SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

De la simplicité en morphologie

Tribout, Delphine[1], Barque, Lucie[2], Haas, Pauline[3], Huyghe, Richard[4]

[1] STL, CNRS & Universités Lille 3 et Lille 1, delphine.tribout@univ-lille3.fr
[2] Alpage, INRIA & Université Paris 13, lucie.barque@univ-paris13.fr
[3] Lattice, CNRS & Université Paris 13, pauline.haas@univ-paris13.fr
[4] CLILLAC-ARP & Université Paris Diderot, rhuyghe@eila.univ-paris-diderot.fr

1 Introduction

En morphologie constructionnelle, le lexique simple n'a jamais fait l'objet d'une étude systématique, comme le rappelle Kerleroux (2012 : 171) : « il reste à comprendre pourquoi le lexique simplex n'a pas constitué un objet pertinent de la recherche linguistique : pourquoi les travaux ne produisent jamais cet objet ? ». De ce point de vue, le travail de Croft (1991) sur le russe, à l'origine de sa théorie sur les catégories syntaxiques, fait figure d'exception. Cette exception est toutefois relative dans la mesure où l'étude de Croft ne concerne que 468 mots russes.

Cet article se donne donc pour but de combler en partie ce manque en étudiant la sémantique des noms morphologiquement simples en français. L'objectif est ici de tester sur un corpus du français de grande envergure l'hypothèse de Croft (1991, à paraître) selon laquelle les noms dénotent prototypiquement des objets, les adjectifs des propriétés et les verbes des actions. Croft (1991, à paraître) propose en effet que les catégories du nom, de l'adjectif et du verbe soient définies par une corrélation prototypique entre une classe sémantique et une fonction pragmatique, selon le Tableau 1. Ainsi, un nom comme VEHICULE a pour fonction pragmatique prototypique de référer et dénote un objet ; un adjectif comme BLANC a pour fonction pragmatique prototypique de modifier et dénote une propriété, tandis qu'un verbe comme DETRUIRE a pour fonction pragmatique de prédiquer et dénote une action dans les termes de Croft. Selon cette hypothèse, c'est ensuite le rôle de la morphologie constructionnelle de fabriquer des unités ayant une fonction pragmatique qui n'est pas prototypiquement celle de leur catégorie lexicale, comme le montre le Tableau 2 traduit de (Croft 1991 : 53). Ainsi, BLANCHEUR, par exemple, dénote une propriété bien qu'étant un nom. L'adjectif BLANC a donc été dérivé morphologiquement pour former le nom BLANCHEUR afin de pouvoir référer au lieu de modifier, tout en continuant de dénoter une propriété. De la même façon, DESTRUCTION, dérivé du verbe DETRUIRE, réfère à une action tout en étant un nom. La morphologie constructionnelle permet ainsi qu'un mot comme BLANC ou DETRUIRE conserve la classe sémantique qui lui est prototypique, mais change de fonction pragmatique, pour référer par exemple, au lieu de modifier ou de prédiquer.

	Nom	**Adjectif**	**Verbe**
Classe sémantique **Fonction pragmatique**	objet référence	propriété modification	action prédication
Exemple	VEHICULE	BLANC	DETRUIRE

Tableau 1 : Catégories lexicales selon Croft (1991)

	Fonction pragmatique		
Classes sémantiques	Référence	Modification	Prédication
Objet	**VEHICULE**	VEHICULAIRE	ETRE UN VEHICULE
Propriété	BLANCHEUR	**BLANC**	ETRE BLANC
Action	DESTRUCTION	DETRUIT	**DETRUIRE**

Tableau 2 : Corrélations prototypiques et non prototypiques entre classes sémantiques et fonctions pragmatiques selon Croft (1991 : 53)

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801182

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

Selon Croft, les corrélations prototypiques entre classes sémantiques et fonctions pragmatiques sont non marquées dans les langues, tandis que les combinaisons non prototypiques sont marquées par le biais d'un ou plusieurs morphèmes. Ainsi, pour le nom BLANCHEUR, l'association non prototypique entre le fait de référer et le fait de dénoter une propriété est effectivement marquée par le suffixe *-eur*. De la même façon, pour DESTRUCTION, l'association non prototypique entre le fait de référer et le fait de dénoter une action est marquée par le suffixe *-ion*. Cependant, dans le domaine nominal, des contre-exemples à cette hypothèse existent, tels que COURAGE qui est simple et dénote une propriété, ou CRIME également simple mais dénotant une action.

