UNE BASE DE DONNÉES DES ENTRÉES ET SORTIES DANS LA NOMENCLATURE D'UN CORPUS DE DICTIONNAIRES : PRÉSENTATION ET EXPLOITATION

Résumé : Dans cet article sont présentées quelques fonctionnalités offertes par une base de données lexicologique que nous avons constituée à partir de la comparaison de trente dictionnaires. Il s'agit d'étudier quelques propriétés des articles entrés dans les Petit Larousse *et les* Petit Robert *parus depuis 1997, et quelques propriétés des articles sortis des mêmes ouvrages ainsi que du dernier* Dictionnaire de l'Académie française. *De cette étude se dégagent des informations sur l'évolution dans le temps des consignes rédactionnelles, qui permettent de constater des différences entre les millésimes successifs d'un même dictionnaire.*

INTRODUCTION

La base de données présentée dans cet article est le fruit d'une lecture comparée manuelle d'un corpus de dictionnaires, menée dans le cadre d'une thèse de doctorat. Cette lecture a commencé fin 2005 par la comparaison des versions papier des *Petit Larousse* 2005 et 2006. Elle s'est poursuivie jusqu'à aujourd'hui, en dehors du cadre du doctorat, et concerne à ce jour 30 dictionnaires : les *Petit Larousse* 1997 à 2010 (ce qui représente 14 volumes), les *Petit Robert* de la même période[1] (14 volumes), enfin les huitième (1932-1935) et neuvième éditions du *Dictionnaire de l'Académie française*, jusqu'au mot *mappemonde*, qui clôt le deuxième et dernier volume paru à ce jour. Cette lecture comparée est appelée à se poursuivre, de façon à continuer d'alimenter la base de données, en prenant en compte non seulement les dictionnaires les plus récents (à commencer par les millésimes 2011), mais aussi ceux parus avant 1996.

La lecture comparée de deux dictionnaires consécutifs consiste à relever le plus de différences possible (idéalement, toutes) dans le texte des

1. Les *Petit Robert* sont millésimés à partir de l'édition 2006, parue en 2005. Nous avons millésimé les éditions qui ne l'étaient pas en suivant cette logique.

deux éditions comparées, à tous les niveaux : nomenclature, définitions, orthographe, mise en pages, etc. Nous avons tenté de relever exhaustivement les données concernant la macrostructure, c'est-à-dire les entrées d'articles, les sorties d'articles, mais aussi les fusions et scissions d'articles. Pour valider les données observées mais surtout pour réparer les éventuels oublis après la comparaison deux à deux des millésimes 1997 à 2007 des *Petit Larousse* et *Petit Robert*, nous avons également comparé ces deux éditions éloignées de dix ans.

Notre méthode de lecture n'a donc de sens que si l'on compare deux éditions du même ouvrage, car les données relevées restent en nombre relativement restreint. À aucun moment il n'a été question de comparer un *Petit Robert* et un *Petit Larousse*, par exemple.

Le matériau constitué par ce relevé constitue un indicateur du travail des lexicographes. D'une édition à l'autre du dictionnaire, les lexicographes conservent inchangée la plus grande partie du texte. Lors d'une mise à jour, leur travail consiste à modifier le texte par une série d'opérations (comme l'ajout d'articles nouveaux) de façon à conserver son intégrité générale tout en lui apportant une touche de nouveauté et d'actualité. Lors d'une refonte, comme celle qui a abouti à la parution du *Petit Larousse* 1998, le travail lexicographique est plus fourni que lors de la mise à jour ; néanmoins l'essentiel du texte reste là encore inchangé. La comparaison représente donc un travail de reconstitution a posteriori de l'œuvre lexicographique.

Les données de la comparaison manuelle sont consignées sur des fiches en papier. L'essentiel des données a été saisi dans une base de données informatisée, qui en permet une meilleure exploitation. Enfin, une partie de cette base de données a été publiée sur Internet[2]. Le but de cet article est de présenter quelques fonctionnalités de la base de données, à travers un parcours parmi quelques-unes des données qu'elle recèle.

