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 un modèle linguistique et informatique permettant de faire émerger la structure morphologique dérivationnelle du lexique à partir des régularités sémantiques et formelles des mots qu'il contient. Ce modèle est radicalement lexématique. La structure morphologique est constituée par 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 forment des paradigmes analogiques. La modélisation a été testée sur le lexique du français en utilisant le dictionnaire informatisé TLFi.

Abstract. The paper presents a linguistic and 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 word-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 : Morphologie dérivationnelle, morphologie lexématique, similarité morphologique, analogie formelle.

Keywords: Derivational morphology, word-based morphology, morphological relatedness, formal analogy.

1 Introduction

L'objectif de cet article 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 ?

Cette étude, encore exploratoire, s'inscrit dans le cadre d'une théorie morphologique et d'un modèle informatique en cours de co-élaboration (section 1). Elle porte sur une mesure de similarité morphologique (section 3) et une méthode d'acquisition de relations dérivationnelles (section 4), deux composants fondamentaux d'un système qui permettra à terme de créer un réseau morphologique global à partir de ressources lexicographiques et de corpus. Nous présentons également quelques résultats préliminaires (section 5), une évaluation détaillée de ces deux composants étant pour l'heure prématurée.

Théorie morphologique. Cette étude s'inscrit dans le cadre d'une théorie morphologique lexématique dans laquelle les atomes ne sont pas des morphèmes mais des mots. Dans cette théorie, les mots n'ont pas de structure. La structure morphologique est conçue comme un niveau d'organisation du lexique, basé sur les relations de partage de propriétés sémantiques et formelles entre mots (Bybee, 1985). Dans la théorie proposée ici, cette structure se compose notamment des relations morphologiques :

– 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 *compilation*, *éducation*, *variation*, etc.

Chacune de ces relations est intégrée à un ensemble d'analogies permettant de la caractériser sur le plan sémantique et formel. Par exemple, la relation entre *dérivation* et *dérivable* fait partie d'une série d'analogies incluant *dérivation*:*dérivable*::*variation*:*variable* [1], *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*.

Naturellement, la morphologie ne se réduit pas à cette organisation lexicale et toutes les constructions produites par la première n'ont pas vocation à entrer dans la seconde (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.

Modèle informatique. La distinction entre morphologie morphématique et morphologie lexématique se retrouve sur le plan informatique. Dans une conception morphématique, l'analyse d'un mot consiste à le découper en une séquence de morphèmes (Déjean, 1998; Gaussier, 1999; Schone & Jurafsky, 2000; Goldsmith, 2001; Creutz & Lagus, 2002; Bernhard, 2006). Par exemple, *dérivabilité* est analysé en $[[[\text{dériv-}]_V \text{ -able}]_A \text{ -ité}]_N$. Dans une conception lexématique, elle est de découvrir les relations que le mot entretient avec les autres unités du lexique. Ces relations permettent de déterminer sa position dans le lexique, notamment d'identifier sa famille morphologique, sa série dérivationnelle et les analogies auxquelles il participe. Par exemple, on considérera que l'analyse du mot *dérivation* est satisfaisante si elle le met en relation avec un nombre suffisant d'éléments de sa famille morphologique et de sa série dérivationnelle.

L'article propose une méthode originale permettant de faire émerger une partie de la structure morphologique dérivationnelle du lexique. L'apport principal de la méthode est de combiner analogie formelle et mesure de proximité morphologique. Dans un premier temps, la similarité morphologique est utilisée pour sélectionner des quadruplets de mots susceptibles d'être morphologiquement apparentés. Ces candidats sont ensuite filtrés au moyen de l'analogie. Ces deux techniques sont complémentaires : les voisinages morphologiques peuvent être calculés

1. 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.

en grand nombre, 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. La méthode a été testée en utilisant le *Trésor de la Langue Française informatisé* [2] (TLFi).

Les caractéristiques principales 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 morphème ; (2) qu'il intègre de manière uniforme les informations sémantiques et formelles ; (3) que l'appartenance aux familles et aux séries est graduelle, permettant par exemple de rendre compte du fait que *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 l'analogie formelle entre chaînes de caractères (Lepage, 1998; Stroppa & Yvon, 2005; Langlais & Patry, 2007) dont les algorithmes ne font pas intervenir la notion de morphème.