S'il n'est pas anormal de trouver des contre-exemples, la question à laquelle souhaite répondre cet article, en se limitant au domaine nominal, est celle de la quantification des contre-exemples, afin de déterminer dans quelle mesure l'hypothèse de Croft se vérifie ou non. L'objectif de cette étude est donc de tester l'hypothèse de Croft sur un corpus de noms simples du français. Pour mener à bien cette tâche, nous avons tout d'abord constitué un corpus de noms simples en français, puis annoté chacun de ces noms au moyen de tests permettant leur catégorisation sémantique. Ce faisant, nous avons été confrontés à deux difficultés majeures : la première fut de définir ce qu'est un nom simple français, et la deuxième fut d'élaborer des tests efficaces permettant d'identifier les N d'objet, d'action et de propriété. La section 2 présente la constitution du corpus et les problèmes empiriques et théoriques rencontrés. La section 3 présente les tests utilisés pour l'annotation sémantique. Enfin, la section 4 analyse les résultats de l'annotation.

2 Constitution d'un corpus de noms simples

Pour étudier la sémantique des noms simples du français, nous avons constitué un corpus de noms simples à partir du lexique libre *Lexique 3* (http://www.lexique.org/). Ce lexique comprend 135 000 formes fléchies correspondant à 55 000 lemmes, ainsi que de nombreuses informations pour chaque forme, telles que la catégorie lexicale, le genre et le nombre pour les noms et adjectifs, le temps, le mode, la personne et le nombre pour les verbes, la transcription phonétique, etc. *Lexique 3* fournit également la fréquence des formes fléchies et des lemmes dans deux corpus, l'un étant un sous-ensemble de textes littéraires récents tirés de *Frantext*, et l'autre étant un corpus de sous-titres de films.

2.1 Méthodologie

Le corpus de noms simples a été constitué de manière automatique à partir de *Lexique 3*, puis a été validé manuellement. Après avoir récupéré les 30 630 lemmes de *Lexique 3* catégorisés comme des noms, avec leurs fréquences additionnées dans les deux corpus (textes littéraires et sous-titres de films), nous avons supprimé tous les noms qui pouvaient être identifiés de manière automatique comme construits. Pour cela, nous nous sommes basés sur les propriétés formelles des noms. Nous avons ainsi écarté tous les noms présentant un trait d'union (ABAT-JOUR) ou un espace (AYANTS DROIT) qui peuvent être le signe d'une composition ou d'une lexicalisation de syntagme. Ont ensuite été exclus tous les noms se terminant graphiquement par un suffixe lié à une nominalisation, tels que *-age*, *-ment*, *-ion*, *-ade*, *-ure*, *-ette*, *-ier*, *-isme*, *-erie*, *-esse*, *-ité,* etc. ou par un élément de composition néo-classique tels que *-graphe*, *-philie*, *-scopie*, *-thèque,* etc. à condition que la chaîne de caractères correspondant au lexème moins la finale soit supérieure à deux caractères. Cette contrainte sur la taille du lexème a permis de ne pas exclure des noms tels que PLAGE ou STADE dont les finales sont identiques à des suffixes mais qui ne sont pas construits. Cette méthode d'identification automatique des noms construits fondée sur la reconnaissance des finales permet de traiter non seulement les suffixés construits sur des bases régulières comme LAVAGE ou DIVISION (dérivés de LAVER et DIVISER), mais également les dérivés construits sur des bases allomorphiques tels que PRODUCTION, dérivé de PRODUIRE et construit sur la base allomorphique *product*, ou PERCEPTION construit sur la base allomorphique *percept* du verbe PERCEVOIR. D'autre part, nous avons également décidé d'écarter tous les noms pouvant être en relation de conversion avec un verbe ou un adjectif. En effet, selon Tribout (2010 : 139-196), il n'est pas toujours possible de déterminer l'orientation de la dérivation dans le cas d'une conversion. C'est pourquoi, afin de limiter le risque d'avoir dans notre

