Verbum XXXIV, 2012, nº 2, 195-212

UNE APPROCHE DISTRIBUTIONNELLE POUR L'ANALYSE DES COMPOSÉS NÉOCLASSIQUES

Marine LASSERRE
CLLE-ERSS, CNRS & Université de Toulouse

Fabio MONTERMINI
CLLE-ERSS, CNRS & Université de Toulouse

RÉSUMÉ

Cet article présente une analyse des mots construits en français et en italien à partir des éléments latins -cida/-cidium *('tuer'). Ces lexèmes sont intéressants en ce qu'ils mettent en jeu, dans les deux langues, des propriétés sémantiques et formelles différentes, notamment avec un syncrétisme en français entre le sens agentif / instrumental et le sens abstrait (actionnel). Les mots construits par ce procédé se répartissent en trois classes principales, qui correspondent à trois schémas qui se sont développés dans les deux langues à des époques différentes. L'hypothèse fondamentale qui est développée ici est que chaque nouvelle formation peut être analysée en fonction de sa proximité avec l'une de ces classes. La méthode utilisée s'inspire de l'analyse distributionnelle en sémantique et vise à caractériser le sens des mots complexes non pas sur la base de leurs propriétés intrinsèques, mais sur la base des contextes syntaxiques qu'ils partagent.*

ABSTRACT

In this paper, we present an analysis of the words constructed in French and in Italian from the Latin elements -cida/-cidium *('kill'). These lexemes are interesting because they display, in the two languages, different semantic and formal properties, with in particular a syncretism in French between the agentive / instrumental and the abstract (actional) meanings. The words constructed by means of this process can be divided into three main classes corresponding to three schemes which emerged in the two languages at different periods. The main hypothesis investigated is that every new formation can be analysed according to its proximity with one of these classes. The method adopted is inspired by distributional analysis which is current in semantics and aims at characterising the meaning of complex words not on the basis of their intrinsic properties, but on the basis of the syntactic contexts they share.*

1. INTRODUCTION

> In 620 ore di discussioni, i 600 delegati dell'O.N.U. si sono trovati d'accordo una volta sola, per coniare una parola nuova. La nuova parola è *genocidio* e significa : sterminio di un raggruppamento umano.
>
> 'En 620 heures de discussion, les 600 délégués de l'ONU se sont trouvés d'accord seulement une fois, pour créer un nouveau mot. Le nouveau mot est *génocide* et il signifie : extermination d'un regroupement humain'.
>
> *La Stampa*, 3 novembre 1948

Dans cet article, nous présentons une étude préliminaire que nous avons menée sur un groupe de lexèmes construits par un procédé de composition néoclassique en français et en italien, en nous intéressant à la manière dont ces mots complexes construisent leur sens. Nous partons du présupposé que l'étude de la construction du sens des composés néoclassiques peut nous renseigner sur la construction du sens des lexèmes construits en général. Nous présentons en particulier une analyse des mots construits en français et en italien à partir des éléments latins *-cida*/*-cidium*, à leur tour liés au verbe CAEDO ('tuer'). Ces lexèmes sont intéressants en ce qu'ils mettent en jeu, dans les deux langues, des propriétés sémantiques et formelles différentes, notamment avec un syncrétisme en français, qui est dû aux transformations phonologiques que cette langue a subies au cours des siècles, entre le sens agentif / instrumental (*Mark Chapman est l'homicide de John Lennon*) et le sens abstrait (actionnel : *l'homicide de John Lennon a eu lieu le 8 décembre 1980*). Nous suggérons que les mots construits par ce procédé peuvent se répartir en trois classes distinctes, qui correspondent à trois schémas qui se sont développés dans les deux langues à des époques différentes et nous considérons que chaque nouvelle formation peut être analysée en fonction de sa proximité ou de sa distance avec l'une de ces classes. La méthode que nous utilisons s'inspire des analyses distributionnelles qui sont courantes en sémantique (cf. Lenci 2008 pour un aperçu). Elle vise à caractériser le sens des mots complexes non pas sur la base de leurs propriétés intrinsèques, mais sur la base des contextes syntaxiques qu'ils partagent. L'article est organisé comme suit : en 2., nous présentons quelques préliminaires théoriques à l'analyse que nous proposons, en mettant l'accent sur la distinction, en morphologie constructionnelle, entre règle et schéma ; en 3., nous discutons du statut des composés néoclassiques et des éléments qui les constituent, en les comparant aux lexèmes 'canoniques' et aux affixes ; en 4., nous présentons les données des deux langues considérées et les corpus sur lesquels nous avons travaillé ; 5. présente l'analyse effectuée, la méthodologie employée et quelques résultats ; 6., enfin, contient quelques conclusions.

2. ENJEUX THÉORIQUES

Pendant longtemps, au moins depuis Aronoff (1976), la construction de lexèmes complexes a été envisagée en termes de règles. Aujourd'hui on préfère considérer qu'il s'agit de schémas, ou patrons, ce qui permet de dépasser certaines des rigidités des règles, telles qu'elles étaient établies dans la morphologie lexicaliste issue de la linguistique générative. Regarder la construction de mots en termes de schémas permet de contourner les problèmes posés par la nature essentiellement sélective et déterministe des règles et de rendre compte de la dynamique lexicale de la langue et de sa diversité. Les schémas permettent, par exemple, d'abandonner l'hypothèse compositionnelle pour la sémantique des mots construits, selon laquelle le sens d'un mot complexe est entièrement prédictible, étant le résultat d'une opération sémantique réalisée par la règle de formation de mots sur le sens d'un lexème base. En effet, il est facilement observable que des mots complexes peuvent être attestés avec un grand nombre de sens différents, parfois connectés, mais pas nécessairement. Pour rendre compte de ce fait, certains linguistes adoptent ce que Booij (2010 : 78) appelle une « approche monosémique » : ils attribuent à chaque règle de formation de mots un sens très général, abstrait, qui permet de rendre compte du sens de chaque item et laissent à des facteurs extra-morphologiques (pragmatiques, stylistiques, etc.) le soin de déterminer la signification spécifique de chaque lexème. On obtient ainsi une liste de sens spécifiques découlant d'un sens général sous-spécifié. Ainsi, Corbin (1989, 1991) oppose clairement un sens attesté et un sens prédictible. Le sens prédictible est « a compositional function of the morphological structure of a constructed word » (Corbin 1989 : 32), il est le résultat de l'opération sémantique réalisée par une règle sur une base donnée. Selon Corbin, pour un mot construit, « les éventuelles distorsions entre sa structure morphologique profonde et sa forme apparente ou entre son sens prédictible et son sens lexicalisé doivent être explicables par des mécanismes réguliers permettant de passer de l'un(e) à l'autre » (Corbin 1991 : 17). A cette approche monosémique, nous préférons l'« approche polysémique régulière » que lui oppose, par exemple, Booij (2010 : 78) : les différentes interprétations que peuvent recevoir les outputs d'un processus de formation de mots sont distinctes même si potentiellement liées.

