POLYMOTS : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques

Nuria Gala (1), Véronique Rey (2)

(1) LIF – CNRS, Aix – Marseille Univ., 163 av. de Luminy, 13288 Marseille
nuria.gala@univ-provence.fr
(2) SHADYC – CNRS, Aix – Marseille Univ., 29 av. R. Schuman, 13100 Aix
veronique.rey@univ-provence.fr

Résumé Cet article présente POLYMOTS, une base de données lexicale contenant huit mille mots communs en français. L'originalité de l'approche proposée tient à l'analyse des mots. En effet, à la différence d'autres bases lexicales représentant la morphologie dérivationnelle des mots à partir d'affixes, ici l'idée a été d'isoler un radical commun à un ensemble de mots d'une même famille. Nous avons donc analysé les formes des mots et, par comparaison phonologique (forme phonique comparable) et morphologique (continuité de sens), nous avons regroupé les mots par familles, selon le type de radical phonologique. L'article présente les fonctionnalités de la base et inclut une discussion sur les applications et les perspectives d'une telle ressource.

Abstract In this paper we present POLYMOTS, a lexical database containing eight thousand common nouns in French. Whereas most of the existing lexicons for derivational morphology take affixes as starting point for producing paradigms of words, we defend here the idea that it is possible to isolate a morpho-phonological stem and produce a paradigm of words belonging to the same family. This point leads us to describe three types of stems according to their phonological and morphological form. The article presents the different features of such a lexical database and discusses the applications and future work using and enriching this resource.

Mots-clés : ressource lexicale, morphologie dérivationnelle, traitement automatique des familles de mots.

Keywords: lexical resource, derivational morphology, word families processing.

1 Introduction

Le travail, décrit dans cet article, s'inscrit dans la problématique de la construction et de l'utilisation de ressources lexicales. L'approche proposée, bien que liée à la morphologie dérivationnelle, s'en éloigne dans la mesure où les formes de base, que nous appelons « radicaux phonologiques », ne sont pas forcément des lemmes correspondant à des mots de la langue, mais plutôt des formes phonologiques communes à un ensemble de mots de la même famille, ayant ou non un sens établi.

Nous pensons que la productivité morphologique dans la construction lexicale dépend non seulement de la structure de surface des mots (forme sonore immédiate des mots) mais aussi de la structure profonde (règle sous jacente rendant compte des constructions irrégulières). Nous rejoignons ainsi les travaux de Corbin (1987) sur la morphologie dérivationnelle. Elle dit ainsi :

« Toute personne qui a acquis la connaissance d'une langue a intériorisé un système de règles qui détermine des connexions son-sens pour une infinité de ***mots construits****. (…) C'est ce système de règles qui le rend capable de produire et d'interpréter des* ***mots construits*** *qu'il n'a jamais rencontré auparavant.(…) Chacun comprend et produit des* ***mots construits*** *nouveaux sans aucune conscience de leur nouveauté (…) » p. 47-48*

Par ailleurs, Rey-Debove (1982 et 2004) propose un dictionnaire où les mots seraient classés selon l'analyse morphologique. En 1982, l'approche est résolument synchronique ; en 2004, histoire de la langue et analyse synchronique sont présentées simultanément. Selon Rey-Debove (2004 : 193), « la composante lexicale a résisté à l'étude synchronique et la lexicologie théorique est surtout historique ».

Il s'agit de l'analyse des mots en synchronie en radicaux, préfixes et suffixes. L'analyse morphologique est de type concaténatoire et ne prend pas en compte les phénomènes structuraux, ces derniers étant décrits selon une approche diachronique.

Cependant, Corbin (1987) montre que l'analyse morphologique ne doit pas séparer les deux dimensions : des informations historiques peuvent fournir certains renseignements parfois nécessaires et/ou compléter une défaillance dans la compétence du morphologue. De plus, les mots n'ont pas le même âge dans le lexique d'aujourd'hui : certains, comme « père », sont très anciens, d'autres, comme « informatique », sont plus récents.