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

corpus d'étude des noms qui ne sont pas simples mais qui sont en réalité convertis de verbes ou d'adjectifs, nous avons préféré écarter tous les noms en relation de conversion avec un autre lexème. Pour réaliser cela, nous avons supprimé tous les noms présents dans le corpus de Tribout (2010) pour les conversions nom~verbe, ainsi que tous les noms identiques à un adjectif présent dans *Lexique 3* pour les conversions nom~adjectif. Lors de la validation manuelle, nous avons également écarté tous les noms identiques à une préposition (POUR, CONTRE), un adverbe (DESSUS, DESSOUS) ou une interjection (ALLELUIA, CLIC).

Grâce à cette méthode automatique, nous avons obtenu un corpus de 4 542 noms *a priori* simples, que nous avons ensuite validés manuellement. La phase de validation manuelle a permis de supprimer les noms préfixés, tels que DEPLAISIR ou MESENTENTE, qui n'avaient pas été traités de manière automatique, ainsi que d'écarter les noms construits récupérés à tort. Parmi les noms considérés comme simples se trouvaient en effet des noms suffixés mais dont les suffixes n'avaient pas été listés parmi les finales indiquant une nominalisation, comme CLIENTELE (< CLIENT) ou TOMBEAU (< TOMBE), ainsi que des composés verbe-nom, comme TOURNEVIS, ou néo-classiques, comme EPIDERME, qui ne présentent pas d'indice formel de construction (comme un trait d'union par exemple). Les noms complexes non construits (Corbin 1987), tels que FUNAMBULE ou HABITACLE, ont quant à eux été conservés dans le corpus de noms simples. En effet, ces noms n'étant pas issus d'une règle de formation de lexèmes, l'identification d'une partie (*-ambule*, *habit-*) ne suffit pas à les considérer comme construits. La validation manuelle a également permis d'exclure les noms construits par des procédés extragrammaticaux (Fradin, Montermini et Plénat 2009), tels que le verlan (MEUF < FEMME), le louchébem (LAMEDE < DAME), la reduplication (JOUJOU), la siglaison (SIDA, DRH) ou les mots valises (MODEM, BOBO). Nous avons également écarté les noms apparaissant uniquement dans des locutions, comme FUR, jamais employé seul mais toujours dans la locution *au fur et à mesure*, ou encore FOR n'apparaissant que dans la locution *for intérieur*. Enfin, la phase de validation manuelle a permis de supprimer des erreurs de catégorisation de *Lexique 3*, telles que TARD ou TANTOT catégorisés comme des noms, ou encore des erreurs de segmentation comme *lamedu* (pour *lame du*).

La validation manuelle, parce qu'elle permet de supprimer beaucoup de bruit résultant de la classification automatique, est donc une étape indispensable de la constitution du corpus. Mais, outre la vérification des données, cette phase de validation manuelle nous a confrontés à un certain nombre de cas constituant des problèmes empiriques et théoriques pour lesquels il a fallu déterminer une démarche à suivre.

2.2 Problèmes empiriques et théoriques

Les principaux problèmes auxquels nous nous sommes confrontés sont les apocopes comme VELO, les étymologies populaires comme PEAGE, les emprunts comme HASARD, et les antonomases comme POUBELLE. Ces différents cas constituent un problème empirique pour notre tâche de classification des noms comme construits ou non construits. Mais ils constituent en outre un problème théorique dans la mesure où ils imposent de fixer les limites du lexique simple. Le problème particulier posé par chacun de ces cas est discuté ci-dessous avec la solution que nous avons adoptée.

2.2.1 Apocopes

Les apocopes font partie des procédés extragrammaticaux (Fradin, Montermini et Plénat 2009) de formation de lexèmes. Un des critères proposés par les auteurs permettant de considérer une unité comme construite par un procédé extragrammatical est le caractère conscient de la création. En effet, selon eux, le propre de la morphologie grammaticale productive est d'opérer de façon non consciente, comme cela se passe en syntaxe, tandis que la morphologie extragrammaticale, à l'inverse, opère de façon consciente. De ce point de vue, la plupart des apocopes comme MANIF, TELE, RESTO, TRAUMA, etc. sont bien des constructions extragrammaticales car elles sont toujours perçues par les locuteurs comme des formes réduites de lexèmes existants. Cependant, la question se pose pour certaines unités qui ne sont plus perçues par certains locuteurs comme des formes tronquées de lexèmes. C'est notamment le cas de VELO (< VELOCIPEDE) ou STYLO (< STYLOGRAPHE) dont on pourrait en effet se demander s'ils n'ont pas