Nous estimons, de plus, que, contrairement à ce qui est souvent fait dans les travaux de morphologie dérivationnelle, pour rendre compte de la construction du sens d'un lexème construit la prise en compte du contexte est essentielle. Selon son contexte d'apparition, un mot construit par un processus de formation de lexèmes apparemment unique peut avoir différents sens. Une donnée objective dont nous disposons pour étudier le sens de chaque mot construit est sa distribution. L'approche distributionnaliste, ancrée dans une tradition qui remonte à la linguistique structurale des années 1950 (en particulier aux travaux de Z. Harris), est aujourd'hui développée

notamment en sémantique et en Traitement Automatique du Langage (cf. Lenci 2008 ; Sahlgren 2008 pour un aperçu). Ses principes sont fondés sur l'idée qu'il existe une corrélation entre la similarité du sens et la similarité de la distribution de deux mots. En partant de cette hypothèse, nous postulons qu'on peut accéder au sens d'un lexème complexe en prenant en compte sa distribution. Si un processus de construction de lexèmes apparemment unique donne lieu à différentes interprétations au niveau de ses outputs, on peut en rendre compte en étudiant les distributions respectives de ces outputs.

3. LA NATURE DES FORMANTS NÉOCLASSIQUES

Comme nous l'avons dit, le travail que nous présentons se base sur l'analyse de deux séries de composés néoclassiques en français et en italien. Quelques critères récurrents sont généralement mis en avant pour définir les éléments entrant en jeu dans ce type de formations, qui existent dans la plupart des langues européennes de culture (cf. par exemple Amiot & Dal 2007) : ils sont originellement des lexèmes en grec ancien et/ou en latin mais ne sont plus autonomes dans les langues modernes ; ils servent en général à former des lexèmes appartenant au vocabulaire scientifique ou technique ; enfin, il y a le plus souvent une voyelle (graphiquement *-o-* pour les formants d'origine grecque et *-i-* pour ceux d'origine latine) qui lie un formant néoclassique à l'autre constituant du composé dans lequel il entre en jeu. Les traitements qui en sont proposés dans la littérature sont variés. Pour certains, les éléments néoclassiques auraient conservé un sens lexical et seraient donc à différencier des affixes. Corbin (1999) parle ainsi d'*archéoconstituants* qui, au même titre que les affixes, sont des unités infralexicales mais dont la principale distinction se fait au niveau du sens : sens lexical pour les premiers, instructionnel pour les seconds. En ce sens, ils seraient des éléments de composition au même titre que les lexèmes natifs. L'existence de formants néoclassiques directement affixés (ex. *hydrique*) tendrait à confirmer cette hypothèse. Villoing (2012) souligne de plus que la plupart des formants néoclassiques peuvent être reliés à des lexèmes français et peuvent donc être analysés comme des bases allomorphiques ou supplétives. Il nous semble cependant que différencier les formants néoclassiques des éléments de composition natifs suppose qu'on prête aux locuteurs une conscience étymologique que peu d'entre eux possèdent. Par contre, si on s'intéresse au rôle que jouent ces formants dans la construction des lexèmes, on s'aperçoit qu'ils correspondent à la manifestation formelle d'opérations sémantiques et syntaxiques faites sur des lexèmes qui ne sont pas différentes dans la substance de celles effectuées par l'affixation canonique : leur spécification phonologique consiste en l'adjonction d'une séquence stable à droite ou à gauche du thème d'un lexème, ils établissent une relation entre deux catégories et ont une instruction sémantique, voire plusieurs, qui sont parfois

prédictibles. Quelques travaux récents sur des langues différentes du français suggèrent que la ligne de partage n'est pas entre affixes et éléments de composition néoclassique (et a fortiori lexèmes) : d'une part, les frontières entre ces classes sont floues, d'autre part, lorsque les classifications des exposants de schémas de constructions de lexèmes se basent sur des critères explicites, les classes identifiées sur une base étymologique ne tiennent pas. Lüdeling *et al.* (2002), par exemple, proposent l'utilisation des traits SELECTING et BOUNDNESS qui permettent une distinction des exposants de construction de mots et des lexèmes qui n'a rien à voir avec leur origine ou leur caractère plus ou moins savant, au moins en allemand (cf. aussi Bauer 2005 : 105 sur l'anglais). Si les exposants de schémas de constructions de lexèmes ont souvent été analysés comme étant quasi exclusivement des affixes, la distance entre exposants et lexèmes est aujourd'hui considérée comme plus floue. Ainsi, Booij (2010 : 61-62) prend l'exemple du nom néerlandais HOOFD 'tête' qui est utilisé comme exposant dans des constructions comme *hoofd*-X_N et peut prendre des interprétations différentes mais reliées, conduisant à des sous-schémas hiérarchisés : le sens 'X de grande importance' a donné lieu d'un côté à celui 'X au sommet de la hiérarchie' et de l'autre à celui 'principal X'. En français aussi nous trouvons des lexèmes par ailleurs autonomes qui rentrent de manière récurrente dans des constructions au même titre que des éléments qui sont incontestablement des affixes, par exemple THÉRAPIE (*théâtrothérapie*, *curiethérapie*, *vertébrothérapie*, cf. Amiot & Dal 2008), DÉPENDANT (*alcoolodépendant, caféinodépendant*), etc. Chacun de ces lexèmes partage avec les formants néoclassiques la capacité de former des composés 'allogènes'[1], c'est-à-dire des composés ayant la tête sémantico-syntaxique à droite, contrairement aux composés natifs du français. Ils donnent également lieu à des séries de lexèmes formés selon le même schéma constructionnel, exactement comme le font un grand nombre de formants néoclassiques.