Notre démarche teste l'hypothèse suivante : à partir d'un corpus de mots extraits d'un dictionnaire, peut-on construire des familles de mots reposant uniquement sur leur analyse structurelle, en tenant compte des contraintes morpho-phonologiques ? L'histoire des mots de la langue n'est pas donc à l'origine de leur segmentation en radicaux et affixes, et certains rapprochements pourront surprendre le lecteur. Huit mille mots ont été ainsi étudiés.

L'article est organisé en trois parties : dans la première (section 2) nous montrons quelques systèmes existants en TAL pour la morphologie dérivationnelle et nous en dégageons leurs spécificités ; par la suite (section 3) nous décrivons notre approche, en nous focalisant sur les choix linguistiques adoptés. Enfin, la section 4 est consacrée à la description des fonctionnalités de la base et à son évaluation, tout en mettant en perspective les applications visées à terme : outil pédagogique pour les enseignants de l'école primaire (famille de mots et didactique de l'orthographe) ; outil pour construire des bases lexicales pour des études en

psycholinguistique et outil pour construire des exercices de remédiation auprès d'orthophonistes.

2 Construction de ressources pour la morphologie et TAL

Dans le domaine de la morphologie dérivationnelle en TAL, la littérature est riche d'exemples d'outils et de méthodes. On peut distinguer les approches visant à construire (semi)-automatiquement une liste de mots dérivés à partir d'une racine, et les approches permettant d'obtenir le lemme racine à partir d'un mot dérivé. Quelle que soit l'optique adoptée, il existe des méthodes à base de connaissances et des méthodes à base d'analogies à partir de corpus.

2.1 Méthodes utilisant des listes de référence

Dans cette approche, on utilise des dictionnaires (Savoy, 93) ou des bases lexicales de référence (listes de mots) à partir desquelles, moyennant des règles de correspondance, on découpe les mots candidats selon un lemme de base et une suite d'affixes. Ainsi par exemple, DériF (Dal et al., 99) est un analyseur morphologique qui effectue l'analyse dérivationnelle d'un mot étiqueté avec des règles élaborées à partir d'hypothèses linguistiques. Ces règles découpent le mot selon des suffixes et/ou préfixes et produisent une séquence également étiquetée avec le mot base (suffixé) et le(s) affixe(s). Par exemple pour « inexplicable » l'analyseur donnera [in [[expliquer VERBE] able ADJ] ADJ].

Dans Cordial (Synapse), logiciel proposant différents traitement linguistiques (correction orthographique, grammaticale), une option « Famille de Mots » est proposée parmi l'ensemble de ressources accessibles (dictionnaires de synonymes, de citations, de locutions de contextes d'apparition de mots, etc.). Pour un mot donné, base de la famille, ce dictionnaire donne le détail des constituants des familles (ou liste de termes dérivés). La recherche se fait à partir d'un mot base existant dans la langue. Cette ressource est constituée de 15.000 noms propres (exemple, « Victor Hugo » produit « hugolien ») et 85.000 noms communs (exemple, « parent » est lié à « apparentage », « apparenté », etc.). A chaque terme dérivé lui est associée sa catégorie morphosyntaxique.

2.2 Méthodes utilisant des corpus

Les méthodes à base d'analogies cherchent à vérifier sur un corpus l'existence de formes dérivées crées semi-automatiquement à partir d'une combinaison de mots base et de suffixes donnés. Par exemple, DéCor (Dal et al., 99) est un outil de construction d'un lexique dérivationnel qui permet de calculer des bases possibles (d'un point de vue statistique) à partir de formes lexicales contenant des suffixes. Il permet de relier de façon systématique des dérivés avec un terme de base existant dans la langue. Par exemple, « imperturbable » sera relié à « perturber » car « perturbable » n'est pas attesté.