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

remplacé dans le lexique les lexèmes dont ils sont issus, ces derniers n'étant plus, ou très peu, usités. Dans le corpus de *Lexique 3*, VELO (64,03 millions d'occurrences) apparaît ainsi près de 120 fois plus que VELOCIPEDE (0,54 million d'occurrences), et STYLO (30,5 million d'occurrences) 64 fois plus que STYLOGRAPHE (0,48 million d'occurrences). La question de l'intégration de ces noms dans le lexique simple s'est donc posée. Toutefois, en étudiant les rapports de fréquence entre les formes apocopées et leur lexème base, nous nous sommes heurtés à deux obstacles. Tout d'abord, il n'est pas certain qu'un simple rapport de fréquences corrobore notre intuition sur le fait qu'une forme apocopée ait remplacé son lexème base dans le lexique. Par exemple, le nom KILO (46,73 millions d'occurrences) apparaît 150 fois plus dans le corpus que son lexème base KILOGRAMME (0,31 million d'occurrences), sans pour autant que l'on ait l'intuition que KILO a remplacé sa base dans le lexique. En outre, Haspelmath (2006 : 47) note que les formes courtes sont, en raison de l'économie à l'œuvre dans le langage, volontiers plus fréquentes que les formes plus longues. Il est donc attendu que les apocopes soient souvent plus fréquentes que leurs bases. Par ailleurs, même si les rapports de fréquences entre forme apocopée et lexème base constituaient des indices fiables, il resterait à déterminer un seuil à partir duquel il est possible de considérer que la forme apocopée a pris la place de son lexème base dans le lexique. Or, fixer un tel seuil nous a paru impossible sans que soient menées des expériences psycholinguistiques sur le sujet. En conséquence, nous avons considéré toutes les apocopes comme des noms construits : les 118 apocopes rencontrées ont été écartées du corpus de noms simples.

2.2.2 Étymologies populaires

Le second problème que nous avons rencontré est celui posé par les étymologies populaires, c'est-à-dire les cas de motivation d'un lexème par rapport à un autre alors même qu'étymologiquement, les deux lexèmes n'entretiennent aucune relation. Un exemple bien connu est le cas de l'adjectif OUVRABLE dans la formule *jours ouvrables* qui est généralement considéré comme lié au verbe OUVRIR et comme signifiant "jours où les magasins sont ouverts". La motivation de OUVRABLE par rapport à OUVRIR est favorisée par le fait que le verbe base OUVRER n'est plus que très peu usité actuellement et est en outre réservé à des vocabulaires techniques. Lors de la constitution du corpus, un problème similaire a été rencontré avec PEAGE. En effet, ce dernier peut être ressenti comme sémantiquement et morphologiquement dérivé du verbe PAYER. Sémantiquement car il désigne une taxe (c'est-à-dire une chose que l'on paie) ou le lieu de perception de cette taxe (c'est-à-dire le lieu où l'on paie) et morphologiquement car sa finale en *-age* l'intègre dans la liste des noms déverbaux suffixés en *-age*. Cependant, étymologiquement PEAGE ne dérive pas du verbe PAYER mais est, d'après le TLFi, un ancien dérivé de PIED. La question s'est donc posée de conserver PEAGE comme nom simple ou au contraire de le considérer comme dérivé de PAYER. Notre étude se voulant résolument synchronique, et dans la mesure où ce nom semble analysable aujourd'hui comme dérivé du verbe PAYER au moyen de la règle de suffixation en *-age*, nous avons décidé de ne pas garder ce nom dans la liste des noms simples. À l'inverse, nous avons considéré les noms originellement construits, mais dont la construction n'est plus analysable en français contemporain comme des noms simples. Ainsi DEPART, construit en ancien français sur le verbe DEPARTIR aujourd'hui disparu, a été considéré comme un nom simple.

