Acquisition morphologique à partir d'un dictionnaire informatisé

Nabil Hathout

TALN 2009, Senlis, 24–26 juin 2009

Acquisition morphologique à partir d'un dictionnaire informatisé

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

Résumé. L'article propose une modélisation computationnelle permettant de faire émerger la structure morphologique dérivationnelle du lexique à partir des régularités formelles et sémantiques des mots qu'il contient. Cette modélisation, purement lexématique, consiste à découvrir les relations que chaque mot entretient avec les autres unités du lexique et notamment avec les mots de sa famille morphologique et de sa série dérivationnelle. Ces relations sont complétées par un ensemble d'analogies auxquelles le mot participe. La modélisation a été testée sur le lexique du français en utilisant le dictionnaire informatisé TLFi.

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.

Mots-clés : Analyse morphologique dérivationnelle, morphologie lexématique, similarité morphologique, analogie formelle.

Keywords: Morphological derivational analysis, lexeme-based morphology, morphological relatedness, formal analogy.

1 Morphologie morphématique *vs* lexématique

L'objectif de cette communication est d'apporter quelques éléments de réponse à la question suivante : Comment réaliser une analyse morphologique dérivationnelle dans le cadre d'une morphologie lexématique, c'est-à-dire sans recourir aux notions de morphème, d'affixe ni d'exposant morphologique ?

La morphologie est traditionnellement considérée comme la branche de la linguistique qui étudie la structure des mots. Selon cette conception, les mots sont constitués de morphèmes qui se composent selon des règles de flexion, de dérivation ou de composition. Ils ont une structure généralement représentées de façon arborescente. Par exemple, l'analyse classique d'un mot comme *dérivabilité* lui attribue la structure [[[dériv-]$_V$ -able]$_A$ -ité]$_N$. La morphologie morphématique offre un cadre théorique élégant et facile à mettre en œuvre mais présente des inconvénients importants. Les modèles théoriques proposées pour la remplacer relèvent de la mor-

phologie lexématique dans laquelle les atomes ne sont plus des morphèmes mais des mots. Les mots n'ont alors plus de structure. La structure morphologique devient alors un niveau d'organisation du lexique. basée sur le partage de propriétés sémantiques et formelles. Dans sa version adoptée ici, cette structure se compose des relations morphologiques qui s'établissent entre les mots, notamment :

– entre les formes d'un même lexème. Par exemple, la forme verbale *dérivons* appartient à l'ensemble des formes fléchies du verbe *dériver*, qui contient également *dérive*, *dériverez*, *dérivaient*, *dérivées*, *dérivions*, etc.
– entre les formes d'une même série flexionnelle. Par exemple, *dérivons* appartient à une série de formes verbales à l'indicatif présent première personne du pluriel qui inclut *acclimatons*, *compilons*, *éduquons*, *localisons*, *varions*, etc.
– entre les mots d'une même famille morphologique. Par exemple, la famille morphologique de *dérivation* contient *dériver*, *dérivable*, *dérivatif*, *dérivationnel*, *dérivabilité*, etc.
– entre les mots d'une même série dérivationnelle. Par exemple, *dérivation* appartient à une série de noms en *-ion* qui rassemble également *acclimatation*, *compilation*, *éducation*, *localisation*, *variation*, etc.

Naturellement, la morphologie ne se réduit pas à cette organisation lexicale et toutes les constructions qu'elle produit n'ont pas vocation à y entrer (par exemple *anti petit morveux qui ne connaissent plus que le mot pikachou*). Dans le reste de l'article, nous nous intéressons uniquement à la composante dérivationnelle de cette structure.