Pour toutes ces raisons, nous proposons de considérer les éléments de composition néoclassique, au moins ceux du type étudié ici, comme des exposants de schémas de construction de lexèmes, au même titre que les affixes et que certains lexèmes, comme THÉRAPIE. Ces exposants peuvent éventuellement être repartis en sous-classes, mais vraisemblablement sur la base de propriétés différentes que leur étymon ou leur caractère plus ou moins savant. Par ailleurs, pour ce qui est de la construction du sens, nous ne considérons pas que la distinction entre sens instructionnel et sens lexical est pertinente et opérationnelle dans tous les cas, puisqu'il existe des affixes qui véhiculent un sens que l'on pourrait définir 'lexical' (par exemple les évaluatifs) à côté de nombreux lexèmes (cf. par exemple le cas de HOOFD ci-dessus) qui, dans certaines constructions, ont incontestablement un sens instructionnel. En revanche, et c'est ce que nous souhaitons montrer dans ce

[1] Nous devons cette définition à Michel Roché (c.p.).

qui suit, la construction du sens des lexèmes complexes dans lesquels ces exposants interviennent est sensiblement semblable.

4. LES DONNÉES

Dans ce travail, nous nous intéressons plus précisément aux lexèmes construits, en français et en italien, au moyen des continuateurs des mots latins en *-cida*/*-cidium*[2]. Il s'agit, en latin, de mots composés présentant, en deuxième position, le verbe CAEDO ('battre, rompre, tuer') – qui possède un thème apophonique *cid* pour les formes préfixées ou composées (cf. Brucale 2012 : 104-105) – et qui peuvent être des noms [+humain] de la première déclinaison désignant des agents (*homicida* 'homicide$_{[+h]}$', *tyrannicida* 'tyrannicide$_{[+h]}$'), ou bien des noms neutres [–humain] de la deuxième déclinaison désignant des actions (*homicidium* 'homicide$_{[–h]}$', *tyrannicidium* 'tyrannicide$_{[–h]}$')[3]. Comme le montrent les exemples cités, le nom régi (l'élément de gauche) se réfère à l'argument interne de l'action désignée par le verbe. En ce sens, les composés latins en question sont exactement parallèles aux composés VN dans les langues romanes modernes (mis à part, bien entendu, l'ordre des constituants), qui désignent le plus souvent des agents (*rabat-joie*), ou des instruments (*ouvre-boîte*), et plus rarement des actions (*baise-main*) (cf. Villoing 2009 sur le français). Du point de vue sémantique, il est très intéressant de noter que, à l'intérieur des différents sens possibles pour le verbe CAEDO[4], deux composés seulement, parmi ceux que nous avons trouvés dans des dictionnaires, sélectionnent celui de 'couper, fendre' (*lapi(di)cida* 'tailleur de pierres' et *lignicida* 'bûcheron'), alors que tous les autres sélectionnent le sens de 'tuer' et ont comme premier élément un nom [+humain] ou [+animé] qui désigne la victime de l'action indiquée par le composé en entier. Ces faits montrent que, déjà en latin, s'il s'agissait bien de composés, ceux-ci étaient soumis aux contraintes sémantiques et de sélection des bases[5] qui sont typiques des schémas constructionnels, en démontrant, une fois de plus, que même la construction du sens des lexèmes composés ne se fait pas de manière purement additive.

---

2 Désormais, nous indiquons avec *-cid-* l'ensemble des exposants issus de *-cida* et *-cidium* latins lorsque nous voulons nous référer sans distinction aux deux langues, avec *-cide* uniquement l'élément français et avec *-cida/io* l'ensemble des deux éléments italiens ; de plus, nous ajoutons les étiquettes [+h] ou [–h] aux mots français en *-cide*, uniquement lorsque cette distinction est pertinente pour notre discussion.

3 Notons, sans nous étendre sur ce point, pour des raisons de place, que les composés en question présentent les mêmes problèmes d'analyse que les composés synthétiques dans d'autres langues, par exemple en anglais (*troublemaker* 'fauteur de troubles', *whistle-blower* 'souffleur', cf. Ackema & Neeleman 2010), mais aussi – plus rarement – en français (cf. *francocentrique*) (sur le latin, cf. Brucale 2012 : 98-99 ; sur les composés synthétiques en général cf., entre autres, Melloni & Bisetto 2010).

4 Gaffiot (1934) cite, pour ce verbe, six sens différents.

5 Pour l'emploi de la notion de 'base' y compris pour la composition cf. Roché (2009).

Dans les langues romanes modernes, et dans quelques autres langues, ces composés ont donné naissance à deux séries parallèles de lexèmes : des noms [+humain] qui désignent un agent et qui fonctionnent également comme des adjectifs (fr. *homicide*$_{[+h]}$, it. *omicida*) et des noms [–humain] qui désignent une action (fr. *homicide*$_{[–h]}$, it. *omicidio*). En français, en particulier, les changements phonologiques que la langue a subis au cours du temps ont neutralisé toute distinction formelle entre les deux types de lexèmes, différence qui s'est maintenue dans d'autres langues, comme l'italien ou l'espagnol. Le résultat est que nous avons, en italien, une différenciation qui est à la fois formelle, catégorielle et sémantique (*-cida* pour les noms [+humain], *-cidio* pour les noms [–humain][6]), alors qu'en français nous avons une variété d'emplois différents pour des lexèmes qui partagent la même séquence finale /sid/. Du point de vue sémantique, même s'il existe quelques lexèmes dans lesquels *-cid-* a un sens différent (cf. fr. *denticide*, 'qui s'ouvre en forme de dents'), le sens majoritairement représenté est celui de 'tuer' (donc de 'tueur' pour les mots [+humain] et de 'meurtre' pour ceux [–humain])[7]. Une comparaison entre italien et français est donc intéressante à plusieurs égards : elle nous permet notamment d'observer la différente distribution des propriétés formelles, catégorielles et sémantiques que ces lexèmes ont héritées de leurs ancêtres latins.

L'analyse que nous proposons ici s'appuie sur deux listes de lexèmes se terminant par *-cide* en français et par *-cida/io* en italien ayant la signification générale de 'tuer'. Les deux listes contiennent des lexèmes attestés et fréquents ainsi que des néologismes, voire des occasionalismes. Ils ont été recueillis, pour le français, à partir du *Grand Robert*, du *TLFi* et de Google Ngrams[8], et pour l'italien à partir du *Grande dizionario italiano dell'uso* (*Gradit*) et du corpus CORIS[9]. Nous avons élargi ces bases par des recherches aléatoires et en cherchant les éléments issus de *-cid-* en combinaison avec les bases qui nous semblaient les plus plausibles. Enfin, des recherches croisées effectuées sur le Web à partir des corpus initiaux nous ont permis d'augmenter ultérieurement le nombre d'occurrences dans

6 Notons, de plus, que les mots [+humain] en *-cida* appartiennent, lorsqu'ils sont masculins, à une classe flexionnelle marginale et qui ne comprend quasiment que des lexèmes d'origine grecque ou latine – celle des noms en *-a* au singulier et *-i* au pluriel.