1 ARTICLES ENTRÉS DANS LES DICTIONNAIRES

1.1 Fiches de la base de données

Dans la période 1997-2010, 2 983 articles sont entrés dans les *Petit Larousse* et 2 166 dans les *Petit Robert* (désormais PL et PR). Nous avons consigné ces 5 149 articles dans une section de la base de données. Sur chacune des 5 149 fiches réalisées, de nombreuses informations complémentaires sont renseignées. Voici deux fiches tirées de la base de données :

2. On trouve à l'adresse <http://orthogrenoble.net/page-de-camille-club-orthographe-grenoble.html> les listes des articles entrés dans les 30 dictionnaires évoqués, mais aussi les listes des articles sortis, ainsi que quelques autres ressources.

Entrée	Gosser
Dictionnaire	PR
Millésime	2007
Page	1167
Lignes	5
Cat. gram.	v.tr. et v.intr.
Pluriel	-
Marque diatopique	régionalisme (Canada)
Marque diastratique	Familier
Marque diachronique	-
Marque diatechnique	-
Étymologie	mot de l'Ouest (Vendée, Poitou), d'origine inconnue
Première attestation	1842
Type d'article	Normal

Entrée	3. casse ➔ 1. cassier
Dictionnaire	PL
Millésime	1998
Page	184
Lignes	1
Cat. gram.	n.f.
Pluriel	-
Marque diatopique	-
Marque diastratique	-
Marque diachronique	-
Marque diatechnique	Botanique
Étymologie	-
Première attestation	-
Type d'article	Renvoi

Sous chaque entrée sont listées plusieurs informations qui constituent une source secondaire de renseignements sur les dictionnaires, informations destinées à établir des analyses tant chiffrées que qualitatives. Des éléments

non encore saisis (comme le nombre de définitions, la liste des renvois, etc.) pourront être ajoutés sur chaque fiche au fur et à mesure de l'alimentation de la base de données. Pour l'heure, les données essentielles sont saisies : l'entrée elle-même bien sûr, le dictionnaire qui accueille cette nouvelle entrée, la page précise qui reçoit le mot nouveau, sa catégorie grammaticale, son pluriel s'il est indiqué, ce qui est le cas général pour les noms composés, les marques présentes dans l'article – diatopiques[3] (Belgique, Sud-Ouest, etc.), diastratiques (familier, soutenu, etc.), diachroniques (vieux, ancien, etc.), diatechniques (zoologie, médecine, etc.) –, puis l'étymologie telle qu'elle est donnée dans l'article, la date de première attestation (que l'on trouve dans les *Petit Robert*), enfin le type d'article. Mis à part ce dernier renseignement, toutes les données récoltées sont directement et objectivement recopiées dans les dictionnaires, sans traitement.
L'exploitation informatique de ces données fournit des renseignements intéressants sur les mots nouveaux qui entrent dans les dictionnaires.

1.2 Nombre d'articles entrés dans les dictionnaires

Le premier renseignement que l'on peut exploiter est le nombre d'articles entrés dans chaque millésime, indication du programme lexicographique et de son évolution au fil des ans. Le tableau suivant rassemble les données chiffrées à ce sujet.

millésime	nombre d'articles entrés dans le PL	nombre d'articles entrés dans le PR
1998	1463	80
1999	37	51
2000	287	53
2001	113	127
2002	135	108
2003	103	142
2004	56	86
2005	304	67
2006	103	73
2007	114	513
2008	65	460
2009	109	218
2010	94	188

Les chiffres n'étant pas ronds, on suppose déjà que le nombre d'articles à ajouter dans chaque édition n'est pas strictement programmé. Cela dit, l'ampleur de la mise à jour ou de la refonte souhaitée par les lexicographes détermine une estimation du nombre d'articles à ajouter.

