Dispersion sémantique dans des familles morpho-phonologiques : éléments théoriques et empiriques

N. Gala[1] V. Rey[2] L. Tichit[3]
(1) LIF, Marseille, CNRS UMR 6166
(2) SHADYC, Marseille, CNRS et EHESS, UMR 8562
(3) IML, Marseille, CNRS UMR 6206
nuria.gala@lif.univ-mrs.fr, veronique.rey@univ-provence.fr,
laurent.tichit@univmed.fr

Résumé. Traditionnellement, la morphologie lexicale a été diachronique et a permis de proposer le concept de famille de mots. Ce dernier est repris dans les études en synchronie et repose sur une forte cohérence sémantique entre les mots d'une même famille. Dans cet article, nous proposons une approche en synchronie fondée sur la notion de continuité à la fois phonologique et sémantique. Nous nous intéressons, d'une part, à la morpho-phonologie et, d'autre part, à la dispersion sémantique des mots dans les familles. Une première étude (Gala & Rey, 2008) montrait que les familles de mots obtenues présentaient des espaces sémantiques soit de grande cohésion soit de grande dispersion. Afin de valider ces observations, nous présentons ici une méthode empirique qui permet de pondérer automatiquement les unités de sens d'un mot et d'une famille. Une expérience menée auprès de 30 locuteurs natifs valide notre approche et ouvre la voie pour une étude approfondie du lexique sur ces bases phonologiques et sémantiques.

Abstract. Traditionally, lexical morphology has been diachronic and has established the notion of word families. This notion is reused in synchronic studies and implies strong semantic coherence within the words of a same family. In this paper, we propose an approach in synchrony which highlights phonological and semantic continuity. Our interests go on morpho-phonology and on the semantic dispersion of words in a family. A first study (Gala & Rey, 2008) showed that the semantic spaces of the families displayed either a strong semantic cohesion or a strong dispersion. In order to validate this observation, we present here a corpus-based method that automatically weights the semantic units of a word and a family. An experience carried out with 30 native speakers validates our approach and allows us to foresee a thorough study of the lexicon based on phonological and semantic basis.

Mots-clés : morpho-phonologie lexicale, traitement automatique des familles dérivationnelles, espaces sémantiques.

Keywords: lexical morpho-phonology, derivational families processing, semantic spaces.

1 Introduction

Il existe à l'heure actuelle des ressources linguistiques ayant comme particularité commune le regroupement de mots en familles. L'objectif de ces outils est fondamentalement pédagogique ; cependant, la construction et la constitution de telles ressources présente des intérêts indéniables pour le traitement automatique des langues. Parmi ces ressources, la notion de 'famille' est prise en compte selon des points de vue très variés : étymologique (dictionnaire de familles de mots de Synapse[1]), analogique (outil pédagogique du Centre Collégial de Développement de Matériel Didactique du Québec[2]) ou encore thématique (JeuxdeMots [3](Lafourcade, 2007)). Notre contribution illustre une quatrième approche, celle des familles morpho-phonologiques, et permet d'apporter un éclairage nouveau dans l'analyse morphologique et sémantique du lexique du français.

Notre intérêt porte sur la morphologie lexicale. Les approches fondées sur les regroupements de synonymes et les associations thématiques s'écartent de nos objectifs du fait que les mots appartenant à une même famille ne partagent pas de lien de forme. La première approche s'éloigne également de notre démarche. En effet, en dépit d'une continuité de sens entre des formes voisines (dérivation régulière), notre proposition est complètement en synchronie et s'ancre dans une théorie de la morphologie lexicale qui tient compte de la phonologie (Kiparsky, 1982). Ainsi, nos familles morpho-phonologiques partagent une continuité de forme et de sens, ce sens pouvant parfois être très proche ("table" *vs* "tablette") ou bien très dispersé ("lune" *vs* "lunettes").