La distinction morphématique / lexématique se retrouve sur le plan computationnel. L'objectif de l'analyse morphologique d'un mot consiste à le découper en une séquence de morphèmes dans une conception morphématique (Déjean, 1998; Gaussier, 1999; Schone & Jurafsky, 2000; Goldsmith, 2001; Creutz & Lagus, 2002; Bernhard, 2006). Il est de découvrir les relations que le mot entretient avec les autres unités du lexique dans une conception lexématique. Ces relations permettent notamment d'identifier sa famille morphologique, sa série dérivationnelle ainsi que les différentes analogies auxquelles il participe. Par exemple, on considérera que l'analyse du mot *dérivation* est satisfaisante si elle met en relation *dérivation* avec un nombre suffisants de mots de sa famille morphologique et de sa série dérivationnelle. Chacune de ces relations est intégrée à un ensemble d'analogies permettant de la caractériser sur le plan sémantique et formelle. Par exemple, la relation entre *dérivation* et *dérivable* doit faire partie d'une série d'analogies incluant *dérivation*:*dérivable*::*variation*:*variable*, *dérivation*:*dérivable*::*modification*:*modifiable*, etc. De façon analogue, *dérivation* et *variation* entrent dans une série d'analogies comme *dérivation*:*variation*::*dériver*:*varier*, *dérivation*:*variation*::*dérivationnel*:*variationnel*, *dérivation*:*variation*::*dérivable*:*variable*. Les analogies morphologiques rendent explicite la structure d'hypercube du lexique.

La suite de l'article est organisée comme suit. Nous présentons d'abord les grandes lignes de notre méthode et nous la comparons à quelques travaux connexes. La section 4 décrit la mesure de similarité morphologique que nous proposons. Nous nous intéressons ensuite à l'analogie formelle et à son utilisation comme filtre sur les voisinages morphologiques (section 5). Enfin, Nous présentons en § 6 quelques résultats préliminaires.

2 Associer la similarité morphologique et l'analogie formelle

L'article propose une méthode permettant de faire émerger la structure morphologique dérivationnelle du lexique à partir des régularités formelles et sémantiques des mots qu'il contient.

Elle a été testée sur le lexique du français en utilisant le *Trésor de la Langue Française informatisé* [1] (TLFi). La méthode repose sur une mesure de similarité morphologique qui rapproche les membres des familles morphologiques et des séries dérivationnelles. Elle repose aussi sur la découverte d'analogie formelles entre voisins morphologiques. L'utilisation de l'analogie est courante en morphologie computationnelle (Lepage, 1998; Stroppa & Yvon, 2005). Le principal apport de notre méthode est de la combiner avec une mesure de proximité morphologique. Dans un premier temps, la similarité morphologique est utilisée pour sélectionner des quadruplets de mots susceptibles d'être morphologiquement apparentés. Les candidats sont ensuite filtrés au moyen de l'analogie. Ces deux techniques sont complémentaires : les voisinages morphologiques peuvent être calculer en grand nombres, mais ils sont trop grossiers pour discriminer entre les mots qui sont effectivement morphologiquement apparentés et ceux qui ne le sont pas ; l'analogie formelle permet un filtrage fin mais elle est coûteuse à calculer.

Les caractéristiques essentielles de notre modèle sont (1) que la découverte de relations morphologiques entre les mots ne fait intervenir à aucun moment la notion de morphème ni aucune représentation de morphèmes ; (2) qu'il intègre de manière uniforme les informations sémantiques et formelles ; (3) que l'appartenance aux familles et aux séries est graduelle, permettant ainsi de rendre compte du fait que, par exemple, *dériveur* est morphologiquement et sémantiquement plus proche de *dérive* que ne l'est *dérivationnellement*, bien que les trois mots appartiennent clairement à la même famille morphologique. Le modèle permet d'articuler la représentation du lexique sous la forme d'un graphe et son exploitation au moyen de parcours aléatoires dans la lignée des travaux de Gaume (2002) avec les travaux sur les analogies formelles entre chaînes de caractères (Lepage, 1998; Stroppa & Yvon, 2005) dont les algorithmes ne font pas intervenir la notion de morphèmes.

3 Travaux connexes

Un grand nombre de recherches en morphologie computationnelles visent à découvrir des relations entre des unités lexicales. Toutes s'appuient en premier lieu sur les similarités entre les formes graphémiques des mots. Ces relations sont généralement préfixales ou suffixales. Deux exceptions peuvent être signalées : (Yarowsky & Wicentowski, 2000) et (Baroni *et al.*, 2002) utilisent les distances d'édition pour estimer la similarité formelle des mots. À notre connaissance, tous les autres réalisent d'une façon ou d'une autre une segmentation, y compris ceux comme (Neuvel & Fulop, 2002) dont l'objectif n'est pas de découvrir des morphèmes. Notre modèle se distingue de ces approches par le fait que la proximité graphémique sans segmentation et de façon globale, à l'échelle de la totalité du lexique.