Un deuxième exemple de ce type de méthodes à base de corpus est Webaffix (Tanguy et Hathout, 02), un outil d'acquisition de couples de lexèmes morphologiquement liés. Ici on utilise une liste de référence de termes et on cherche sur le Web des formes lexicales nouvelles (absentes de cette liste initiale) en fonction de leur terminaison (suffixes permettant de construire des noms d'action).

Dans cette approche on situe, enfin, les travaux de morphologie orientés vers la recherche d'information. En tenant compte de cette application précise, le but est l'obtention de mots dérivés à partir d'une racine de façon a être utilisés, par exemple, pour étendre les requêtes (Jacquemin, 97).

Ces différentes approches rendent compte de phénomènes linguistiques de surface, à savoir la structure immédiate du lexème. Cependant, la structure profonde des mots n'est pas problématisée. Par exemple, est-il possible de rapprocher dans une même famille morphologique, des mots comme « filature » et « défilé » ? Si l'on s'en tient au sens contemporain des mots, non. De plus, on a tendance à considérer qu'un rapprochement relèverait davantage de la diachronie, de l'histoire de la langue. Nous avons fait l'hypothèse qu'il serait possible d'appréhender la structure profonde des mots tout en restant dans la langue contemporaine.

3 Choix linguistiques

Le travail réalisé jusqu'à présent, repose sur une liste de vingt mille mots communs extraits des 59 000 entrées du dictionnaire Larousse (2000). Ces mots désignent soit un objet (concret ou abstrait), soit une activité, soit une qualité (Rey & Stoffel, 2006). Il s'agit donc d'unités de désignation. Les noms scientifiques et les noms propres n'ont pas été intégrés dans cette base. En revanche, dans les familles de mots, il est possible de retrouver des dérivés de noms communs de type adjectif (« anachronique », « charbonneux », etc.), verbe (« converser », « soustraire », etc.) et adverbe (« unanimement », « accessoirement », etc.).

A partir de cette liste initiale de vingt mille mots communs, nous avons analysé manuellement huit mille entrées, de façon à les décomposer et à en extraire leur forme de base, au niveau structurel. L'objectif était d'isoler une forme sonore commune aux membres d'une même famille et de lister les affixes composant une entrée donnée. Cette approche nous a permis d'obtenir, à ce jour, six cents familles de mots (voir tableau 1, § 3.2).

Jusqu'à présent, de nombreux travaux en dérivation ont essentiellement traité des affixes et de leur génération automatisée. Ainsi, Moeschler et Auchlin (2000) relèvent 60 préfixes et 150 suffixes en langue française. Cependant, si l'analyse repose sur une nouvelle segmentation des mots (analyse structurelle), il est tout à fait possible que l'inventaire et le nombre d'affixes soient différents. L'approche que nous proposons serait alors originale de par le fait que nous travaillions sur des « radicaux phonologiques » et que ces derniers donnent un éclairage nouveau sur la notion d'affixe.

3.1 Les radicaux phonologiques

Le principe de base de la comparaison repose sur le principe de comparaison de forme et de sens (quand ce dernier cas est possible). Soit une forme sonore « brume ». Par dérivation, on obtient les mots suivants « embrumer, brumeux, embrun ». Ce dernier mot, « embrun », a un radical phonologique « brun » que l'on entend dans la série « brune, brunette, bruni, brunâtre ». Mais il y a rupture de sens entre la brume et la couleur : la brume n'est pas brune. Nous proposons alors de considérer le mot « embrun » uniquement dans la première série. Cet exemple est simple. Dans le cas du mot « fin », entre la fin d'une histoire et la tranche fine de pain, nous proposons une continuité de sens (il y a de fait une continuité de forme) et d'après nous, une continuité sémantique. Parler de polysémie serait une solution trop rapide : une

« définition » est quelque chose d'achevé, de fini et également de précis, de fin. Mais ceci, dans le cadre de cette première exploration, dépend de la subjectivité linguistique du linguiste. Nous souhaiterions à l'avenir établir des traits sémantiques permettant un travail plus systématique.