7 En italien, il existe aussi quelques mots en *-cidio* dans lesquels ce dernier élément a le sens de 'tomber' (cf. *stillicidio* 'égouttement') et est lié non pas au verbe latin CAEDO, mais à CADO. Cette ambiguïté, par ailleurs, existait déjà en latin. Notons aussi que, à la différence du latin, en italien l'hypothèse que les noms en *-cidio* soient dérivés par suffixation, par exemple de ceux en *-cida* n'est pas tenable, puisqu'il n'existe plus de suffixe actionnel *-io* qui soit productif en italien.

8 http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

9 Le CORIS (environ 130 millions de mots) est un des plus grands corpus de langue écrite actuellement disponibles en italien (Rossini Favretti 2000) et peut être consulté à partir de l'adresse http://corpora.dslo.unibo.it/coris_ita.html.

chacun des deux, pour un total de 306 *-cide* en français, 344 *-cida* et 345 *-cidio* en italien (la plupart desquels, naturellement, sont construits à partir de la même base).

A partir de ces deux bases de données, nous pouvons faire une série d'observations préliminaires, en nous concentrant principalement ici sur les aspects sémantiques. Les premières attestations de mots en *-cid-* dans les deux langues, tous directement empruntés au latin, ne comportent que des bases non autonomes ayant une forme latine (fr. *fratricide*, *régicide*). Cependant, comme le montre l'exemple de *tyrannicida/ium*, cité ci-dessus, même en latin cette construction pouvait prendre comme bases des lexèmes non natifs, une possibilité qui s'est développée et est aujourd'hui massivement exploitée dans les deux langues. A part les cas de lexèmes bien intégrés, actuellement les locuteurs ont la possibilité de choisir entre une base savante et une base 'native', comme le montrent les couples it. *avicidio* / *uccellicidio* (< *uccello* 'oiseau'), fr. *arachnicide* / *araignicide.* Base 'savante', cependant, ne veut pas nécessairement dire latin, puisque, comme nous l'avons observé, il existe aujourd'hui un nombre important de lexèmes en *-cid-* où le premier élément est d'origine grecque, et même des triplets (latin, grec, natif), comme it. *equi(ni)cidio*, *ippocidio*, *cavallicidio* (< *cavallo* 'cheval'). Du point de vue sémantique, il est possible de dégager au moins trois classes principales, qui correspondent à des ensembles de lexèmes de catégories différentes et à des types de bases préférentiels. Le Tableau 1 illustre la situation pour le français et l'italien (les classes sémantiques des bases sont données en ordre descendant de préférence).

	Français	Italien	Cat.	Sens	Types de bases
i.	homicide	omicida	$N_{[+h]}$ / A	'tueur de X'	$N_{[+humain]}$ (*clodocide*) $N_{[+animé]}$ (*chaticide*) $N_{[+abstrait]}$ (*tempslibricide*)
		omicidio	$N_{[-h]}$	'meurtre de X'	
ii.	insecticide	insetticida	$N_{[-h]}$ / A	'produit qui sert à tuer X'	$N_{[+animé]}$ (*raticide*) $N_{[-animé]}$ (*herbicide*) $N_{[+humain]}$ (*crétinicide*)
iii.	génocide	genocidio	$N_{[-h]}$	'tuerie à grande échelle de X'	$N_{[+abstrait]}$ (*genricide*) $N_{[+collectif]}$ (*arabicide*) $N_{[+animé]}$ (*pigeonicide*[10])

Tableau 1. – Distribution sémantico-catégorielle des dérivés en *-cid-* en français et en italien

Le type (i) est celui qui, comme nous l'avons vu, était représenté en latin et est, sans surprise, le plus ancien aussi bien en italien qu'en français. Tous les mots en *-cid-* qui sont apparus dans les deux langues jusqu'au XIX[e] siècle

10 Cf. (2c) ci-dessous.

relèvent de ce groupe sémantique et sont, pour la plupart des emprunts directs ou des calques du latin[11]. Le type (ii) apparaît dans le langage médical au XIX^e^ siècle et sert à construire d'abord des adjectifs et successivement des noms, dans le sens de 'substance qui tue X'. En (1) nous indiquons les premières attestations de lexèmes de ce groupe sémantique (adjectifs et noms) en français et en italien que nous avons trouvées grâce à Google books :

(1)

	A	N
f	la qualité **insecticide** du camphre [*Bulletin de pharmacie et des sciences accessoires*, 1814]	la double propriété d'adoucissante et d'**insecticide** de ce liquide [*Séance publique de la Société Royale de Médecine de Toulouse*, 1834]
i	bolo **vermicida** [*Biblioteca italiana*, 1818]	il primo di questi è un **insetticida** [*Nuovo giornale de' letterati*, 1832]

Quant au type (iii), il possède un prototype bien identifié, à savoir le mot *génocide*, qui a été d'abord créé en anglais par le juriste polonais Raphael Lemkin dans son ouvrage *Axis Rule in Occupied Europe*, publié en 1944 (dans lequel il propose aussi, comme alternative, *ethnocide*). Il s'agit, semble-t-il, de la première création 'hybride' dans laquelle *-cid-* est précédé d'une base grecque et non pas latine, et qui a donné lieu à un schéma qui présente aujourd'hui une certaine disponibilité, en donnant légitimité, entre autres, à *-o-* comme élément de liaison mineur, mais possible, pour les mots en *-cid-*. Du point de vue diachronique, nous avons donc trois schémas qui sont apparus et se sont développés à des époques différentes, bien que de manière assez parallèle dans les deux langues. Chacun des schémas identifiés se distingue des autres, nous l'avons vu, par ses propriétés sémantiques, catégorielles (cf. le Tableau 1) et par les types de bases préférés, du point de vue non seulement de leur sens, mais aussi de leur origine et de leur statut dans la langue : bases latines pour le type (i), natives pour le type (ii) et grecques pour le type (iii). Cette propriété aussi nous pousse à assimiler ce type de constructions à des dérivations affixales plus typiquement reconnues comme telles, puisque c'est un fait avéré depuis longtemps que des opérations affixales différentes peuvent sélectionner les thèmes auxquels ils s'appliquent selon leur caractère plus ou moins savant (y compris dans le cas de thèmes qui sont manipulés pour leur donner une forme pseudo-savante, cf. Plénat 2008 ; Roché 2011a : 109-110).

