *orientable*, *orientement* 'orientation', *orienteur* 'orientor', 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* 'synchronization', *pointage* 'checking', etc. The general schema is illustrated in figure 4. This representation corresponds precisely to the Network Model of Bybee (1995).

We use a bipartite graph mainly for two reasons: (1) We can spread an activation synchronously into the formal and the semantic sub-graphs. (2) It contains representations of the formal and the semantic properties of the lexemes which, for instance, could be used in order to describe the semantics of the *-able* suffixation or the characteristic endings of the boat names (*-ier*, *-eur*, etc.). However, the bipartite structure is not essential and we only need

to be able to compute morphological distances between words.

5 Random walks

The computational side of the method is based on the estimation of the proximity between words represented in a lexical graph (Gaume et al., 2002; Gaume et al., 2005; Muller et al., 2006). The graphs used in this approach are slightly different from the ones presented above. All their vertices represent words and the edges describe semantic relations such as synonymy. The proximity is computed by simulating the spreading into the graph of an activation initiated at a vertice. Following the spreading, the nodes which are most excited are regarded as being the closest to the initial vertice.

The same method can be used to estimate the morphological proximity between words that are described in a bipartite graph like the one we propose (see figure 4). It then connects words that have the same semantic and formal features. One has just to propagate the activation into the bipartite graph for an even number of times. When the graph is heavily redundant, two steps of propagation are sufficient to obtain the intended proximity estimations.

In the example in figure 4, the morphological neighbors of the noun *orientation* are identified by activating the vertice which represents it. In the first step, the activation is spread toward the vertices which represent its formal and semantic features. In the second step, the activation located on the feature vertices is spread toward the headword vertices. For instance, *orienter* becomes activated via the formal features `$or`, `$ori`, `orien` and *fermentation* through the formal feature `entati` and the semantic feature `N.résultat X.de X.ce`. The greater the number of features shared by a headword with *orientation*, the stronger the activation it receives.

The spreading of activation is simulated as a random walk in the lexical graph, classically computed as a multiplication of the stochastic adjacency matrix. More precisely, let $G = (V, E, w)$ be a weighted graph consisting of a set of vertices $V = \{v_1, \ldots, v_n\}$, a set of edges $E \subset V^2$ and of a weight function $w : E \rightarrow \mathbb{R}$. Let A be the adjacency matrix of G, that is a $n \times n$ matrix such that $A_{ij} = 0$ if $(v_i, v_j) \notin E$ and $A_{ij} = w(v_i, v_j)$ if $(v_i, v_j) \in E$. (In the experiment, $w(e) = 1, \forall e \in E$.) We normalize the rows of A in order to get a stochastic matrix M. M^n_{ij} is the probability of reaching node v_j from the node 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 vertice v_i.

In the experiment presented in this paper, the activation is spread for one half toward the semantic feature and for the other toward the formal features. The edges of the bipartite graph can be divided in 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 values of M are defined as follows:

- if $e_{ij} = (v_i, v_j) \in J$, $M_{ij} = \frac{A_{ij}}{2 \sum_{e_{ih} \in J} A_{ih}}$ if v_i is connected to a semantic feature and $M_{ij} = \frac{A_{ij}}{\sum_{e_{ik} \in J} A_{ik}}$ otherwise.
- if $e_{ik} = (v_i, v_k) \in K$, $M_{ik} = \frac{A_{ik}}{2 \sum_{e_{ih} \in K} A_{ih}}$ if v_i is connected to a formal feature and $M_{ik} = \frac{A_{ik}}{\sum_{e_{ih} \in K} A_{ih}}$ otherwise.
- if $e_{il} = (v_i, v_l) \in L$, $M_{il} = \frac{A_{il}}{\sum_{e_{ih} \in L} A_{ih}}$.

6 Lexical neighborhood

The graph used in the experiment has been built from the definitions of the TLFi. We only removed the definitions of non standard uses (old, slang, etc.). The extraction and cleaning-up of the definitions have been carried out in collaboration with Bruno Gaume and Philippe Muller. The bipartite graph has been created from 225 529 definitions describing 75 024 headwords (lexemes). We then removed all the features associated only with one headword. This reduces the size of the graph significantly without changing the connections that hold between the headwords. Table 1 shows that this reduction is stronger for the semantic feature (93%) than it is for the formal ones (69%). Indeed, semantic descriptions show greater variability than formal ones.