Par comparaison phonologique (forme phonique comparable) et morphologique (continuité de sens), nous avons regroupé les mots par familles, sous un « radical phonologique » commun. Nous définissons ce terme comme une unité récurrente dans une même famille de mots qui peut avoir -ou pas- un sens dans la langue.

Fradin (2003) montre que le concept de morphème n'épuise pas les unités minimales de construction. En effet, des mots peuvent se construire en dérivation avec des bases sans morphème (si l'on considère que le morphème a toujours une unité de sens). Par exemple, le mot « bikini » a permis la construction du dérivé « monokini » et pourtant "-kini" n'est pas un morphème et ne fait référence à aucune base linguistique historique (latine ou grecque). Dans ce cas précis, ce sont les affixes qui produisent le sens du mot. Il est donc pertinent de repérer les bases phonologiques qui permettent de créer des familles de mots. Ces dernières partagent, dans un continuum, un sens, mais le radical n'a pas nécessairement une signification.

3.2 Typologie

L'analyse morphologique des huit mille entrées (repérage manuel des radicaux phonologiques et des différents affixes) nous a conduit à distinguer trois types de mots : les transparents, les alternants et les opaques.

Les entrées dont la forme de base est transparente (TSP) constituent le premier groupe. Pour ces mots, le radical a une réalisation phonique et un sens. Par exemple, la forme de base « fil » a un sens et permet de construire soixante-treize dérivés (« effiler », « filature », « filon », « surfilage », etc.)

Dans le deuxième groupe on retrouve les mots dont la forme de base est phonologiquement alternante (ALT). Le radical a une variation phonique lors des constructions en dérivation mais maintient le sens. Par exemple, le /f/ de la forme « actif » peut être transformée en /v / en dérivation (« activité », « active », « réactivité », etc.) ; le /a/ alterne avec /e/ dans des mots comme « pacification » et « paix » ; le /ʃ/ et le /k/ alternent dans « monarch/monarc » (« monarchie », « monarque »), etc. La liste des alternances vocaliques et consonantiques est disponible dans Sabater et Rey (2005).

Les mots TSP et ALT ont donc des radicaux morpho-phonologiques puisque ces radicaux ont toujours un sens.

Enfin, on distingue les mots dont la forme de base est opaque (OPAK). Pour ce groupe le radical a une réalisation phonique mais n'a pas de sens aujourd'hui. Par exemple, « voc » ne signifie rien mais cette forme se retrouve dans « vocal », « vocaliser », « évoquer », « invoquer », « révoquer », « vocation », « provoquer », etc. Il en va de même pour « panta » dans « pantalon » et « pantacourt » ; pour « cid » dans « coïncidence », « décision », etc.

Les huit mille mots traités jusqu'à présent sont distribués de la façon suivante entre les trois catégories :

	TSP	ALT	OPAK	total
Mots dérivés	2940	3062	2044	8046
Distribution	36,54%	38,06%	25,40%	
Radicaux phonologiques	288	155	158	601
Distribution	47,92%	25,79%	26,29%	

TAB 1. Distribution des types de radicaux morphologiques

On constate qu'il n'y a pas de différence significative entre les trois groupes (test de Fisher TSP vs ALT, p = .58 ; TSP vs OPAK p=.78 ; ALT vs OPAK p= .40). Cela signifie donc que le nombre de mots est également réparti dans les trois groupes.

La catégorie de mots la plus représentée est la catégorie ALT : les formes de base sont très productives puisque 150 mots produisent 3062 mots, ce qui veut dire qu'une forme de base peut produire environ 19 dérivés. Viennent ensuite les mots OPAK, où une forme de base produit environ 13 mots dérivés (2044 pour 158 radicaux).

Enfin, les moins productifs sont les TSP avec 288 bases pour 2940, c'est-à-dire qu'un radical produit environ 10 mots dérivés. Comme l'écriture des formes de base est souvent maintenue dans toute la famille de mots, cette productivité est non seulement intéressante en langue orale, mais également pour l'apprentissage de l'orthographe lexicale.