Dans la suite de cet article (section 2) nous décrirons la notion de 'famille morpho-phonologique' et nous nous intéresserons à l'impact de la phonologie dans la morphologie dérivationnelle, tout en préservant la notion de continuité de forme et de sens dans les familles. La section 3 décrit une méthode qui utilise des corpus structurés et permet l'ajout d'informations sémantiques associées aux mots. Enfin, la section 4 étudie la dispersion sémantique de quelques familles de mots et compare les résultats obtenus empiriquement à ceux obtenus suite à une expérience menée avec 30 locuteurs natifs.

2 Éléments théoriques

Le principe est de rassembler les mots dans une même famille. Nous proposons de définir le mot non pas comme une unité qui "manque de rigueur" (Dubois *et al.*, 1994) , mais comme une unité de désignation d'un objet, d'une réalité, d'un état, d'une action, d'une qualité : le mot, en associant une forme sonore à un "morceau" de réalité, est une unité qui montre quelque chose. De nombreux morphèmes (les clitiques, les prépositions, ...) ne sont pas alors des mots. Dans la suite de ce travail, le terme "mot" est préféré au terme "lexème".

Le concept de famille de mots n'est pas en soi très novateur. La méthodologie morpho-phonologique requise est, elle, plus novatrice. A la suite des travaux de Corbin (Corbin, 1987), la description synchronique des structures morphologiques cherche à segmenter les morphèmes constituant un mot. Cependant l'objectif n'est pas de dégager des règles génératrices de nouvelles constructions morphologiques. Classiquement, les formes différentes attestées pour un

[1]http ://www.synapse-fr.com/produits/Famille.htm
[2]http ://www.ccdmd.qc.ca/fr/jeux_pedagogiques/id=1089&action=animer
[3]http ://www.jeuxdemots.org

mot ne sont pas liées à la dérivation mais à des éléments grammaticaux (genre, nombre, conjugaison...). Or, au moment des constructions dérivationnelles, les mots sont aussi transformées par des principes d'alternances consonantiques ou vocaliques. "Chaleur" et "chaud" appartiennent à la même famille ; ils présentent une alternance al/au que l'on retrouve dans d'autres familles ("cheval", "chevaucher"). Sur ce point, notre approche rejoint la théorie de la morphologie et de la phonologie lexicale (*Lexical morphology and phonology theory*) (Kiparsky, 1982). Trois principes sont établis : 1/ les règles de la formation des mots et les règles phonologiques s'appliquent ensemble dans un même composant isolé ; 2/ la périodicité est une propriété des règles phonologiques ; l'application cyclique est un mode d'application qui n'est pas une propriété inhérente de la grammaire ; 3/ l'application cyclique des règles phonologiques devrait s'établir à partir de l'organisation du lexique. Parallèlement, Corbin (Corbin, 1987) démontre que les mots n'épuisent pas toutes les constructions potentielles : certaines dérivations pourraient exister, mais elles ne sont pas attestées dans la langue ("refaire" existe, mais "redanser" n'est pas attesté). Nous retenons deux éléments : la morpho-phonologie a toute sa pertinence pour appréhender les constructions dérivationnelles ; cependant, les mots, reflétant des pratiques sociales hétérogènes, ne permettent pas l'application "aveugle" de dérivation phonologique cyclique et systématique. Deux critères méthodologiques sont alors retenus : la continuité sémantique entre deux mots et les variations phonologiques participant à la construction d'une famille de mots.

L'analyse permet de rendre compte que de nombreux mots sont construits avec des syllabes sans signification. Par exemple, "soustraction", "abstraction", "tracter", "tracteur" ont-ils un radical commun ? La technique de la commutation (garder le plus petit minimum commun) conduit à proposer "tract". Comment cela est-il possible ? Premier réflexe, ce rassemblement est erroné et ne veut rien dire. Deuxième réflexe, en regardant l'histoire écrite de la langue, des attestations peuvent ou non accréditer cette approche. Troisième réflexe, cela pose un problème linguistique en synchronie, car ce sont les usagers qui fabriquent, emploient, modifient les formes des mots et leur sens. Par conséquent, nous ne retiendrons pas l'approche diachronique des dictionnaires. Notre propos est de rendre compte d'une organisation lexicale de la langue française contemporaine sur un double principe organisationnel : phonologique et sémantique.