Notre méthode peut être également comparer aux approches qui combinent indices formels et sémantiques. Ces informations sont généralement acquises à partir de corpus en utilisant par exemple l'analyse sémantique latente comme (Schone & Jurafsky, 2000), l'information mutuelle comme (Baroni *et al.*, 2002) ou la co-occurrence à l'intérieur d'une fenêtre de mots (Xu & Croft, 1998; Zweigenbaum & Grabar, 2003). Notre approche s'en distingue par le fait que nous utilisons une ressource lexicographique et que les similarités sémantiques sont établies sur la base de parcours aléatoires dans un graphe lexical. Notre approche peut également être comparée à (Hathout, 2002), où des informations sémantiques fournies par des dictionnaires de synonymes sont utilisées pour acquérir des connaissances morphologiques.

1. `http://www.atilf.fr/tlfi.htm`

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

FIGURE 1 – Traits formels associés à *orientation*.

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 2 – Traits sémantiques induits par la définition *action d'orienter, de s'orienter ; résultat de cette action*

4 Similarité morphologique

Nous adoptons ici une définition classique de la parenté morphologique : deux mots sont morphologiquement apparentés s'ils partagent à la fois des propriétés phonologiques et sémantiques. Le TLFi ne fournissant pas la prononciation de toutes les entrées, nous utilisons les propriétés graphémiques à la place des propriétés phonologiques. La similarité morphologique est estimée en utilisant un bigraphe qui contient un ensemble de sommets qui représentent les lexèmes et un autre de sommets qui représentent leurs propriétés formelles et sémantiques (voir figure 3).

Les traits formels associées à un lexème sont les n-grammes de lettres qui apparaissent dans son lemme. Nous imposons aux n-grammes une taille minimale ($n \geq 3$). Le début et la fin des lemmes sont marqués par des $. La figure 1 présente une partie des n-grammes associés au mot *orientation*. Signalons que cette description n'accorde de statut privilégié à aucun des n-grammes. Tous y jouent le même rôle : leur seule fonction est de rapprocher les mots qui contiennent les mêmes sons.

De façon similaire, les traits sémantiques qui décrivent les lexèmes sont les n-grammes de mots qui apparaissent dans leurs définitions. Les traits qui contiennent des ponctuations sont éliminés : nous n'utilisons que les n-grammes qui se trouvent dans des segments compris entre deux ponctuations. Par exemple, les traits sémantiques induits par la définition *action d'orienter, de s'orienter ; résultat de cette action* du lexème *orientation* sont présentés en figure 2. Les mots des définitions sont catégorisés et lemmatisés. Les étiquettes utilisées sont : A pour les adjectifs ; R pour les adverbes ; N pour les noms ; V pour les verbes ; X pour toutes les autres catégories. Cette représentation très grossière de la sémantique des mots est inspirée des segments répétés (Lebart *et al.*, 1998). Elle présente plusieurs avantages : (1) elle est fortement redondante afin de capter les ressemblances qui existent entre les définitions ; (2) les n-grammes permettent d'intégrer des informations de nature syntagmatique sans réaliser une véritable analyse syntaxique des définitions ; (3) elle gomme légèrement les variations qui existent dans le traitement lexicographique des mots, notamment dans les découpages en sous-sens et dans les définitions.

Le bigraphe est construit en connectant de façon symétrique chaque mot à l'ensemble de ses traits formels et sémantiques (voir figure 3). La structure de graphe bipartie n'est pas essentielle mais elle est utile car elle permet de propager de façon synchrone d'une activation dans les sous-graphes formels et sémantiques.