2 Travaux connexes

Un grand nombre de recherches en morphologie informatique 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 et / 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 celles 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 est estimée sans segmentation, de façon globale, à l'échelle de la totalité du lexique.

Notre méthode peut être également comparée aux approches qui combinent indices formels et sémantiques. Ces derniers sont généralement acquis 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 comme (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 proposition peut être rapprochée de celle de (Hathout, 2002) qui utilise un dictionnaire de synonymes ou de (Claveau & L'Homme, 2005) qui établissent des relations morphologiques au sein d'une terminologie.

3 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 pro-

2. `atilf.atilf.fr/tlf.htm`

priétés graphémiques à la place des propriétés phonologiques [3]. 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 1).

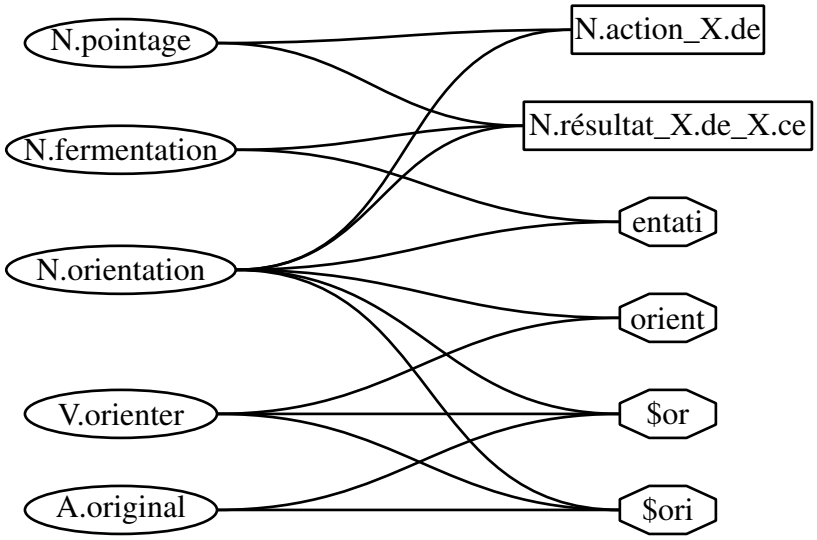

FIGURE 1 – Extrait du bigraphe qui représente le lexique. Les lexèmes se trouvent dans des ovales, les traits formels par des octogones et les traits sémantiques par des rectangles. Le graphe est symétrique.

Les traits formels associés 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 2 présente une partie des n-grammes associés au mot *orientation*. Signalons que dans cette description aucun n-grammes n'a le statut de morphème. Tous les n-grammes jouent le même rôle. Leur seule fonction est de rapprocher les mots qui contiennent les mêmes sons.

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

FIGURE 2 – Liste partielle des traits formels associés au mot *orientation*.

De façon similaire, les traits sémantiques qui décrivent un lexème sont les n-grammes de mots qui apparaissent dans ses définitions. Les n-grammes qui contiennent des ponctuations sont éliminés. 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 3. 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, pour capter les ressemblances qui existent entre les définitions ; (2) les n-grammes permettent

3. Nous envisageons de reprendre l'expérience présentée ici en utilisant des représentations phonologiques calculées par LliaPhon (Béchet, 2001).

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 la formulation des définitions.

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

Le bigraphe est construit en connectant de façon symétrique chaque mot à l'ensemble de ses traits formels et sémantiques (voir figure 1). La structure de graphe bipartie n'est pas essentielle mais elle est utile car elle permet de propager de façon synchrone une activation dans les sous-graphes formel et sémantique.

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 1, 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 les traits formels et sémantiques de *orientation*. Lors de la deuxième étape, l'activation qui se trouve au niveau des traits est propagée vers les mots. Ainsi *orienter* se trouve activé via les traits formels `$or`, `$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 récupère 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 (i.e. peu fréquents). L'hypothèse sous-jacente est que le niveau d'activation est une estimation du degré de parenté morphologique.