The use of the graph is illustrated in figure 4. It shows the 20 nearest neighbors of the verb *fructifier* for various propagation configurations. The examples in (a) and (b) show clearly that formal features are the more predictive ones while semantic features are the less reliable ones. The example in (c) illustrates the contribution of the semantic

(a) 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

(b) 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*

(c) 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*

Figure 4: The 20 nearest neighbors of the verb *fructifier* when the activation is spread (a) only toward the formal features, (b) only toward the semantic ones, (c) toward both the semantic and formal features. Words that do not belong to the family or series of *fructifier* are emphasized.

graph	complete	reduced
formal features	1 306 497	400 915
semantic features	7 650 490	548 641

Table 1: Number of the semantic and formal features coming from TLFi.

features. They reorder the formal neighbors and introduce among them the nearest semantic neighbors. We see in the lists in (a) and (c) that the family members are the nearest neighbors and that the members of the series come next.

7 Analogy

The members of the series and families are massively involved in the analogies which structure the lexicon. A word x belonging to a family F_x participates in several analogies with a large number of other members of F_x. The analogies that involve two words $(x, y) \in F^2$ include two other words (z, t) that belong to one same family F'. On the other hand, if x is a complex word that belongs to a series S_x, then $z \in S_x$, $x \in S_z$, $y \in S_t$ and $t \in S_y$. For instance, the couple of words *fructifier* and *fructification* form analogies with of members of other families (*rectifier*, *rectification*), (*certifier*, *certification*), (*plastifier*, *plastification*), etc. Moreover, the first elements of these couples belong to series of *fructifier* and the second ones to the series of *fructification*.

In a dual manner, a word u belonging to a series S participates in a set of analogies with a large number of other members of S. The analogies that involve two elements of the same series are made up with words which themselves belong to a same series. For instance, *fructifier* and *sanctifier* form analogies with the members of other series (*fructificateur*, *sanctificateur*), (*fructification*, *sanctification*) or (*fructifiant*, *sanctifiant*). These couples are respectively made of members of the families of *fructifier* and *sanctifier*.

7.1 Analogies and neighborhoods

The analogies that involve members of families and series can be used to efficiently filter the morphological neighbors that are identified by the method presented above. If v is a correct morphological neighbor of w, then it is either a member of the family of m or a member of its series. Therefore, it exists another neighbor v' of w (v' belong to the family of w if v belongs to the series of w or vice versa) such that it exists a neighbor w' of v and of v' such that $w : v :: v' : w'$.[2] Therefore, we have two configurations:

1. if $v \in F_w$, then $\exists v' \in S_w, \exists w' \in S_v \cap F_{v'}, w : v :: v' : w'$
2. if $v \in S_w$, then $\exists v' \in F_w, \exists w' \in F_v \cap S_{v'}, w : v :: v' : w'$

The first case is illustrated by the above examples with $w =$ *fructifier* and $v =$ *fructification*, and the second one with $w =$ *fructifier* et $v =$ *rectifier*.

7.2 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

[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.

and b are the same as the ones between c and d. It can be exemplified 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 5. They are identical for the two couples of words.

ϵ	k	a	t	a	b	a
ma	k	ϵ	t	ou	b	on

ϵ	f	a	3	a	l	a
ma	f	ϵ	3	ou	l	on

Figure 5: Formal analogy `kataba`: `maktoubon`::`fa3ala`:`maf3oulon`. The differences are locates in frame boxes.

More generally, formal analogies can be defined in terms of factorization (Stroppa and 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 the 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 for 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 $(f(a), f(b), f(c), f(d)) \in L^{\star 4}$ 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 5).