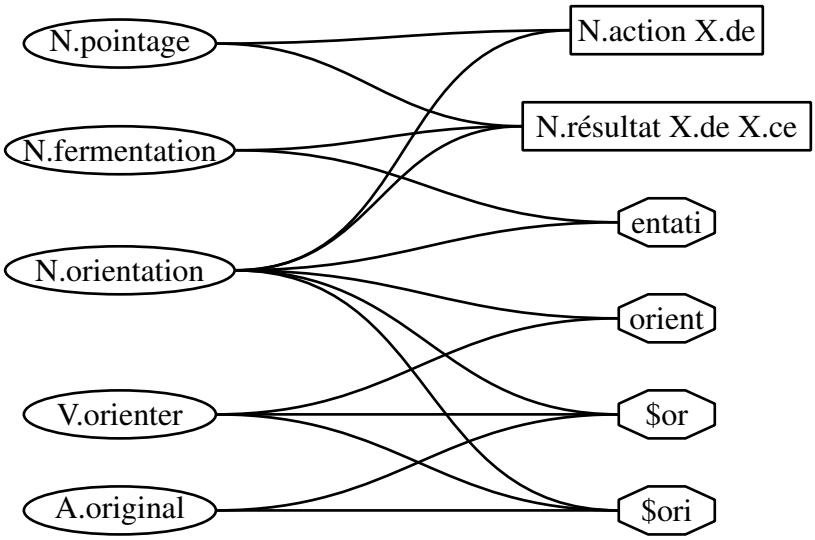

FIGURE 3 – Extrait du bigraphe qui représente le lexique. Les lexèmes se trouvent dans des ovales, les traits formels par des octogones et les traits sémantiques par des rectangles. Le graphe est symétrique.

4.1 Parcours aléatoires

La similarité morphologique est estimée en propageant une activation dans le bigraphe un nombre pair de fois. Dans un graphe fortement redondant comme celui que nous venons de décrire, une propagation de longueur 2 (des mots vers les traits puis des traits vers les mots) permet d'obtenir les proximités visées. La propagation est simulée par des parcours aléatoires et calculée de façon classique en multipliant la matrice d'adjacence stochastique du graphe (Gaume *et al.*, 2002; Muller *et al.*, 2006).

Dans l'exemple de la figure 3, les voisins morphologiques du mot *orientation* sont identifiés en générant une activation au niveau du sommet qui représente ce mot. Lors de la première étape, l'activation est propagée vers l'ensemble des sommets qui représentent ses traits formels et sémantiques. Lors de la deuxième étape, l'activation qui se trouve au niveau des traits est propagée vers les mots. Ainsi *orienter* se trouvera activé via le trait formel `$or`, `$ori`, `orien` et *fermentation* par l'intermédiaire du trait formel `entati` et du trait sémantique `N.résultat X.de X.ce`. L'activation que l'on obtient au niveau de chaque mot est d'autant plus grande que le nombre de traits qu'il partage avec *orientation* est élevé et que ces traits sont spécifiques. L'hypothèse sous-jacente est que le niveau d'activation est une estimation de la parenté morphologique.

4.2 Voisinage lexical dans le graphe du TLFi

Le graphe que nous utilisons est construit à partir des définitions du TLFi. Nous en avons éliminé celles qui concernent les emplois non standards (archaïsmes, argotiques, etc.). Le bigraphe a ainsi été créé à partir de 225 529 définitions décrivant 75 024 entrées (lexèmes). Nous avons supprimé les traits qui n'apparaissent que dans un mot. On réduit ainsi fortement la taille du graphe sans modifier les connexions qui s'établissent entre les mots. On peut voir dans le tableau 1 que la réduction est plus forte pour les propriétés sémantiques qu'elle ne l'est pour les propriétés formelles.

traits	complet	réduit	hapax
formels	1 306 497	400 915	69%
sémantiques	7 650 490	548 641	93%
total	8 956 987	949 556	90%

TABLE 1 – Nombre des traits sémantiques et formels issus du TLFi.

(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 R.infructueusement A.infructueux V.transsubstantifier V.substantifier V.stratifier V.schistifier V.savantifier V.refortifier V.ratifier V.quantifier**
(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 N.conductibilité N.alevinage V.profiter **A.fructifiant** N.pouvoir V.agir N.opération V.placer N.rentabilité N.jouissance
(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 N.police N.hangar N.dîme N.ban A.fruitier V.affruiter A.instructif A.obstructif A.destructif A.constructif

FIGURE 4 – Les 30 voisins les plus proches du verbe *fructifier* lorsque l'on utilise seulement les traits formels (a), seulement les traits formels (b), à la fois les traits formels et sémantiques (c). Les mots qui appartiennent à la famille ou à la série de *fructifier* sont en gras.