Voisinage lexical dans le graphe du TLFi. Le graphe que nous utilisons est construit à partir des entrées et des définitions du TLFi. Nous en avons éliminé celles qui concernent des emplois non standards (archaïques, argotiques, etc.). Le bigraphe est ainsi créé à partir de 225 529 définitions décrivant 75 024 entrées. 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 voit dans le tableau 1 que l'élagage est plus fort pour les propriétés sémantiques qu'il ne l'est pour les propriétés formelles.

La figure 4 illustre l'utilisation du graphe. Elle présente les 30 premiers voisins du verbe *fructifier* pour différentes configurations de propagation. 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*. On constate également en comparant (a) et (b), que les traits formels sont les plus prédictifs et que les traits sémantiques sont les moins fiables.

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 formels et sémantiques issus du TLFi. La première colonne indique le nombre initial des traits, la deuxième, le nombre des traits associées à deux mots ou plus et la dernière, le taux de réduction.

(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 (a) seulement les traits formels, (b) seulement les traits sémantiques et (c) à la fois les traits formels et sémantiques. Les mots qui appartiennent à la famille ou à la série de *fructifier* sont en gras.

4 Analogies

Les éléments des séries et des familles sont massivement impliqués dans les analogies qui structurent le lexique. Par exemple, le couple (*fructifier*, *fructification*) forme des analogies avec plusieurs couples composés d'éléments des 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 à la série de m ou vice versa) tel qu'il existe w voisin de v et de v' tel que $m : v :: v' : w$[4]. 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 série. La configuration 1 est illustrée par les exemples ci-dessus avec m = *fructifier* et v = *fructification*, et la 2 avec m = *fructifier* et v = *rectifier*.

4. 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.

Analogies formelles. Une analogie formelle est une relation $a : b :: c : d$ qui s'établit entre quatre formes telles que les différences graphémiques qui existent entre a et b sont les mêmes que celles qui existent entre c et d. La figure 5 en présente un exemple adapté de (Lepage, 1998; Lepage, 2003). On constate que les différences entre les deux premières formes et les deux dernières sont bien identiques.

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 entre les transcriptions de quatre mots arabes, kataba: maktoubon::fa3ala:maf3oulon. Ces mots sont respectivement la forme de citation du verbe 'écrire', du nom de résultat 'écrit', du verbe 'faire' et du nom de résultat 'effet'. Les différences sont situées dans les parties encadrées.

Les analogies formelles peuvent être définies en utilisant la notion de factorisation (Stroppa & Yvon, 2005). 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 $f_1, \ldots, f_n$ dont la concaténation est égale à a. 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 $(f(a), f(b), f(c), f(d)) \in (L^{\star n})^4$ 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 $n = 7$.

Mise en œuvre. Les analogies formelles peuvent être vérifiées en comparant les séquences d'opérations d'édition permettant de transformer une chaîne de caractères en une autre. Ces séquences peuvent être déduites simplement des tables de distances d'édition de Levenshtein (Jurafsky & Martin, 2000). Chaque séquence d'opérations permettant de transformer une première chaîne de caractères en une seconde correspond à un parcourir dans le tableau qui part de la dernière case et remonte 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. Cette séquence peut être simplifiée en fusionnant les identités contiguës : ((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 de la sous-séquence d'identités : ((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 séquences peuvent être rendues strictement identiques si ces sous-chaînes ne sont pas spécifiées. 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 $\sigma(a, b) = \sigma(c, d)$. Insistons sur le fait que le calcul des signatures ne fait intervenir ni morphème ni représentation de morphème.

	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 $f(a)$ et $f(b)$ des deux chaînes de caractères. 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.

5 Premiers résultats

Nous avons implémenté le modèle informatique qui vient d'être présenté et réalisé une première expérience consistant à déterminer les 100 plus proches voisins de chaque entrée pour les trois configurations présentées en figure 4, puis à calculer les analogies formelles qui s'établissent entre cette entrée et les mots qui se trouvent dans ces voisinages. Nous avons ensuite révisé manuellement les analogies induites par un échantillon de 22 entrées appartenant à 4 familles morphologiques de la tranche *fr-* du *TLF*[5]. Une analogie $a : b :: c : d$ est considérée correcte si $b \in F_a, c \in S_a, d \in S_b \cap F_c$ ou si $b \in S_a, c \in F_a, d \in F_b \cap S_c$. Voici quelques exemples de quadruplets corrects et erronés :

– 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 quantité des quadruplets varie fortement en fonction du type de parcours.