Dans tous les cas, on observe que la procédure de dérivation est également pertinente dans les trois groupes. Les noyaux TSP génèrent moins de constructions dérivationnelles et confirment par ce fait l'intérêt d'aborder la structure profonde des mots pour rendre compte de la dérivation.

3.3 Informations associées aux entrées lexicales

La base ne donne pas, pour l'instant, d'autres informations sur les mots de type strictement linguistique. Les verbes ne présentent pas une construction dérivationnelle particulière. Les rares cas de verbes conjugués devenus des mots (comme « rendez-vous ») ne donnent pas de dérivation. De même, le pluriel ne génère pas une organisation spécifique (par exemple, « cheval » et « chevaux » donnent respectivement « chevaucher » et « chevalier »). De plus, les règles syntaxiques ne s'appliquent pas sur les principes de la construction lexicale : comme le soulignait déjà Benveniste (1974), le mot « maintenir » place l'objet « main » avant le verbe « tenir », construction qui est impossible dans le cadre phrastique. Le travail portant uniquement sur la construction du lexique, les informations catégorielles (nom, verbe, adjectif) ou grammaticales (nombre, genre, …) ne sont donc pas mentionnées.

Par ailleurs, ne s'agissant pas d'un dictionnaire au sens « classique », la base ne contient pas de définitions ou de liens lexicaux entre les entrées (hyperonymes, synonymes, etc.). Les seules informations pertinentes associées aux entrées correspondent aux jeux de construction, c'est-à-dire à la formation morpho-phonologique de chaque entrée, ainsi qu'aux liens entre les formes d'une même famille. Quelques indices quantitatifs (nombre de mots appartenant à une même famille, fréquence d'apparition d'un affixe donné, etc.) sont aussi proposés.

4 Fonctionnalités de recherche dans la base POLYMOTS

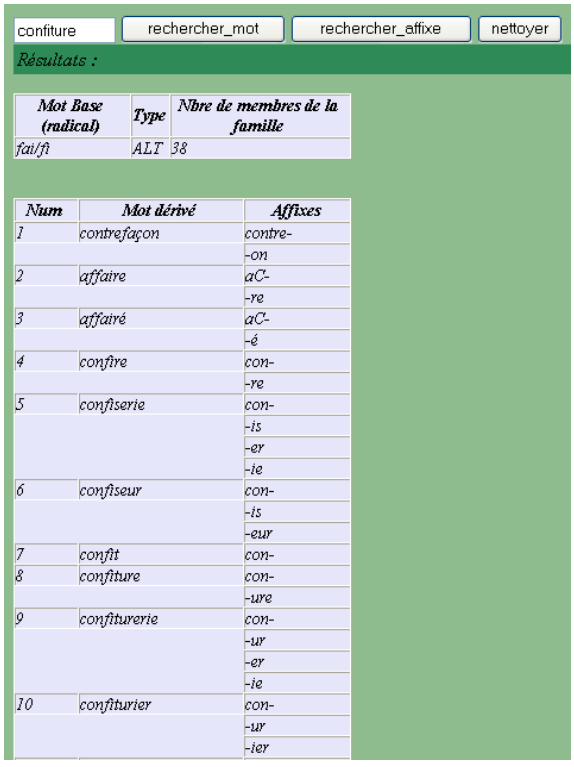
POLYMOTS propose deux types de fonctionnalités : la recherche par mots et la recherche par affixes.