3 Éléments empiriques

3.1 Le rideau est ridé : justifications phonologiques et sémantiques

Nous avons appliqué une grille d'analyse pour construire des familles de mots sur un principe morpho-phonologique. Le résultat est une ressource lexicale, Polymots, permettant la consultation des familles à partir de mots ou d'affixes (Gala & Rey, 2008). L'analyse mise en oeuvre a montré qu'effectivement les familles obtenues contiennent une moyenne de 10 mots pour un même radical ou *mot base* (pouvant aller jusqu'à plus de 50). De plus, beaucoup de mots sont construits à l'aide d'un radical sans signification lorsqu'il est isolé -nous l'appelons mot base *opaque*- (1/3 des entrées, exemples "clam", "duct", "dict", "opt", etc.).

Une fois résolu le traitement morpho-phonologique des unités, la question de la variation lexicale au sein d'une même famille est alors posée : avons-nous le droit de rassembler dans une même famille, des mots partageant un même radical morpho-phonologique mais appartenant à des champs sémantiques apparemment différents (comme "abstraction" et "soustraction") ? Si oui,

pourquoi ? Si le concept de famille de mots repose sur un radical commun et une signification commune, alors il s'agit, d'après nous, d'une construction très réductrice. En effet, comme unité de désignation, le mot peut comporter plusieurs traits sémantiques (dans une perspective structuraliste) générant des mots très différents. Ce phénomène est très bien documenté d'un point de vue historique. Il suffit pour cela de lire les très belles pages du dictionnaire historique de langue française de A. Rey. Notre hypothèse est qu'il en est de même dans l'actualité de la langue.

Entre les mots "boule", "boulotte" et "boulon", le trait sémantique de 'rondeur' est commun ; cela est donc relativement transparent pour le lecteur. Cependant, le mot "bouleverser" pose question : il y a bien la forme "boule" ; il y a également la forme "verse" qui indique un mouvement de retournement (on ne peut verser que vers le bas). La construction de ces deux mots, si on accepte cette analyse, conduit à la désignation d'un état d'âme. Le regroupement de mots sur un principe morpho-phonologique interroge donc la continuité sémantique[4]. Tel est le propos de cette étude.

3.2 Acquisition d'informations sémantiques

La question qui nous intéresse est l'étude de la 'dispersion sémantique' d'une famille de mots ; l'idée est d'associer à chaque entrée lexicale de Polymots un ensemble de termes constituant son 'espace sémantique'. La constitution et l'enrichissement de ressources reste une tâche particulièrement difficile en ce qui concerne la méthode mise en œuvre. Sans évoquer les ressources construites manuellement (dont le coût est significatif), il existe un certain nombre de méthodes semi-automatiques qui se servent des ressources existantes (dictionnaires, listes de synonymes, ontologies comme WordNet, etc.). Beaucoup d'entre elles utilisent plus particulièrement les définitions de dictionnaires pour différents objectifs (Ide & Véronis, 1990), (Brun *et al.*, 2001), (L'Homme, 2003). Les définitions ont été, aussi, notre point de départ.

3.2.1 Méthode générale

Dans un souci de diversification de corpus, mais confrontés à la difficulté d'obtenir facilement des ressources structurées, nous avons choisi d'utiliser des sources lexicographiques et encyclopédiques accessibles librement, à savoir, Wiktionnaire et Wikipédia. Nous avions aussi à notre disposition le disctionnaire Hachette-XML. La méthodologie que nous avons utilisée repose sur l'extraction semi-automatique des définitions (hors exemples) dans le cas des dictionnaires et du texte introductif (avant la table de matières) dans le cas de Wikipédia. Nous avons fait l'hypothèse de la présence de termes significatifs dans ces extraits[5], c'est-à-dire des termes que nous appellerons par la suite des *unités de sens*, caractérisant sémantiquement chaque mot m donné. Par exemple, à partir de la première définition de "bras" du Wiktionnaire :

"(Anatomie) Partie du membre superieur des humains (et des bipèdes en général) qui s'étend depuis l'épaule jusqu'au coude"

nous obtenons "partie, membre, supérieur, humains, bipèdes" etc.