3. F. J. Hausmann, 1989, p. 651.

Dans la colonne du nombre d'articles entrés dans le PL, les orientations récentes de ce dictionnaire se dessinent clairement : depuis la refonte de 1998 et l'ajout de près de 1 500 articles nouveaux, les mises à jour se suivent avec une certaine régularité. En 2000 et 2005, ces mises à jour étaient importantes. À l'inverse, en 1999, très peu d'articles nouveaux ont fait leur entrée, comme si l'alimentation du PL en néologismes passait par un trou d'air après la refonte. Une nouvelle refonte semble se profiler.
Dans les PR, les mises à jour se sont succédé dans la période 1998-2006. Quant à la refonte annoncée par l'éditeur dans le millésime 2007, elle est d'un genre particulier. Il s'agissait surtout de mettre en place une nouvelle maquette, bien plus souple que les précédentes, autorisant par la suite des mises à jour conséquentes. En 2007, dans le cadre d'une refonte, assez peu d'articles ont été ajoutés ; mais les mises à jour suivantes sont de grande ampleur. Ajoutons que si l'on compte le nombre de lignes que représentent ces articles ajoutés (en laissant de côté les lignes ajoutées par l'ajout de définitions dans des articles préexistants), la mise à jour de 2008 est plus grande que celle de 2007, avec 1 675 lignes ajoutées contre 1 409. En tous les cas, la série des PR 2007 et suivants marque un tournant dans l'histoire éditoriale de ce dictionnaire.

1.3 Nombre de lignes des articles ajoutés

La prise en compte du nombre de lignes de chaque article nouveau apporte des renseignements complémentaires.
Ainsi, on constate que les articles ajoutés au PR2007 sont bien plus courts (2,74 lignes chacun en moyenne) que ceux ajoutés dans les millésimes 2003 et suivants, 2008 et suivants (de 3,4 à 3,84 lignes en moyenne). Dans les PL, au cours de la période 1997-2010, la tendance est à l'augmentation progressive du nombre moyen de lignes dans les articles ajoutés : on passe des valeurs basses de 2000 et 2003 (2,39 et 2,52 lignes de moyenne) aux valeurs hautes de 2008 et 2010 (3,55 et 3,49 lignes). On peut lire dans l'évolution de ces chiffres un changement dans les pratiques rédactionnelles, qui serait à préciser.
Enfin, en répartissant les articles ajoutés par nombre de lignes, on observe que la majorité des articles ajoutés dans les dictionnaires sont très courts. Le tiers des articles entrés dans le PL entre 1997 et 2010 mesurait 2 lignes ; et la moitié des articles entrés dans le PR durant la même période mesurait 3 ou 4 lignes[4]. Quant aux articles de 6 lignes et plus, qui sont pourtant fréquents dans les dictionnaires, ils sont rarement ajoutés : ils ne représentent 8 % des articles entrés dans le PR et 5 % des articles entrés dans le PL. Cela témoigne du fait que les néologismes ne nécessitent qu'un petit espace de description, peu d'entre eux étant polysémiques. C'est plutôt au moment où des sens nouveaux sont ajoutés dans des articles préexistants, donc quand il est fait

4. On ne peut pas ici comparer directement le PL et le PR, car les lignes sont très inégales entre les deux ouvrages. Pour rendre cette description possible, il faudrait compter le nombre de signes de chacun des articles ajoutés.

état de leur évolution dans le sens de la polysémie, que le volume de ceux-ci augmente, éventuellement au-delà des 6 lignes.
À l'inverse de la tendance générale, quelques articles nouveaux sont très longs, à l'instar de *GEIE* (14 lignes, entré dans le PL 1998) ou *interopérabilité* (9 lignes, PR 2009). On peut penser que l'espace accordé à ces mots par les lexicographes est proportionnel à l'importance qu'ils attribuent à leur description.