4.1 Recherche par mots

Il est possible de rentrer une forme lexicale et d'en obtenir plusieurs informations, à savoir, son radical phonologique, le type de radical et la liste de mots partageant ce radical de base, c'est-à-dire l'ensemble des mots de la famille. Nous avons fait le choix d'écrire le radical phonologique et les mots dérivés selon l'orthographe conventionnelle. Deux raisons motivent ce choix: d'une part la lecture est plus aisée en écriture orthographique et d'autre part, il s'avère qu'il y a souvent un lien entre le radical phonologique et son écriture orthographique. Dans la grande majorité des cas, l'écriture est maintenue dans la construction dérivationnelle. L'écriture du français comprend en effet selon Doneux (2001), une part aléatoire et une part prédictible : l'écriture des mots « brun » et « cousin » est prédictible car le « u » s'entend dans « brune », et le « i » dans « cousine ». L'écriture du mot « second » est aléatoire car le « c » ne correspond pas à la sortie phonique attendue et ne peut être entendue en dérivation.

En l'état des travaux, les affixes ont été écrits selon l'orthographe. Par exemple, le préfixe inversif « in- » aura ainsi plusieurs graphies : im (impossible), in (inlavable), il (illisible). Un travail à venir devra également proposer des formes structurelles profondes pour les affixes.

Pour chaque mot dérivé, il est possible de visualiser sa structure morphologique, décomposée en préfixes et suffixes.


confiture | rechercher_mot | rechercher_affixe | nettoyer

Résultats :

Mot Base (radical)	Type	Nbre de membres de la famille
fai/fi	ALT	38

Num	Mot dérivé	Affixes
1	contrefaçon	contre-
		-on
2	affaire	aC-
		-re
3	affairé	aC-
		-é
4	confire	con-
		-re
5	confiserie	con-
		-is
		-er
		-ie
6	confiseur	con-
		-is
		-eur
7	confit	con-
8	confiture	con-
		-ure
9	confiturerie	con-
		-ur
		-er
		-ie
10	confiturier	con-
		-ur
		-ier

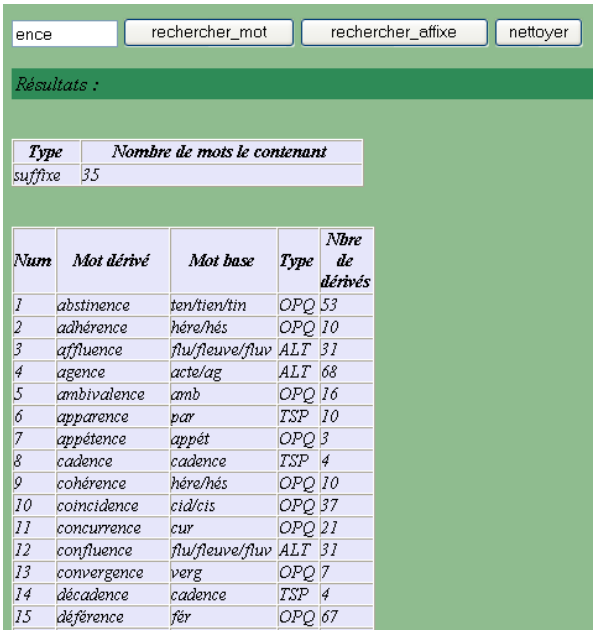
FIG 1. Résultats de la recherche pour le mot « confiture ».

La figure 1 montre le résultat pour les dix premiers mots[1] de la famille ayant comme radical phonologique « fai/fi » (pour une recherche avec « confiture »). L'alternant fait/fit est attesté dans le paradigme de la conjugaison. La graphie « C » dans le préfixe « aC » correspond à une consonne qu'on ne peut pas déterminer car elle dépend de la consonne du radical ; il s'agit du même préfixe avec une assimilation régressive, on n'est donc pas en mesure de spécifier la forme phonologique de « C » (exemple, « aC » avant la base « fai/fi » fera doubler le « f » ; dans « accusateur », avec la base « cause/cus », ce sera le « c » etc.).

4.2 Recherche par affixes

A partir d'un affixe donné, la base permet d'obtenir des renseignements sur sa productivité, c'est-à-dire, la liste de mots le contenant. Il est aussi possible de lancer une requête avec deux affixes ; par exemple pour "ana" et "ique" simultanément, on obtiendra « anarchique », « anatomique », « anamorphique », etc.