[4] Nous attirons l'attention sur le fait que la dérivation flexionnelle permet une économie mnésique car le locuteur peut créer des néologismes à partir de mots connus. Ceci n'est pas possible dans toutes les langues.

[5] Noms (hors gloses et autres termes lexicographiques), verbes, adjectifs, quelques adverbes.

La liste obtenue est appelée l'*espace sémantique* du mot et est notée $\alpha(m)$. Soit un corpus (filtré avec un 'antidictionnaire' et lemmatisé avec TreeTagger) constitué de plusieurs définitions lexicographiques et d'informations encyclopédiques concernant un multi-ensemble $\mathcal{M}$ de mots (qui peuvent être des mots base ou des mots dérivés). Il s'agit, pour chaque mot $m \in \mathcal{M}$, d'extraire sa liste $\alpha(m)$ d'unités de sens :

$$\alpha(m) = (u_1, u_2, \ldots, u_i, \ldots, u_n)$$

3.2.2 Pondération

Nous attribuons à chaque unité de sens u_i présente dans $\alpha(m)$ un poids $\omega(u_i)$ en fonction de la distance de chaque unité u_i au mot m, et en tenant compte du nombre total de mots n dans chaque définition, avec $0 \leq i < n$:

$$\forall u_i \in \alpha(m), \omega(u_i) = 1 - i/n$$

Nous considérons que l'importance d'une unité de sens u_i diminue en fonction de sa distance au mot m. De fait, l'unité de sens la plus proche de m aura un poids de 1, et à chaque éloignement d'une unité de sens, le poids ω de l'unité de sens suivante sera divisé selon un écart constant (exemple, $1/4$ si la définition contient quatre unités de sens, etc.). Nous obtenons donc, pour chaque mot m, un espace sémantique $\beta(m)$ pondéré.

$$\beta(m) = \{(u_1, \omega(u_1)), (u_2, \omega(u_2)), \ldots, (u_i, \omega(u_i)), \ldots, (u_n, \omega(u_n))\}$$

3.2.3 Sommation et harmonisation

Il est courant que le même mot apparaisse plusieurs fois dans le corpus initial, constitué de définitions extraites de plusieurs sources. $\mathcal{M}$ est donc multi-ensemble. Chaque occurrence m_j de m aura donc son propre espace sémantique pondéré $\beta(m_j)$, composé de couples $(u, \omega_j(u))$. Pour obtenir le poids global Ω d'une unité de sens d'un mot m donné, il est donc nécessaire d'additionner, pour chaque unité de sens u donnée, les poids $\omega_j(u)$. Soit p le nombre d'occurences du mot m dans $\mathcal{M}$:

$$\forall u \in \bigcup_{j=1}^{p} \alpha(m_j), \Omega(u) = \sum_{j=1}^{p} \omega_j(u)$$

Afin de pouvoir comparer les valeurs des poids entre différents mots donnés, nous considérons que le poids maximal d'une unité de sens doit être égal à 1 (le terme u est très porteur de sens pour m), et le poids minimum strictement supérieur à 0. Nous harmonisons donc, pour chaque mot m, les poids calculés en divisant l'ensemble des valeurs par le poids maximal. Grâce à ce type de normalisation, notre méthode possède l'avantage d'être indépendante de la taille globale du corpus. La figure 1 illustre le début de l'espace sémantique obtenu pour "embrasser" :

```
[serrer 1] [contenir 0.666] [saisir 0.666] [bras 0.585] [attacher
0.444] [entourer 0.444] [étendre 0.314] [regard 0.314] [adopter
0.296] [baiser 0.248] [englober 0.166][étreindre 0.148] ...
```

FIG. 1 – Extrait du résultat pour le mot "embrasser".

Les espaces sémantiques obtenus contiennent une moyenne de 105 unités de sens. Comme l'illustre la Figure 1 pour "embrasser", on y retrouve des termes synonymes ("serrer", "englober", "étreindre" etc.) et des termes liés thématiquement ("bras", "regard", "baiser")[6].