1.4 Rubrique étymologique et date de première attestation

La rubrique étymologique est facultative dans la microstructure des articles du PL. Dans le PR, elle est supposée être obligatoire. Nous avons calculé le taux de remplissage de la rubrique étymologique des articles entrés dans chaque millésime des deux dictionnaires.
Les données concernant le PL nous renseignent, une fois de plus, sur l'évolution des pratiques rédactionnelles dans ce dictionnaire. Alors que parmi les articles entrés dans les millésimes 1998 à 2005, seuls 21 à 33 % possédaient une rubrique étymologique, ce pourcentage augmente considérablement à partir de 2006, en avoisinant régulièrement les 50 %, jusqu'à atteindre 60 % dans le millésime 2010. D'édition en édition, le taux de rubriques étymologiques parmi les articles nouveaux est donc susceptible de varier du simple au triple. On ne peut que saluer le fait que les auteurs du PL aient choisi, ces dernières années, d'améliorer les descriptions lexicographiques par davantage de données étymologiques.
Dans les PR, le taux de remplissage de la rubrique étymologique oscille entre 72 et 95 %. Cela vient du fait que dans notre base de données, nous avons traité 1° les articles-renvois, 2° les articles dédiés à des éléments de composition grecs et latin, 3° les sous-articles contenus à l'intérieur d'un article principal (comme *truandage* contenu dans l'article *truander*), comme des articles à part entière. Or il n'est pas prévu de rubrique étymologique dans la microstructure de ces articles particuliers (ou alors à titre exceptionnel). Reste que, dans la plupart des articles, une rubrique étymologique est bien présente, précédée d'une date de première attestation à laquelle nous allons maintenant nous intéresser.
Cette date de première attestation, qui renseigne sur le statut néologique ou non des articles ajoutés, est susceptible de prendre diverses formes. Si elle est le plus souvent une date fixe (par exemple « 1985 » pour le verbe *ringardiser*), elle peut aussi être moins précise : des indications telles que « XIX^{e} s. », « av. 1990 », « début XX^{e} », « répandu 1990 », etc., pour être exploitées, doivent être interprétées. Ainsi, dans l'optique de calculer la date moyenne de première attestation de l'ensemble des articles entrés dans un millésime donné du PR, nous avons procédé à un double calcul, celui des limites d'une fourchette temporelle. Soit par exemple cet échantillon d'articles entrés dans le PR2010, avec leur date de première attestation telle qu'elle est indiquée dans ce dictionnaire.

entrée	1re attestation
performeur, euse	1985 ; 1984 au Québec
réseauter	1997 ; autre sens 1985
trie	XVIIIe, dans le Bordelais
tue-l'amour	v. 1980
uranite	1790

À partir de ces données, nous avons calculé une limite inférieure et une limite supérieure de la date moyenne de première attestation, en attribuant une double valeur aux dates complexes : à l'indication « XVIIIe », nous avons attribué la valeur minimale « 1701 » et la valeur maximale « 1800 » ; à l'indication de l'article *réseauter*, nous avons attribué la valeur minimale « 1985 » et la valeur maximale « 1997 », etc. Les valeurs ayant servi au calcul sont exposées dans le tableau suivant.

entrée	1re attestation : valeur minimale	1ère attestation : valeur maximale
performeur, euse	1984	1985
réseauter	1985	1997
trie	1701	1800
tue-l'amour	1975	1985
uranite	1790	1790
moyenne	1887	1911

Il apparaît dans ce tableau que la date moyenne de première attestation de l'échantillon considéré est située dans la fourchette 1887-1911.
Nous avons donc mené ce calcul sur l'ensemble des articles entrés dans les PR depuis 1997. Il en résulte un âge moyen des articles entrés dans chaque millésime (qui tient compte du décalage d'un an entre la date de parution du dictionnaire et son millésime). Les résultats sont présentés dans le tableau suivant.

millésime	moy. minimale	moy. maximale	âge moyen des entrées
PR1998	1959	1962	35-38 ans
PR1999	1951	1958	40-47 ans
PR2000	1941	1952	47-58 ans
PR2001	1957	1962	38-43 ans
PR2002	1958	1961	40-43 ans
PR2003	1953	1961	41-49 ans
PR2004	1938	1947	56-65 ans
PR2005	1939	1947	57-65 ans

millésime	moy. minimale	moy. maximale	âge moyen des entrées
PR2006	1958	1963	42-47 ans
PR2007	1882	1894	112-124 ans
PR2008	1919	1926	81-88 ans
PR2009	1909	1921	87-99 ans
PR2010	1923	1931	78-86 ans