7.3 Implementation

A formal analogy $a : b :: c : d$ can be easily checked by comparing the sequences of string edit operations between (a, b) and between (c, d). Both sequences must minimize Levenshtein edit distance (i.e. have a minimal cost). Each sequence corresponds to a path in the edit lattices of the couple of words. The lattice are represented by a matrix computed using the standard string edit algorithm (Jurafsky and 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. Only three directions are allowed: upward (deletion), to the left (insertion) or in the upper left diagonal direction (substitution). Figure 6 shows the sequence of edit operations for the couple `fructueux`:`infructueusement`. Sequences of edit operations can be simplified by merging the series of identical character matchings. The sequence in figure 6 then becomes ((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`)). This simplified sequence is identical to the one for the couple `soucieux`:`insoucieusement` except for the matching operation: ((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. The resulting sequence can then be assigned to both couples as their edit signatures (σ). The formal analogy `fructueux`:`infructueusement`:: `soucieux`:`insoucieusement` can be stated in terms of identity the edit signatures:

$\sigma($`fructueux`, `infructueusement`$) =$
$\sigma($`soucieux`, `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 $\sigma(a, b) = \sigma(c, d)$ or $\sigma(a, c) = \sigma(b, d)$.

7.4 First results

The computational model we have just presented has been implemented and a first experiment has been carried out. It consists in determining the 100 closest neighbors of every headword for the three configurations presented in § 6. All the formal analogies that hold between these words have then been collected. We have not been able to do a standard evaluation in terms of recall and precision because of the lack of morphological resources for French. However, we have manually checked the analogies of 22 headwords belonging to 4 morphological families. 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.

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

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 6: Sequence of edit operations that transform `fructueux` into `infructueusement`. The type of each operation is indicated on the first line: D for deletion, I for insertion, M for matching and S for a substitution by a different character.

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 results are summarized in table 2. Their quality is quite satisfactory. However, the number of analogies strongly depends on the configuration of propagation. The best trade-off is a simultaneous propagation toward the semantic and formal features. Here are some of the correct and erroneous analogies collected:

- R.fructueusement:R.affectueusement:: A.infructueux:A.inaffectueux
- N.fructification:N.identification:: V.fructifier:V.identifier
- N.fruiterie:N.fruitier::N.laiterie:N.laitier
- * N.fruit:N.bruit::V.frusquer:V.brusquer

The first example is particularly interesting because it involves on one side suffixed words and on the other prefixed ones.

The performance of the method strongly depends on the length of the headwords. Table 3 presents the number of analogies and the error rate for 13 groups of 5 words. The words of each group are of the same length. Lengths range from 4 to 16 letters.

length	analogies	correct	errors
4	29	15	51.7%
5	22	8	36.4%
6	8	1	12.5%
7	10	2	20.0%
8	55	1	1.8%
9	29	2	6.9%
10	30	0	0.0%
11	32	0	0.0%
12	19	0	0.0%
13	11	0	0.0%
14	35	0	0.0%
15	63	0	0.0%
16	39	0	0.0%

Table 3: Number of the analogies and error rate for headwords of length 4 to 16.

8 Conclusion

We have presented a computational model that makes the morphological structure of the lexicon emerge from the formal and semantic regularities of the words it contains. The model is radically lexeme-based. It integrates the semantic and formal properties of the words in a uniform manner and represents them into a bipartite graph. Random walks are used to simulate the spreading of activations in this 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 each word are then identified among its lexical neighbors by means of formal analogies.

This is work in progress and we still have to separate the members of the families from the members 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 and Lu-

dovic Tanguy for their comments and suggestions. All errors are mine.

References

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

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

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, University of Pennsylvania, Pennsylvania, USA.

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

Bernhard, Delphine. 2006. Automatic acquisition of semantic relationships from morphological relatedness. In *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, volume 4139 of *Lecture Notes in Computer Science*, pages 121–13. Springer.

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

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

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

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

Gaume, B., F. Venant, and B. Victorri. 2005. Hierarchy in lexical organization of natural language. In Pumain, D., editor, *Hierarchy in natural and social sciences*, Methodos series, pages 121–143. Kluwer.

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

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

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

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

Lepage, Yves. 1998. Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume 2, pages 728–735, Montréal, Canada.

Lepage, Yves. 2003. *De l'analogie rendant compte de la commutation en linguistique*. Mémoire de HDR, Université Joseph Fourier, Grenoble.

Muller, Philippe, Nabil Hathout, and Bruno Gaume. 2006. Synonym extraction using a semantic distance on a dictionary. In Radev, Dragomir and Rada Mihalcea, editors, *Proceedings of the HLT/NAACL workshop Textgraphs*, pages 65–72, New York, NY. Association for Computational Linguistics.

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

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

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

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

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

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