Pour illustrer l'utilisation du graphe, nous présentant en figure 3 les 30 premiers voisins du verbe *fructifier* pour différentes configurations de propagation : en (a), avec une propagation réalisée uniquement vers les traits formels ; en (b), uniquement vers les traits sémantiques ; en (c), pour moitié vers les traits formels et pour moitié vers les traits sémantiques. On voit en (a) que les membres de la famille morphologique tendent à être plus proches que ceux de la série dérivationnelle, en l'occurrence les verbes en *-ifier*. (a) et (b) montrent clairement que les traits formels sont les plus prédictifs, que les traits sémantiques le sont les moins.

5 Analogies

Les éléments des séries ou des familles sont massivement impliqués dans les analogies qui structurent le lexique. Par exemple, le couple *fructifier* et *fructification* forme des analogies avec plusieurs couples d'éléments appartenant respectivement aux séries de *fructifier* et de *fructification* comme (*rectifier*, *rectification*), (*certifier*, *certification*), (*plastifier*, *plastification*), etc. De façon duale, *fructifier* et *sanctifier* forment des analogies avec les membres de leurs familles respectives comme (*fructificateur*, *sanctificateur*), (*fructification*, *sanctification*) ou (*fructifiant*, *sanctifiant*).

Les analogies permettent de filtrer efficacement les voisinages morphologiques. Si v est un voisin morphologique **correct** de m, c'est soit un élément de la famille de m, soit un élément de sa série. Il existe alors un autre voisin v' de m (v' appartient à la famille de m si v appartient

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 – Analogie formelle kataba:maktoubon::fa3ala:maf3oulon. Les différences sont situées dans les parties encadrées.

à la série de m ou vice versa) tel qu'il existe w voisin de v et de v' tel que $m : v :: v' : w$ [2]. Il n'existe ainsi que deux configurations possibles :

1. si $v \in F_m$, alors $\exists v' \in S_m, \exists w \in S_v \cap F_{v'}, m : v :: v' : w$
2. si $v \in S_m$, alors $\exists v' \in F_m, \exists w \in F_v \cap S_{v'}, m : v :: v' : w$

où F_x représente la famille de x et S_x sa famille. La première est illustrée par les exemples ci-dessus avec m = *fructifier* et v = *fructification*, et la seconde avec m = *fructifier* et v = *rectifier*.

5.1 Analogies formelles

Une analogie formelle ou graphémique est une relation $a : b :: c : d$ qui s'établit entre quatre formes telle que les différences graphémiques qui existent entre a et b sont les mêmes que celles qui existent entre c et d. Elle peut être illustrée, en s'inspirant de l'un des exemples proposés par (Lepage, 1998; Lepage, 2003), par les quatre formes arabes kataba:maktoubon::fa3ala:maf3oulon, transcriptions de la forme de citation du verbe 'écrire', du nom de résultat 'écrit', du verbe 'faire' et du nom de résultat 'effet'. Les différences entre les deux premières formes et les deux dernières sont présentées en figure 5. Elles sont identiques pour les deux couples.

Les analogies formelles peuvent être définies en utilisant la notion de factorisation (Stroppa & Yvon, 2005; Langlais & Patry, 2007). Soit L un alphabet et $a \in L^\star$ une chaîne de caractères définie sur L. On appelle factorisation de a de longueur n une séquence de n chaînes de caractères $f_1, \cdots, f_n$ dont la concaténation est égale à a. Plus formellement, $f \in L^{\star n}$ est une factorisation de a de longueur n si $f = (f_1, \cdots, f_d)$ et $a = f_1 \cdot f_2 \cdots f_d$. Par exemple, (ma, k, ϵ, t, ou, b, on) est une factorisation de longueur 7 de maktoubon. On peut alors définir l'analogie formelle comme suit. Soit $(a, b, c, d) \in L^{\star 4}$ quatre chaînes de caractères. $a : b :: c : d$ constitue une analogie formelle ssi il existe un entier $n \in \mathbb{N}$ et quatre factorisations de longueur n des quatre chaînes $(f(a), f(b), f(c), f(d)) \in L^{\star 4}$ telles que $\forall i \in [1, n], (f_i(b), f_i(c)) \in \{(f_i(a), f_i(d)), (f_i(d), f_i(a))$. Dans le cas de l'analogie kataba:maktoubon:: fa3ala:maf3oulon, la propriété est vérifiée pour $n = 7$.