---

[11] Les premiers mots construits en français et en italien semblent être, respectivement, *uxoricide* (1531 selon *Le Grand Robert*) et *vericida* ('menteur', XVI^e^ siècle, selon le *Gradit*).

Les lexèmes en *-cid-* constituent donc un terrain d'analyse intéressant en ce qu'un procédé apparemment unique et uniforme correspond, en réalité, à un ensemble de schémas dérivationnels différents qui, à cause de leur émergence et de leur développement, s'articulent de manière différente avec le reste du lexique : avec un mot leader défini (sur concept de mot leader, cf. par exemple Rainer 2003, Roché 2011b : 86-87) qui sert d'attracteur pour le type (iii), de façon plus distribuée pour les types (i) et (ii).

5. VERS UNE ANALYSE DISTRIBUTIONNELLE DES PROCÉDÉS MORPHOLOGIQUES

Les trois types sémantiques que nous avons identifiés de manière informelle en 3 ne doivent pas être vus comme des classes fermées mais plutôt comme des pôles d'attraction dans lesquels les lexèmes observés peuvent rentrer de manière plus ou moins claire. D'une part, il existe un petit nombre de lexèmes que l'on peut difficilement faire rentrer dans l'un des trois types (fr. *légicide* 'qui ne respecte pas la loi', it. *cervellicida* '< *cervello* 'cerveau' 'abrutissant') ; d'autre part, des mots construits à partir de la même base peuvent appartenir à plusieurs des classes identifiées[12] :

(2) a. Qui a déjà commis des **pigeonicides** ?
[http://www.jeuxvideo.com/forums/1-51-13364722-1-0-1-0-qui-a-deja-commis-des-pigeonicides.htm]

b. C'est une structure **pigeonicide** à double circonvolution broyeuse.... C'est pas énorme ça comme concept ???
[http://www.youtube.com/all_comments?threaded=1&v=fc_E2EeifY0]

c. un **pigeonicide** est organisé par le conseil municipal sous le nom barbare de "dépigeonnisation" ça fait froid dans le dos, non ?
[http://actualites.forum.orange.fr/messages/index/40198/benoit-xvi-discrimination.html]

Si l'on suit l'hypothèse distributionnaliste, on peut faire l'hypothèse que, lorsqu'un schéma dérivationnel construit plusieurs types sémantiques de dérivés, des lexèmes construits relevant du même type sont plus à même de partager des contextes syntaxiques identiques. Comme nous l'avons indiqué en 1, cette idée correspond à une des hypothèses de base de l'analyse distributionnelle en sémantique. Bien que quelques travaux récents prennent en compte des données issues de la morphologie constructionnelle dans des analyses distributionnelles (cf. Guevara 2009 ; Lazaridou *et al.* 2013), ce domaine reste relativement peu exploré. De plus, les travaux cités se sont plutôt concentrés sur la similarité entre les lexèmes comportant le même affixe,

[12] Nous laissons de côté la question, qu'il serait intéressant d'étudier plus dans le détail et sur un nombre plus grand de données, de savoir si les exemples de (2) constituent un seul lexème ou trois lexèmes différents.

alors que notre perspective est plutôt celle d'explorer la variation sémantique des lexèmes appartenant à la même série.

Quant à la méthodologie employée, nous avons considéré les contextes d'apparition des lexèmes les plus fréquents de nos corpus, en supposant que ceux-ci sont les plus susceptibles d'être des attracteurs dans une opération de construction de lexèmes par analogie. Pour le français, nous avons pris les 20 lexèmes les plus fréquents sur Google[13]. Pour l'italien nous avons pris tous les mots en *-cida*/*-cidio* ayant une fréquence supérieure à 5 dans le CORIS[14]. Pour chacun de ces lexèmes nous avons extrait une liste de contextes syntaxiques fréquents, en nous servant, pour le français, de Frantext, des *voisins du Monde* et des *voisins de Wikipédia*[15] et, pour l'italien, de CORIS. Lorsque nous parlons de 'contexte' nous nous référons non seulement aux mots contigus dans les corpus considérés, mais également aux unités reliées syntaxiquement (par exemple par des relations sujet-prédicat, prédicat-objet ou nom-adjectif). Les *voisins du Monde* et *de Wikipédia* permettent déjà d'extraire des contextes de ce type. Pour les autres ressources nous avons procédé à des extractions manuelles. Pour l'italien, nous avons considéré tous les contextes avec une fréquence supérieure à 10 dans le CORIS (31 types de contextes au total). Pour le français, les ressources employées ne nous permettant pas de mesurer la fréquence d'apparition des contextes au sein d'un corpus, nous avons croisé les distributions les plus fréquentes données par chaque ressource, puis avons restreint les contextes à ceux qui nous paraissaient dévolus à un seul type (par exemple, un contexte comme « coupable de X », fréquent pour les types (i) et (iii), ne nous permettrait pas de faire une quelconque différence entre ces deux types). Dans le Tableau 2 nous donnons l'ensemble des contextes que nous avons testés parmi les plus fréquents pour chacun des types identifiés dans le Tableau 1 ci-dessus[16].

---

13 Le lexème *suicide* était le plus fréquent (27 650 000 occurrences le 22/10/12) mais n'a pas été retenu pour l'analyse, se comportant, selon nous, de manière différente par rapport aux autres lexèmes du corpus.

14 En ce qui concerne l'italien, nous avons considéré, pour *-cida*, la somme des occurrences de la forme se terminant en *-a* et de celle en *-e* (correspondant, respectivement, au singulier – masculin et féminin – et au féminin pluriel, ex. *omicida*, *omicide*), et pour *-cidio* uniquement la forme se terminant en *-o* (*omicidio*). Les formes se terminant en *-i*, en effet, sont des masculins pluriels ambigus entre *-cida* et *-cidio* (en d'autres termes, *omicidi* est aussi bien le pluriel (masculin) de *omicida* que celui de *omicidio*).

15 Les deux dernières ressources, ainsi que des détails, sont disponibles à l'adresse http://redac.univ-tlse2.fr/applications/index.html.