Le résultat donne, pour chaque entrée, son radical phonologique, le type de mot et le nombre de dérivés dans sa famille.


ence rechercher_mot rechercher_affixe nettoyer

Résultats :

Type	*Nombre de mots le contenant*
suffixe	*35*

Num	*Mot dérivé*	*Mot base*	*Type*	*Nbre de dérivés*
1	*abstinence*	*ten/tien/tin*	*OPQ*	*53*
2	*adhérence*	*hére/hés*	*OPQ*	*10*
3	*affluence*	*flu/fleuve/fluv*	*ALT*	*31*
4	*agence*	*acte/ag*	*ALT*	*68*
5	*ambivalence*	*amb*	*OPQ*	*16*
6	*apparence*	*par*	*TSP*	*10*
7	*appétence*	*appét*	*OPQ*	*3*
8	*cadence*	*cadence*	*TSP*	*4*
9	*cohérence*	*hére/hés*	*OPQ*	*10*
10	*coïncidence*	*cid/cis*	*OPQ*	*37*
11	*concurrence*	*cur*	*OPQ*	*21*
12	*confluence*	*flu/fleuve/fluv*	*ALT*	*31*
13	*convergence*	*verg*	*OPQ*	*7*
14	*décadence*	*cadence*	*TSP*	*4*
15	*déférence*	*fér*	*OPQ*	*67*

FIG 2. Résultats de la recherche pour l'affixe « -ence ».

La figure 2 montre les 15 premiers résultats de la requête avec le suffixe « -ence ».

A l'heure actuelle, la base contient plus de six cents affixes répertoriés (environ un quart de préfixes et trois quarts de préfixes), les plus productifs étant représentés dans le tableau suivant :

[1] L'ordre des mots est pour l'instant aléatoire. Dans la version à venir, les mots apparaîtront par ordre alphabétique.

Préfixes		Suffixes	
dé-	468	-er	1460
aC-	450	-ion	622
con-	206	-at	618
in-	198	-ement	594
a-	188	-age	291

TAB 2. Productivité des affixes plus fréquents.

4.3 Evaluation et perspectives

A partir du mois de mars 08, cette base sera évaluée à partir d'une plate-forme de travail d'orthophonistes, afin de valider la pertinence de ces nouvelles familles de mots dans le cadre d'exercies de remédiation.

Par ailleurs, comme les radicaux TSP et OPAK ont une orthographe identique dans toute la famille de mots concernés par ce radical, nous souhaiterions valider cet outil comme un outil pour apprendre l'orthographe des mots. Pour les radicaux ALT, un apprentissage explicite (didactique) devra être réalisé en milieu scolaire (principe de la lettre de rappel pour indiquer le marquage phonologique : "e" dans « européen » car « européenne » ; "u" dans « brun » car « brune » ; "a" dans « paix » car « pacification »…). Ceci se ferait en lien avec l'apprentissage du vocabulaire qui est inscrit dans les recommendations du BO (12 avril 2007, hors-série n°5): "certaines approches sont particulièrement fécondes pour structurer et augmenter le vocabulaire disponible. Ainsi en est-il de l'attention à la construction des mots qui permet d'accroître plus rapidement le vocabulaire disponible. L'enfant découvre le sens et la graphie des mots d'usage quotidien […]".

5 Conclusion

Dans cet article, nous avons presenté une base lexicale de huit mille mots français qui montre des regroupements par familles et permet de visualiser la décomposition des entrées lexicales en diachronie structurelle. Les entrées ont été manuellement analysées à partir de vingt mille mots communs extraits d'un dictionnaire de langue générale. L'approche adoptée nous a permis de proposer trois catégories de radicaux phonologiques, en tenant compte des caractéristiques morpho-phonologiques du radical. La distribution entre ces trois catégories donne trois classes équilibrées, bien que ce sont les mots de type alternant et opaque qui ont une productivité plus élévée (respectivement, 19 et 13, alors que les transparents en produisent une moyenne de 10).