5.2 Mise en œuvre

Les analogies formelles peuvent être vérifiées simplement en comparant des séquences d'opérations d'édition permettant de transformer une chaîne de caractère en une autre. Cette séquence peut être simplement déduite de la table de distances d'édition de Levenshtein des deux chaînes (Jurafsky & Martin, 2000). Chaque séquence d'opération permettant de transformer la première chaîne de caractères en la seconde correspond à un parcourir du tableau

2. Nous notons $a : b :: c : d$ le fait que (a, b, c, d) forme un quadruplet analogique, c'est-à-dire que a est à b ce que c est d.

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

FIGURE 6 – Séquence d'opérations permettant de passer de `fructueux` (a) à `infructueusement` (b) représentée sous la forme d'une correspondance entre deux factorisations des deux chaînes de caractères $f(a)$ et $f(b)$. L'opération correspondant à un couple $(f_i(a), f_i(a))$ est indiquée sur la première ligne par I pour une insertion, D pour une suppression, M pour une identité et S pour une substitution par un caractère différent.

en partant de la dernière case et en remontant jusqu'à la première. Nous ne nous intéressons ici qu'aux séquences de longueur minimale obtenues en sélectionnant pour chaque case la voisine de coût le plus faible et en cas d'égalité, en préférant la case qui se trouve sur la diagonale (substitution), et à défaut celle de gauche (insertion) puis celle du haut (suppression). La figure 6 présente la séquence d'opérations que l'on obtient pour le couple `fructueux:infructueusement`. On peut la simplifier en fusionnant les successions d'identité : ((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)). La séquence similaire pour le couple `soucieux:insoucieusement` est identique à l'exception leurs sous-séquences d'identité : ((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)). Les deux peuvent être rendues strictement identiques si l'on ne spécifie pas ces sous-chaînes. On peut ainsi rendre compte de l'analogie formelle `fructueux:infructueusement::soucieux:insoucieusement` en associant à chaque couple sa séquence comme signature d'édition (σ). En l'occurrence $\sigma(\texttt{fructueux}, \texttt{infructueusement}) = \sigma(\texttt{soucieux}, \texttt{insoucieusement}) =$ ((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)). Plus généralement, quatre chaînes de caractères forment une analogie formelle $a : b :: c : d$ si $\sigma(a, b) = \sigma(c, d)$ ou bien $\sigma(a, c) = \sigma(b, d)$.

6 Premiers résultats

Nous avons implémenté le modèle computationnel qui vient d'être présenté et réalisé une première expérience consistant à déterminer les 100 voisins les plus proches de chaque entrée pour les 3 configurations présentées en § 4.2, puis à calculer toutes les analogies formelles différentes qui s'établissent entre les mots qui se trouvent dans ces voisinages . Nous avons ensuite révisé manuellement ces analogies $a : b :: c : d$ en considérant qu'elles sont correctes si $b \in F_a, c \in S_a, d \in S_b \cap F_c$ ou si $b \in S_a, c \in F_a, d \in F_b \cap S_c$. La révision a été faite sur un ensemble de 22 mots appartenant à 4 familles morphologiques. Voici quelques exemples d'analogies correctes et erronées :

– R.fructueusement:R.affectueusement::A.infructueux:A.inaffectueux
– N.fruiterie:N.fruitier::N.laiterie:N.laitier
– * N.fruit:N.bruit::V.frusquer:V.brusquer
– * A.fruité:A.truité::N.frusquin:N.trusquin

Le premier exemple est particulièrement intéressant car il implique d'un côté des mots suffixés et de l'autre des mots préfixés. Les résultats obtenus sont résumés dans le tableau 2. On observe que la qualité des résultats est très satisfaisante, mais que la quantités des quadruplets varie fortement en fonction du type de parcours.

parcours	analogies	corrects	erreur
form	169	163	3.6%
sém	5	5	0.0%
sém + form	130	128	1.5%

TABLE 2 – Nombre d'analogies et taux d'erreur.