16 Dans le Tableau 2, nous indiquons les lexèmes qui co-occurrent le plus souvent avec les mots en *-cid-*, leur ordre par rapport à ceux-ci, ainsi que la relation syntaxique (qui peut être indiquée par une préposition ou par une étiquette grammaticale comme O(bjet)). Les différences qui s'observent dans les contextes identifiés pour le français et l'italien, y compris dans des cas que nous nous attendrions de retrouver dans les deux langues,

	Français	Italien	
	-cide	-cida	-cidio
i.	X_(in)volontaire commettreO_X geste_X	furia_X follia_X raptus_X volontà_X lotta_X guerra_X conflitto_X maniaco_X	X_colposo tentato_X X_volontario accusaDI_X duplice_X accusareDI_X serieDI_X commettereO_X
ii.	produit_X utilisationDE_X utiliserO_X	attività_X usoDI_X impiegoDI_X resistenzaA_X	
iii.	perpétrerO_X Xcidaire		perpetrare_O X

Tableau 2. – Contextes les plus fréquents pour les trois types de dérivés en *-cid-*

A partir des contextes identifiés, nous avons vérifié, pour la totalité des mots en *-cid-* qui apparaissaient dans nos listes, leurs contextes d'apparition sur le Web, en nous servant du moteur de recherche Google. Pour chacun des contextes nous avons cherché l'ensemble des formes fléchies lorsqu'il s'agissait de noms ou adjectifs, et un échantillon des formes fléchies (infinitif, 3SG/PL PRES IND, participe passé) lorsqu'il s'agissait de verbes.

En ce qui concerne la démarche employée, elle vise principalement à identifier des critères pour classer des nouvelles formations dans les types sémantico-formels qui sont associés à un schéma constructionnel. Cela ne peut se faire qu'en ayant accès à un nombre important de néologismes ou de constructions occasionnelles, et actuellement seul le Web peut fournir ce type de données, en vertu de la quantité énorme de texte qu'il contient. Le nombre de néologismes contenus dans un corpus, y compris de très grandes dimensions, est insuffisant pour avoir une idée ne serait-ce que vague de la compétence morphologique active des locuteurs. A titre d'exemple, en ce qui concerne l'italien, le corpus CORIS (130 millions de mots au total) contient 12 mots en *-cida* et 8 mots en *-cidio* qui n'apparaissent pas dans le *Gradit*, alors que notre base de données (qui n'est certainement pas exhaustive) en contient, respectivement, 256 et 312 qui n'apparaissent dans aucune des deux ressources. Certes, les difficultés liées aux analyses linguistiques réalisées sur le Web, et en particulier à l'utilisation de celui-ci comme un

proviennent certainement de la différence entre les ressources utilisées pour les déterminer.

corpus, sont bien connues (cf. entre autres Lüdeling *et al.* 2007 ; Kilgariff 2007 ; Hathout *et al.* 2008). Dans l'attente d'avoir à disposition une ressource quantitativement comparable au Web qui n'en présente pas les inconvénients, il nous semble que, spécifiquement pour des analyses sur le lexique, cet instrument est aujourd'hui incontournable lorsqu'on veut s'intéresser à la créativité morphologique. Bien entendu, il faut utiliser quelques précautions ; en particulier, il est clair que de pures analyses quantitatives sur Google sont risquées, voire impossibles, mais des analyses qualitatives (qui observeraient par exemple l'existence vs. la non existence d'un certain lexème ou d'un certain contexte) ou des analyses quantitatives moins fines (par exemple qui prendraient en compte les ordres de grandeur plutôt que les nombres absolus, ou bien les proportions relatives d'occurrences de couples de lexèmes) nous semblent possibles et utiles.

Le Tableau 3 donne le nombre de lexèmes qui partagent les contextes compatibles avec chacun des types, en distinguant, pour l'italien, les dérivés qui ont un sens agentif / instrumental (*-cida*) de ceux qui ont un sens abstrait (*-cidio*).

Français		Italien				Types		
-cide		-cida		-cidio				
Nb	% sur le total	Nb	% sur le total	Nb	% sur le total	i	ii	iii
53	17,32%	50	14,49%	159	46,08%	X	–	–
50	16,34%	6	1,74%	–	–	–	X	–
19	6,21%	–	–	–	–	X	X	X
18	5,88%	–	–	4	1,15%	X	–	X
15	4,90%	5	1,45%	–	–	X	X	–
14	4,58%	–	–	1	0,28%	–	–	X
10	3,27%	–	–	–	–	–	X	X
127	41,50%	283	82,27%	177	51,3%	–	–	–

Tableau 3. – Nombre de lexèmes partageant les mêmes contextes en français et en italien

Pour chacune des deux langues il existe un nombre important de lexèmes en *-cid* qui ne donnent pas de résultats utiles. En effet, d'une part, beaucoup des nouvelles formations que nous avons recueillies n'ont que très peu d'occurrences sur Google, ce qui réduit les chances de les retrouver avec les mêmes contextes que d'autres, et d'autre part, nous n'avons pris en considération qu'un nombre limité de contextes ; élargir la liste présentée au Tableau 2, voire conduire une analyse sur la totalité des contextes observables pour chacun des lexèmes de notre corpus, produirait certainement des résultats plus fiables. De plus, il est clair qu'une classification sémantique

plus fine des mots qui forment les contextes serait aussi nécessaire ; en français, par exemple, si *geste* est une des collocations les plus fréquentes avec un *-cide* de type (i), il pourrait être intéressant de considérer aussi des mots synonymes ou sémantiquement proches de *geste*, comme *acte*, *action*, etc., éventuellement en codant les différents contextes selon leur type sémantique ([+humain], [+concret], etc)[17].

Les types qui se dégagent le plus clairement, dans les deux langues, sont ceux qui désignent des agents / instruments ((i) et (ii)). Le type (i) regroupe la quasi totalité des bases $N_{[+humain]}$ et, parmi celles-ci, toutes celles désignant des liens de parenté. Dans ce type, *-cid-* sélectionne également des bases $N_{[+animé]}$ désignant des animaux, mais aussi des bases $N_{[+abstrait]}$. Ces derniers mots construits utilisent le plus souvent des procédés métaphoriques ou métonymiques, comme on le voit dans les exemples de (3)[18] :

(3) a. Cependant, et je t'en parle en connaissance de cause, elle peut se révéler etre l'arme la plus « **amouricide** » qui n'ait jamais existé... en effet, si elle sert ta possessivité, la jalousie peur remettre la confiance ke tu as en ton(ta) copain/copine alors meme qu'il n'y aurait aucune doute à avoir envers l'etre aimé !
[http://www.france-jeunes.net/discut.php ?tid=222&tid2=16142]

b. È passato giusto un decennio e i sogni dei trent'anni hanno dovuto scontrarsi con la realtà, spesso violenta e **sognicida**.
'Il ne s'est passé qu'une décennie et les rêves des trente ans ont dû se heurter à la réalité, souvent violente et tueuse de rêves'.
[http://www.teverenotizie.it/articolo-stampa,711.html]