Ces chiffres doivent être analysés avec prudence, et ce pour trois raisons. Premièrement, ils ne sont pas objectivement tirés des dictionnaires, mais obtenus par un calcul qui, bien que prudent, reste améliorable. Deuxièmement, il arrive que la date de première attestation ne soit pas mentionnée dans un article qui entre dans le dictionnaire, soit que cette date soit inconnue (c'est le cas par exemple pour *frigolite*, entré dans le PR2008 avec la mention « date inconnue »), soit que cet élément de la microstructure ne soit pas renseigné, comme c'est souvent le cas dans les sous-articles. Troisièmement, les dates de première attestation sont susceptibles d'être révisées lors des mises à jour, notamment lorsque de nouvelles attestations plus anciennes sont découvertes ; or nous n'avons tenu compte que des dates indiquées au moment de l'entrée des articles, en ignorant l'évolution future de ces articles. Néanmoins, ces chiffres apportent des éléments de comparaison nouveaux entre tous ces PR successifs. En l'occurrence, on peut segmenter la période étudiée en deux sous-périodes. Dans les éditions 1998 à 2006, les lexicographes ont certes procédé à des mises à jour de moindre envergure (cf. 1.2. ci-dessus), mais l'âge moyen des mots ajoutés était bas : il s'agissait donc, dans la plupart des cas, de néologismes. En revanche, le millésime 2007 marque une rupture éditoriale : à partir de cette refonte, on constate que l'âge moyen des nouveaux mots est beaucoup plus élevé. Les 500 articles ajoutés cette année-là, ainsi que ceux entrés dans les millésimes qui ont suivi, ne sont pas tous des néologismes. En d'autres termes, le programme lexicographique s'est ouvert à partir de la refonte à des ensembles de mots non néologiques ; il s'agit en l'occurrence de mots de la francophonie et de mots techniques. On constate en effet entre autres exemples, en manipulant la base de données, que parmi les 340 articles consacrés à des régionalismes et francophonismes ajoutés à la nomenclature entre 1998 et 2010, seulement 6 l'ont été avant 2007. De même, parmi les 88 articles marqués « chimie » entrés dans les PR 1998 à 2010, 75 ont été ajoutés en 2010. Or les francophonismes, mots de la chimie, et autres groupes de mots ajoutés massivement et ponctuellement dans la nomenclature ont des dates de première attestation plus reculées que les néologismes.

Le calcul de l'âge moyen des mots entrés dans les PR rend mal compte de la diversité des données. Il nous faut signaler que les mots qui entrent dans ce dictionnaire sont pour certains très anciens (ainsi *channe*, attesté dès 1150 et entré en 2007, ou *calendaire*, attesté au XIII^{e} siècle et entré dans l'édition 1998), pour d'autres extrêmement récents, recensés dès leur apparition. *Virophage*, attesté en 2008, entre dans le PR2010. Autant dire qu'ils

sont entrés immédiatement, si l'on précise que ce dictionnaire est paru en juin 2009, que son texte était bouclé quelques mois auparavant, et que les lexicographes ont commencé à sélectionner les nouveaux mots à insérer dès l'été 2008.

On pourrait multiplier les investigations dans la section « entrées » de notre base de données. Mais la section consacrée aux articles sortis des dictionnaires est non moins instructive.

2 ARTICLES SORTIS DES DICTIONNAIRES

2.1 Fiches de la base de données

Dans la période 1997-2010, 4 899 articles sont sortis du PL, et 87 du PR. Nous avons également relevé les articles sortis entre la huitième et la neuvième édition du *Dictionnaire de l'Académie française* (désormais DAF), qui sont au nombre de 486 entre *a* et *mappemonde*, dernier mot de la neuvième édition paru en volume relié.
Les fiches de la section consacrée aux articles sortis des dictionnaires sont très proches de celles présentées précédemment.

entrée	finet, ette
dictionnaire	DAF
année	1932 ➔ 2000
page	546
page	137
cat. gram.	adj.
marque diatopique	-
marque diastratique	familier
marque diachronique	peu usité
marque diatechnique	-

entrée	botswanais, e
dictionnaire	PL
millésime	1997 ➔ 1998
page	152
page	146
cat. gram.	adj. et n.
marque diatopique	-
marque diastratique	-
marque diachronique	-
marque diatechnique	-

Cette fois, pour éviter une ambiguïté, nous indiquons deux dates de publication successives d'un même dictionnaire. Lorsque nous indiquons que l'article *botswanais* est « sorti en 1998 », cela signifie qu'il était présent dans le millésime 1997 et absent du 1998. En conséquence, deux numéros de page sont indiqués sur chaque fiche, le second correspondant à la page sur laquelle le mot supprimé se serait trouvé s'il avait été maintenu.