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

TABLE 2 – Nombre de quadruplets acquis pour un échantillon de 22 entrées et taux d'erreur.

taille	4	5	6	7	8	9	10	11	12	13	14	15	16
quadruplets	29	22	8	10	55	29	30	32	19	11	35	63	39
corrects	14	14	7	8	54	27	30	32	19	11	35	63	39
erreur (%)	51.7	36.4	12.5	20.0	1.8	6.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 3 – Nombres de quadruplets pour 13 échantillons de 5 entrées sélectionnés aléatoirement. Les quadruplets sont acquis dans des voisinages calculés en utilisant à la fois les traits formels et sémantiques. Les entrées dans chaque échantillon sont de la même taille. La taille varie entre 4 et 16 caractères.

Les performances de la méthode dépendent fortement de la longueur des entrées parce qu'elle s'appuie avant tout sur les similarités formelles et que ces dernières sont d'autant plus grandes

5. Les entrées de cette tranche ont été classées en familles morphologiques en 1999 par Josette Lecomte et moi.

que les mots sont longs. Cette corrélation apparaît très clairement dans le tableau 3. On y voit notamment que les quadruplets issus d'entrées de 10 caractères ou plus sont toutes correctes.

6 Conclusion

Nous avons présenté un modèle informatique capable de faire émerger une partie de la structure morphologique du lexique. 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 les voisins lexicaux de chaque entrée. On peut ensuite retrouver parmi ces voisins les membres de la famille morphologique de l'entrée et les éléments de sa série dérivationnelle en constituant des quadruplets analogiques.

Cette étude préliminaire fait partie d'un projet de constitution d'une base de données dédiée à la description morphologique. Nous envisageons dans ce cadre de construire un réseau morphologique en utilisant une méthode par bootstrap à partir des entrées les plus longues. Nous prévoyons également de répéter l'expérience sur l'anglais notamment pour 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 Bruno Gaume et Philippe Muller pour les nombreuses discussions que nous avons eu sur *prox* et sur le nettoyage et l'exploitation du TLFi. Je suis reconnaissant à Gilles Boyé, Olivier Haute-Cœur, Ludovic Tanguy et aux trois relecteurs anonymes de TALN 2009 pour leurs commentaires et suggestions.

Références

BAAYEN R. H., PIEPENBROCK R. & GULIKERS L. (1995). CELEX-2. Cdrom. Penn.: LDC.

BARONI M., MATIASEK J. & TROST H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL Workshop on Morphological and Phonological Learning*, p. 48–57, Philadelphia, Penn.: ACL.

BERNHARD D. (2006). Automatic acquisition of semantic relationships from morphological relatedness. In *Advances in Natural Language Processing*, volume 4139 of *LNCS*, p. 121–132: Springer.

BYBEE J. L. (1985). *Morphology. A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins Publishing Company.

BÉCHET F. (2001). Lia_phon : un système complet de phonétisation de textes. *Traitement Automatique des Langues*, **42**(1), 47–67.

CLAVEAU V. & L'HOMME M.-C. (2005). Structuring terminology by analogy-based machine learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhague.

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.

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

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 LREC-2002*, 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 EMNLP-CoNLL 2007*, p. 877–886, Prague: 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*, p. 728–735, Montréal.

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 *Proceedings of the HLT/NAACL Workshop Textgraphs*, p. 65–72, New York, NY: ACL.

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

SCHONE P. & JURAFSKY D. S. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of CoNLL-2000*, p. 67–72, Lisbonne.

STROPPA N. & YVON F. (2005). An analogical learner for morphological analysis. In *Proceedings of CoNLL-2005*, p. 120–127, Ann Arbor, Michigan: ACL.

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 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.