Dans le type (ii), le procédé dérivationnel en question sélectionne en majorité des N désignant des animaux, des plantes ou des maladies, le plus souvent considérés comme nuisibles ou potentiellement nuisibles. Dans ce cas aussi, le sens 'être nuisible' peut être dérivé métaphoriquement :

(4) … adesso che navigo mi si sono ampiati i miei orizzonti e constato che di idioti ce ne sono a iosa. Non ho la bacchetta magica per farli sparire e neppure un "**idioticida**", seppur chimico.
'… maintenant que je surfe sur le Web mes horizons se sont élargis et je constate qu'il y a des tas d'idiots. Je n'ai pas de baguette magique pour les faire disparaître ni même un « idioticide », même chimique'.
[http://vioadriano.blog.tiscali.it/2004/12/01/il_santo_natale_ipocrisia_ferocе_del_cattolico_cristiano____1710842-shtml/]

Contrairement aux deux types précédents, nos analyses permettent plus difficilement de tracer les contours du type (iii). Les mots leader *génocide* et *genocidio* présentent moins de contextes qui leur sont réservés. Ainsi, par

17 Nous remercions un relecteur anonyme pour cette suggestion.

18 Dans tous les exemples tirés du Web nous avons conservé l'orthographe et la ponctuation des originaux.

exemple en français, seuls trois lexèmes ont comme seul contexte utile, parmi ceux considérés, perpétrerO_X : *linguocide, tutsicide* et *avunculicide* (< *avuncul-* 'oncle maternel'). Quant au contexte Xcidaire, s'il sert à désigner un agent, il est relativement répandu dans les lexèmes de notre corpus, certainement par l'influence de *suicidaire*. Il peut servir tout de même à interpréter une occurrence comme ayant le sens de 'tuerie à grande échelle de X' :

(5) Et vous voulez nous dératiser??? Bande de **raticidaires**! [http://fr.answers.yahoo.com/question/index?qid=20090221164135AAbnBNM]

Le type (iii) est donc plus facilement identifiable par l'observation du type de base, puisqu'il privilégie des noms collectifs ou désignant des concepts abstraits (*glottocide*, *ethnocide*, *genricide*, etc.), que par une analyse uniquement distributionnelle, sans doute aussi à cause du nombre global de lexèmes qu'il forme, plus faible par rapport aux deux autres.

Si l'on se place du côté des bases, certaines d'entre elles se prêtent davantage à des interprétations diverses. C'est le cas pour les bases $N_{[+animé]}$ référant à des animaux. Sans surprise, l'analyse distributionnelle que nous avons réalisée montre, pour le français, que 70% des bases désignant un animal[19] sont sélectionnées dans les contextes définis pour le type (ii) et 50% le sont de manière exclusive. Cependant, il ressort également que 45% de ces bases sont utilisées dans des constructions attribuées au type (i), dont 25% de manière exclusive. Contrairement à ce qui se passe pour le type (ii), l'animal désigné par la base est ici perçu individuellement, et l'acte de le tuer comme unique. Les exemples en (6) montrent cette différence de traitement : en (6a) c'est l'identité de la souris en tant qu'animal de compagnie qui est saillante, alors qu'en (6b) ce sont ses propriétés d'animal potentiellement nuisible qui le sont :

(6) a. Quand à ma souris blanche je l'ai bien assassinée mais c'était un **souricide** involontaire, monsieur le Juge, je vous le jure. [http://www.femiboard.com/threads/nouveau-jeu-2v%C3%A9rit%C3%A9s-et-un-mensonge.40924/page-33]

b. je n'aime pas trop utiliser de **souricide**, les bêtes souffrent terriblement... mieux vaut utiliser une tapette. [http://www.graines-et-plantes.com/index.php?forum=jardin-jardinage&question=campagnols-sous-plancher]

Le contexte syntaxique peut également nous orienter sur l'interprétation de la base qui peut être un collectif ou un individuel. En (7a), où le nom en *-cide-* est le complément du verbe *commettre*, la base est interprétable

19 Ces chiffres ne s'appuient ici que sur les 179 lexèmes qui ont été trouvés avec au moins un des contextes syntaxiques recherchés.

comme pouvant référer à un individu, alors qu'en (7b), où il est le complément du verbe *perpétrer*, elle est interprétable comme un nom collectif :

(7) a. Il est vrai que les **arabicides** ne sont pas nouveaux. Ce qui est nouveau, à mon avis, c'est la fréquence nouvelle des **arabicides** commis par des policiers.
[http://blogs.mediapart.fr/blog/naja/010709/devant-cette-mort-la-serons-nous-aussi-habitues-resignes-soumis-qu-nous-le-dem]

b. Il accuse le FRODEBU de s'être allié au PALIPEHUTU pour perpétrer un « **tutsicide** ». Même si l'auteur reconnaît qu'au Burundi des Tutsi et des Hutu sont morts uniquement à cause de leur appartenance ethnique. [http://repositories.lib.utexas.edu/bitstream/handle/2152/4119/2304.pdf?sequence=1]

6. CONCLUSION

Le but principal de l'analyse que nous avons proposée était d'identifier des instruments pour classer, sémantiquement et formellement, les lexèmes issus d'un certain procédé constructionnel. Les classements de ce type, où par exemple tous les lexèmes comportant le même préfixe ou le même suffixe sont répartis en classes et/ou sous-classes sémantiques sont fréquents. Cependant, ces répartitions sont faites dans la plupart des cas sur des significations abstraites, souvent les plus plausibles ou celles qui sont répertoriées dans les dictionnaires, et consistent à attribuer aux différentes classes identifiées des étiquettes censées représenter des catégories sémantiques générales. Nous considérons que toute classification sémantique, y compris du lexique construit, ne peut faire abstraction, au contraire, des contextes dans lesquels un certain lexème peut être et est employé. Ce qui compte, dans ce cas, pour la caractérisation sémantique d'un lexème, n'est pas tant l'étiquette sémantique abstraite que l'on peut lui attribuer, que sa relation avec d'autres lexèmes qui sont plus ou moins proches de lui. Ce que nous proposons, comme mesure possible, est la distance calculée sur les contextes syntaxiques partagés. Cette conception est également plus compatible avec une vision du lexique dans laquelle les unités qui le constituent ne se distinguent pas en vertu de propriétés discrètes, mais plutôt en vertu de leur distance les unes par rapport aux autres, les unités les plus semblables constituant des regroupements qui peuvent fonctionner comme des pôles d'attraction. A son tour, cette attraction peut prendre des formes différentes, certains de ces regroupements pouvant présenter un mot leader clair (comme *génocide*), d'autres un ensemble plus diffus (comme les types *homicide* et *insecticide*).