7 Conclusion

Nous avons présenté un modèle computationnel capable de faire émerger la structure morphologique du lexique morphologique. Ce modèle purement lexématique intègre de manière uniforme les propriétés sémantiques et formelles des mots au sein d'un bigraphe permettant de simuler la propagation d'une activation dans un réseau lexical. Le niveau d'activation obtenu à la suite de la propagation permet d'identifier des voisins lexicaux parmi lesquels on peut retrouver les membres de la famille morphologique de l'entrée et les éléments de sa famille en constituant des quadruplets analogiques.

Il s'agit d'un travail en cours que nous envisageons de poursuivre dans deux directions. La première consiste à comparer les résultats que nous obtenons avec ceux de systèmes comme *Linguistica* (Goldsmith, 2001) ou comme l'analyseur de (Bernhard, 2006). La seconde consiste à répéter l'expérience sur l'anglais pour pouvoir réaliser une évaluation précise en utilisant la base CELEX (Baayen *et al.*, 1995).

Remerciements

Je remercie L'ATLIF et Jean-Marie Pierrel d'avoir mis à notre disposition le TLFi. Je remercie également Bruno Gaume, Philippe Muller pour les nombreuses discussions que nous avons eu sur le nettoyage du TLFi et son exploitation. Je suis également reconnaissant à Gilles Boyé, Olivier Haute-Cœur et Ludovic Tanguy pour leurs commentaires et suggestions.

Références

BAAYEN R. H., PIEPENBROCK R. & GULIKERS L. (1995). The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, University of Pennsylvania, Pennsylvania, USA.

BARONI M., MATIASEK J. & TROST H. (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*, p. 48–57, Philadelphia: ACL.

BERNHARD D. (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*, p. 121–13: Springer.

CREUTZ M. & LAGUS K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL Workshop on Morphological and Phonological Learning*, p. 21–30, Philadelphia, Penn.: ACL.

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

GAUME B., DUVIGNEAU K., GASQUET O. & GINESTE M.-D. (2002). Forms of meaning, meaning of forms. *Journal of Experimental and Theoretical Artificial Intelligence*, **14**(1), 61–74.

GAUSSIER É. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*, University of Mariland, USA: Association for Computational Linguistics, ACL'99.

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

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

JURAFSKY D. & MARTIN J. H. (2000). *Speech and language processing*. Prentice-Hall.

LANGLAIS P. & PATRY A. (2007). Translating unknown words by analogical learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning, EMNLP-CoNLL 2007*, p. 877–886, Prague, Czech Republic: ACL.

LEBART L., SALEM A. & BERRY L. (1998). *Exploring textual data*. Dordrecht: Kluwer Academic Publishers.

LEPAGE Y. (1998). Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume 2, p. 728–735, Montréal, Canada.

LEPAGE Y. (2003). *De l'analogie rendant compte de la commutation en linguistique*. Mémoire de HDR, Université Joseph Fourier, Grenoble.

MULLER P., HATHOUT N. & GAUME B. (2006). Synonym extraction using a semantic distance on a dictionary. In D. RADEV & R. MIHALCEA, Eds., *Proceedings of the HLT/NAACL workshop Textgraphs*, p. 65–72, New York, NY: Association for Computational Linguistics.

NEUVEL S. & FULOP S. A. (2002). Unsupervised learning of morphology without morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning 2002*, Philadelphia: ACL Publications.

SCHONE P. & JURAFSKY D. S. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, p. 67–72, Lisbon, Portugal.

STROPPA N. & YVON F. (2005). An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, p. 120–127, Ann Arbor, Michigan: Association for Computational Linguistics.

XU J. & CROFT W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems*, **16**(1), 61–81.

YAROWSKY D. & WICENTOWSKI R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the Association of Computational Linguistics (ACL-2000)*, p. 207–216, Hong Kong.

ZWEIGENBAUM P. & GRABAR N. (2003). Learning derived words from medical corpora. In *9th Conference on Artificial Intelligence in Medicine Europe*, p. 189–198, Cyprus.