2.2 Note sur le nombre d'articles sortis

Le nombre d'articles qui sortent des dictionnaires, énoncé plus haut, peut surprendre, et surtout l'écart entre PL et PR. Cette différence est surtout due au choix de la période 1997-2010. En effet, 4 354 des 4 899 articles disparus du PL durant cette période sont sortis lors de la refonte de 1998, dont le mot d'ordre était clairement un renouvellement général de la nomenclature. Rien de comparable dans les PR 1997 à 2010 ; mais il y a fort à parier que les données à tirer de l'observation de la deuxième refonte de ce dictionnaire, parue en 1993, sont similaires.

2.3 Choix des mots sortis du PL1998

La sortie de 4 354 articles du PL en 1998 (soit environ quatre par page) pose question. On se demande notamment sur quels critères certains mots ont été supprimés plutôt que d'autres.
Certaines suppressions sont systématiques. Ainsi, tous les symboles d'éléments chimiques (comme *Ag*, *Tc* ou *Pb*), auxquels était automatiquement consacré un article en 1998, sont supprimés. De façon similaire, de nombreux gentilés ont été supprimés : *roubaisien*, *sarthois*, *maltais*, etc. Ensuite, les mots marqués « Vx », « vieilli » ou « ancien » semblent avoir été des cibles prioritaires, des candidats à la sortie. 356 des articles sortis sont dans ce cas. Enfin, beaucoup de noms de métiers obsolètes sont sortis de la nomenclature : *pompeur*, *libelliste*, *apprêteur* ou encore *bandagiste* disparaissent des colonnes du PL. Avec l'ajout simultané de près de 1 500 articles lors de cette refonte, la nomenclature évolue considérablement, et notamment dans le sens d'une modernisation.
Il est également probable que les lexicographes se soient inspirés du premier volume du *Dictionnaire de l'Académie française*, neuvième édition, paru en 1992, dans lequel 303 articles ont été supprimés. Nous avons en effet repéré 22 mots en commun entre les deux listes, c'est-à-dire 22 mots supprimés de la neuvième édition du DAF en 1992, puis supprimés du PL en 1998 (ainsi *archipresbytéral*, *conglutiner*, *embourrer*, etc.). Quant aux 281 autres articles sortis du DAF, ils ne figuraient pour la plupart déjà plus dans le PL1997.

CONCLUSION

L'intérêt de la constitution d'une base de données des allées et venues dans la nomenclature de nos dictionnaires réside essentiellement dans une meilleure connaissance des matériaux lexicographiques, qui sont familiers car souvent utilisés mais relativement impénétrables. Il s'agit en l'occurrence d'aller plus loin que l'information diffusée dans les médias par les éditeurs lors de

la parution d'un nouveau millésime, information qui consiste bien souvent en une liste incomplète de mots nouveaux, laquelle cache toutes les autres opérations pratiquées régulièrement dans le texte du dictionnaire.

Par ailleurs, la comparaison que nous avons tentée entre les millésimes successifs d'un même dictionnaire, appuyée sur des données chiffrées, apporte de précieuses informations, souvent inattendues, sur le travail des lexicographes et sur l'évolution de la ligne éditoriale et du programme lexicographique de ces ouvrages.

Camille MARTINEZ
Université de Strasbourg
et Université de Cergy-Pontoise
CNRS UMR 7187 (LDI)

BIBLIOGRAPHIE

HAUSMANN, F. J. 1989. « Die Markierung im allgemeinen einsprachigen Wörterbuch : eine Übersicht », dans Hausmann, Reichmann, Wiegand & Zgusta, article 53, p. 649-657, Berlin / New York : De Gruyter.

HAUSMANN, F. J., REICHMANN, O., WIEGAND, H. E., ZGUSTA, L. 1989. *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*, tome 1, Berlin / New York : De Gruyter.

MARTINEZ, C. 2009. *L'évolution de l'orthographe dans les* Petit Larousse *et les* Petit Robert *1997-2008 : une approche généalogique du texte lexicographique*, thèse de doctorat, 778 p., Université de Cergy-Pontoise.