La méthode que nous avons employée est certes à améliorer, notamment en ce qui concerne les données sur lesquelles conduire l'analyse et sur la caractérisation des contextes à prendre en compte, mais il nous semble que les résultats préliminaires qu'elle nous permet d'obtenir nous encouragent à

poursuivre sur la même voie, en prenant en compte plus de paramètres et en envisageant de l'élargir à des ensembles plus larges et diversifiés de mots construits.

BIBLIOGRAPHIE

ACKEMA P., NEELEMAN A. (2010). The role of syntax and morphology in compounding. In : S. Scalise, I. Vogel (eds), *Cross-disciplinary Issues in Compounding*. Amsterdam / Philadelphia : Benjamins, 21-36.

AMIOT D., DAL G. (2007). Integrating neoclassical combining forms into a lexeme based morphology. In : G. Booij, L. Ducceschi, B. Fradin, E. Guevara, A. Ralli, S. Scalise (eds), *Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5) Fréjus 15-18 September 2005*. Bologna : Università degli Studi di Bologna, 323-336.

AMIOT D., DAL G. (2008). Composition néoclassique en français et ordre des constituants. In : D. Amiot (éd.), *La composition dans une perspective typologique*. Arras : Artois Presses Université, 89-113.

ARONOFF M. (1976). *Word-Formation in Generative Grammar*. Cambridge, MA : MIT Press.

BAUER L. (2005). The borderline between derivation and compounding. In : W.U. Dressler, D. Kastovsky, O.E. Pfeiffer, F. Rainer (eds), *Morphology and its Demarcations*. Amsterdam / Philadelphia : Benjamins, 97-108.

BOOIJ G. (2010). *Construction Morphology*. Oxford : Oxford University Press.

BRUCALE L. (2012). Latin compounds. *Probus* 24, 93-117.

CORBIN D. (1989). Form, structure and meaning of constructed words in an associative and stratified lexical component. In : G. Booij, J. van Marle (eds), *Yearbook of Morphology 1989*. Dordrecht : Springer, 31-54.

CORBIN D. (1989). Introduction. La formation des mots : structures et interprétations. *Lexique* 10, 7-30.

CORBIN D. (1989). Pour une théorie sémantique de la catégorisation affixale. *Faits de langue* 14, 65-77.

GAFFIOT F. (1934). *Dictionnaire latin français*. Paris : Hachette [http://www.-lexilogos.com/latin/gaffiot.php].

Gradit : DE MAURO T. (ed.) (1999). *Grande dizionario italiano dell'uso*. Torino : Utet.

Grand Robert : *Grand Robert de la langue française*. Paris : Robert.

GUEVARA E. (2009). Compositionality in distributional semantics : Derivational affixes. Poster présenté au workshop *Words in Action : Interdisciplinary Approaches to Understanding Word Processing and Storage*, Pise, 11-14 octobre 2009.

HATHOUT N., MONTERMINI F., TANGUY L. (2008). Extensive data for morphology : using the World Wide Web. *Journal of French Language Studies* 18.1, 67-85.

KILGARIFF A. (2007). Googleology is bad science. *Computational Linguistics* 33.2, 147-151.

LAZARIDOU A., MARELLI M., ZAMPARELLI R., BARONI M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In : *Proceedings of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*. East Stroudsburg, PA : ACL.

LENCI A. (2008). Distributional semantics in linguistic and cognitive research. A foreword. *Rivista di linguistica* 20.1, 1-30.

LÜDELING A., EVERT S. BARONI M. (2007). Using Web data for linguistic purposes. In : M. Hundt, N. Nesselhauf, C. Biewer (eds), *Corpus Linguistics and the Web*. Amsterdam / New York : Rodopi, 7-24.

LÜDELING A., SCHMID T., KIOKPASOGLOU S. (2002). Neoclassical word formation in German. In : G. Booij, J. van Maarle (eds), *Yearbook of Morphology 2001*. Dordrecht : Springer, 253-283.

MELLONI C., BISETTO A. (2010). Parasynthetic compounds. Data and theory. In : S. Scalise, I. Vogel (eds), *Cross-disciplinary Issues in Compounding*. Amsterdam / Philadelphia : Benjamins, 199-217.

PLÉNAT M. (2008). Le thème L de l'adjectif et du nom. In : J. Durand, B. Habert, B. Laks (éds), *Congrès Mondial de Linguistique Française – CMLF 2008*. Paris : Institut de Linguistique Française, 1613-1626.

RAINER F. (2003). Semantic fragmentation in word-formation : the case of Spanish *-azo*. In : R. Singh, S. Starosta (eds), *Explorations in Seamless Morphology*. New Delhi : Sage, 197-211.

ROCHÉ M. (2009). Pour une morphologie *lexicale*. *Mémoires de la Société de Linguistique de Paris. Nouvelle série* 17, 65-87.

ROCHÉ M. (2011a). Base, thème, radical. *Recherches linguistiques de Vincennes* 39.1, 95-134.

ROCHÉ M. (2011b). Quel traitement unifié pour les dérivations en *-isme* et en *-iste* ? In : M. Roché, G. Boyé, N. Hathout, S. Lignon, M. Plénat, *Des unités morphologiques au lexique*. Paris : Hermès-Lavoisier, 69-143.

ROSSINI FAVRETTI R. (2000). Progettazione e costruzione di un corpus di italiano scritto : CORIS/CODIS. In : R. Rossini Favretti (ed.), *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Roma : Bulzoni, 39-56.

SAHLGREN M. (2008). The distributional hypothesis. *Rivista di linguistica* 20.1, 33-53.

TLFi : *Trésor de la langue française informatisé*. Paris : CNRS Editions [http://atilf.atilf.fr/].

VILLOING F. (2009). Les composés VN. In : B. Fradin, F. Kerleroux, M. Plénat (éds), *Aperçus de morphologie du français*. Saint-Denis : Presses Universitaires de Vincennes, 175-197.

VILLOING F. (2012). French compounds. *Probus* 24, 29-60.