Cela confirme que lors de la construction morphologique, le radical peut avoir ou non une signification. Dans le cas des radicaux OPAK, les affixes sont donc vecteurs de signification du mot. Ce constat contribue à une refléxion plus générale sur l'analyse morphologique et en particulier sur la notion d'affixe. Dans le cas des deux autres types de radicaux (TSP et ALT), le sens et la forme du radical assurent un continuum à travers tous les mots construits. Par exemple, un trait sémantique du radical "fil" s'entend bien dans le mouvement de rue nommé

"défilé". Ce concept de continuité sémantique est une perspective que nous souhaiterions poursuivre. La méthode est donc, d'après nous, validée ; reste à savoir si elle constitue un outil didactique (enseignant) et clinique (orthophoniste) pour l'apprentissage du vocabulaire et de l'orthographe. Des évaluations futures dans ce sens sont sur le point d'être mises en oeuvre. Quant à la base elle-même, nous avons prévu d'y inclure douze mille mots supplémentaires analysés, ainsi que d'autres fonctionalités de recherche (par type de mot, etc.). Une automatisation complète d'une telle analyse morphologique ne nous semble pas possible car chaque mot est une histoire sociale unique.

Références

BENVENISTE E. (1974). *Comment s'est formée une différentiation lexicale en français.* (258-271) in Problèmes de linguistique générale. Tome 2. Paris: Gallimard. 286 p.

CORBIN D. (1987). *Morphologie dérivationnelle et structuration du lexique*. Vol. 1. & 2, Coll. Linguistische Arbeiten, 193. Tübingen : Max Niemeyer.

DAL G., HATHOUT N., NAMMER F. (1999). Construire un lexique dérivationnel : théorie et réalisations. Actes de la Conférence *Traitement Automatique des Langues* (TALN 1999), 115-124. Cargèse, France.

JACQUEMIN C. (1997). Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches, Université de Nantes.

DONEUX J. L. (2001). *L'écriture du français* : Prédictibilité et Aléa. Texte présenté par Véronique Rey. Publications de l'Université de Provence, 200 p.

FRADIN B. (2003). *Nouvelles approches en morphologie*. Paris : PUF. 347 p.

GRUAZ C., HONVAULT R. (2007). Dictionnaire Synchronique des Familles dérivationnelles de mots français. Limoges : éditions Lambert-Lucas. 1300 p.

MOESCHLER J., AUCHLIN A. (2000) – *Introduction à la linguistique contemporaine* – Paris : Cursus, Armand Colin.

REY A. (1998). *Dictionnaire historique de la langue française*. Paris : Le Robert.

REY V., STOFFEL H. (2006). Propos sur la définition du mot : spécificité ou universalité ?, Actes de *SLE 2006*, Bremen (Allemagne).

REY-DEBOVE J. (1982) – *Le Robert Méthodique : Dictionnaire méthodique du français actuel* – Paris, Dictionnaires Le Robert.

REY-DEBOVE J. (2004) – *Le Robert Brio, Dictionnaire général et morphologique : analyse méthodique des mots* – Paris, Dictionnaires Le Robert.

SABATER C., REY V. (2005). De l'ortho-phonie à l'ortho-graphe : le cas de la dictée L2MA. *Glossa*, 92, 4-21.

SAVOY J. (1993). Stemming of French words based on grammatical categories. *Journal of American Society for Information Science*. 44 (1), 1-9.

TANGUY L., HATHOUT N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. Actes de la Conférence *Traitement Automatique des Langues* (TALN 2002), Nancy, France.

ZIEGLER J. C., JACOBS A. M., STONE G. O. (1996). *Statistical analysis of the bidirectional inconsistency of spelling and sound in French*. Behavior research Methods, Instruments and Computers, 28 (4), 504-515. Psychonomic Society Inc.