4 Dispersion sémantique

4.1 Définition et principe

Nous définissons la *dispersion sémantique* comme la présence très faible, voire l'absence, du mot base dans l'espace sémantique d'un mot dérivé de sa famille. Le sens des unités lexicales avec une forte dispersion sémantique a évolué par rapport au mot base, cette évolution se faisant à la manière de la dérivation par métaphore évoquée par (Picoche, 1999) (sens figurés : "vache" dans le sens 'mechanceté', "mine" dans le sens 'explosif'), ou bien par extension d'un seul des traits sémantiques du mot base. Par exemple, d'après nos résultats, le mot base "fil" a un poids de 0,12 dans "défiler", ce qui montre bien une dispersion sémantique du mot dérivé par rapport aux autres membres de la famille ("fil" dans "faufiler" a un poids de 0.50). Pour "défiler", la continuité sémantique existe néanmoins dans la mesure où il est possible d'identifier des traits sémantiques communs ("long" avec 0.32 dans "défilé" et 0.19 dans "fil"). À l'inverse, il existe des familles avec très peu de dispersion sémantique : il y a donc une forte continuité de sens. Dans ces cas, la dérivation a conservé les traits sémantiques 'principaux' du mot base ; la famille de "terre" ou de "bras" en sont des exemples (cf. Figure 1, "bras" dans "embrasser").

4.2 Évaluation des résultats

30 locuteurs francophones ont été testés sur un jugement de dispersion sémantique de mots relevant d'une même famille. Ils devaient donner un score entre 0 et 1 pour caractériser la dispersion sémantique entre la forme de base et le mot dérivé : 0 indique une grande dispersion sémantique (peu de rapport entre deux mots) et 1 une très faible dispersion (un rapport fort). Le tableau 1 présente les résultats pour six familles morpho-phonologiques. Pour chaque couple de mots, nous donnons les moyennes établies par les locuteurs (L) et par le corpus (C). Pour une même famille, la dispersion sémantique perçue par les locuteurs est mise en évidence par l'ordre des colonnes (à gauche forte dispersion, sens éloigné ; à droite faible dispersion, sens proche) :

	L	C		L	C		L	C
arme-alarme	0.17	0.71	arme-armoire	0.41	-	arme-armure	0.90	1
court-courtois	0.22	-	court-raccourcir	0.77	0.37	court-écourter	0.88	0.88
vache-avachir	0.26	-	vache-vacherin	0.66	0.20	vache-vacherie	0.74	1
fil-faufiler	0.49	0.50	fil-défiler	0.69	0.12	fil-filiation	0.74	0.25
bras-embrasser	0.57	0.58	bras-bracelet	0.77	0.36	bras-brassard	0.95	1
terre-terrasse	0.70	0.51	terre-enterrement	0.81	0.43	terre-territoire	0.89	0.58

TAB. 1 – Dispersion sémantique perçue par des locuteurs et sur corpus.

[6]Pour les homonymes, les différents sens sont explicites ("mine" : "aspect", "galerie", "explosif", etc.)

Il s'avère que si le mot base a des traits sémantiques homogènes (par exemple "terre" avec 'espace', 'surface', 'matière', etc.) la dispersion sémantique dans la famille est faible ; le mot base a, d'ailleurs, des poids importants dans les espaces sémantiques de ses dérivés (par exemple, 0.58 de "terre" dans "territoire", 0.51 dans "terrasse", etc.). Si, en revanche, le mot base a plusieurs traits sémantiques (comme "val" qui peut indiquer une forme ou une descente) alors la dispersion est forte. Ainsi, entre "arme" et "alarme" les locuteurs perçoivent mal la continuité sémantique (présente néanmoins dans la définition du dictionnaire et rendue évidente grâce aux corpus). En effet, le mot "alarme" présente des traits sémantiques liés à 'signal', 'alerte', etc. et non pas des traits liés à 'militaire', 'guerre' ; cépendant, les traits 'dispositif' et 'protéger' sont communs.

Une comparaison des résultats obtenus grâce aux locuteurs avec ceux obtenus grâce à l'extraction semi-automatique à partir de corpus confirme la même tendance L-C pour des mots de la famille "vache" et "court", et très similaire pour "bras" et "terre" : il s'agit de familles avec peu de dispersion sémantique. Dans certains cas comme "fil" et "arme" les résultats sont légèrement différents entre L et C soit par l'absence du mot base dans l'espace sémantique d'un des mots dérivés ("arme" dans "amoire"), soit à cause de la perception de certains mots : "défiler" est fortement dispersé selon le corpus (0.12) et peu selon la perception des locuteurs (0.69).

Enfin, le mot base est absent dans tous les mots de type opaque ("val" dans "vallée") ou dans certains mots transparents ("court" dans "courtois"). Dans ces cas-là, des traits sémantiques récurrents sont néanmoins présents dans les espaces sémantiques ('descendre' dans "avaler" et "dévaler", 'fort' dans "glu" et "gluten", etc.).

5 Conclusion

Nous avons présenté une étude de la morphologie lexicale du français fondée sur la notion de continuité de forme et de sens. Dans le cas de familles peu dispersées sémantiquement, la perception des locuteurs et les résultats obtenus empiriquement sont très cohérents, ce qui permet de valider notre approche. Pour les mots (ou les familles) avec une dispersion sémantique importante, la création d'espaces sémantiques à l'aide de corpus s'avère moins évidente. Ceci nous amène à prendre en compte les limites de notre approche : d'une part, l'influence de la qualité et de la taille des corpus lexicographiques et encyclopédiques, ainsi que l'impact des erreurs du TreeTagger ; d'autre part, le traitement principalement statistique des corpus. Ainsi, nous envisageons l'ajout de traitements linguistiques plus approfondis (extraction de dépendances syntaxiques, prise en compte de la négation, etc.) pour une meilleure attribution des poids dans des cas de constructions particulières : négations, phrases relatives, clivées, etc. ('sans vigueur', 'qui a peu de longueur', 'ce qui est séparé d'un ensemble', etc.).

Enfin, outre l'adaptation et l'incorporation des espaces sémantiques dans la base Polymots, nous travaillons à une caractérisation sémantique approfondie des familles morpho-phonologiques qui tient compte du rôle des affixes et des mots base dans la construction des unités lexicales.

Remerciements

Nous remercions les trois relecteurs anonymes pour leurs remarques et critiques pertinentes.

Références

BRUN C., JACQUEMIN B. & SEGOND F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique. *Traitement Automatique des Langues*, **42**(3), 667–691.

CORBIN D. (1987). *Morphologie dérivationnelle et structuration du lexique*, volume 1 & 2. Tubingen : Max Niemeyer, Verlag.

J. DUBOIS, M. GIACOMO, L. GUESPIN, C. MARCELLESI & J. P. MÉVEL, Eds. (1994). *Dictionnaire de la linguistique et des sciences du langage*. Paris, Larousse.

GALA N. & REY V. (2008). Polymots : une base de données de constructions dérivationnelles en français á partir de radicaux phonologiques. In *Traitement Automatique des Langues Naturelles, TALN 2008*, Avignon.

IDE N. & VÉRONIS J. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionnaries. In *13th International Conference on Computational Linguistics, COLING 90*, volume 2, p. 389–394, Helsinki.

KIPARSKY P. (1982). *From cyclic Phonology to lexical Phonology*, In *The structure of Phonological Representations*, volume 1, p. 131–175. V. H. and S. N. New York : Dordrecht.

LAFOURCADE M. (2007). Making people play for lexical acquisition. In *7th Symposium on Natural Language Processing, SNLP 2007*, Pattaya, Thailande.

L'HOMME M. C. (2003). Acquisition de liens conceptuels entre termes à partir de leur définition. *Cahiers de lexicographie*, **83**(2), 19–34.

PICOCHE J. (1999). Dialogue autour de l'enseignement du vocabulaire. *Études de linguistique appliquée*, **16**, 421–434.