2.2.3 Emprunts

Si l'emprunt à une autre langue est un moyen d'augmenter le lexique d'une langue au même titre que la construction de lexèmes, le statut que les lexèmes empruntés ont dans la langue cible n'est pas toujours clair. Les emprunts récents sont généralement ressentis comme tels par les locuteurs, soit parce que la graphie ne correspond pas à l'orthographe de la langue cible, comme pour BUSINESS, soit parce que la prononciation ne correspond pas à la phonologie de la langue, comme c'est le cas avec THRILLER [θrɪlə^{r}]. De tels emprunts sont toujours perceptibles comme ne faisant pas partie du lexique d'origine de la langue, même s'ils peuvent être la base de nouveaux lexèmes construits en français, comme le montrent MAIL et FORWARD qui, bien qu'étant perceptibles comme non français à l'origine du fait de leur prononciation, ont néanmoins donné lieu aux verbes MAILER et FORWARDER. Mais, dans le cas des emprunts plus anciens, il arrive souvent que les locuteurs ne soient pas conscients de leur origine étrangère comme ce peut être le

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

cas par exemple pour HASARD, emprunté à l'arabe *az-zahr* "le dé à jouer" par l'intermédiaire de l'espagnol *azar* "coup défavorable au jeu de dés", ou pour HUSSARD, emprunté au hongrois *huszar* par l'intermédiaire de l'allemand. Dans ces cas, l'origine étrangère des lexèmes peut passer inaperçue d'une part parce que les noms sont conformes à la fois à la graphie et à la phonologie du français, et d'autre part parce qu'ils se comportent morphologiquement comme n'importe quel nom du français, tant en ce qui concerne la flexion (leur pluriel est formé par l'ajout graphique d'un *-s* : *hussards*, *hasards*) que la dérivation (HASARDEUX, HASARDER, HUSSARDESQUE, etc.). De telle sorte qu'il semble difficile aujourd'hui de considérer ces noms comme n'appartenant pas au lexique simple du français. Enfin, entre les deux extrêmes HASARD et THRILLER se trouvent de nombreux cas de figure, plus ou moins perceptibles comme empruntés ou plus ou moins intégrés au lexique simple de la langue.

En ce qui concerne les emprunts trouvés dans *Lexique 3*, la question de leur statut est double. Il s'agit en effet (i) de déterminer s'ils sont bien intégrés au lexique de la langue française ; (ii) si oui, de s'interroger sur le statut morphologique des lexèmes empruntés. Répondre à la première question revient à se demander si les noms en questions sont lexicalisés ou non. Nous avons suivi sur ce point Corbin (1992) qui définit une unité lexicalisée comme une unité codée, disponible pour les locuteurs en tant que présentant des propriétés phonologiques, morphosyntaxiques et sémantiques fixées dans la langue et non modifiables au gré des envies des locuteurs. Nous avons ainsi choisi, comme indice de lexicalisation, de vérifier la présence ou l'absence des noms dans un dictionnaire grand public de référence : *Le Petit Robert électronique* (version 2012). Les noms présents dans le dictionnaire ont ainsi été considérés comme intégrés au lexique français, tandis que les noms absents du *Petit Robert* ont été exclus de notre corpus. Bien que le choix du dictionnaire comporte une part d'arbitraire, cette méthode offre l'avantage de fournir un critère de sélection fiable (présence ou absence) et cohérent. Quant à la deuxième question, nous avons appliqué, pour y répondre, le même critère que pour les autres noms, à savoir que, si un lexème peut être analysé en synchronie comme étant le résultat d'un procédé grammatical ou extragrammatical de formation de lexèmes en français, alors il est considéré comme construit. Sinon, il est simple. Ce critère nous a conduits à intégrer dans notre corpus de noms simples des lexèmes tels que JAZZ, CAFETERIA, CORRIDA, SALSA, YAOURT, YOGA, etc. Mais également des noms comme DRUGSTORE ou GENTLEMAN. En effet, bien que ces derniers soient construits en anglais, ils ne sont pas analysables comme construits en français et sont donc simples du point de vue de la morphologie du français. En revanche, un nom comme TENNISMAN a été exclu car, même s'il semble emprunté, ce nom n'est pas construit en anglais mais bien en français. En effet, le nom TENNISMAN n'existe pas en anglais, un joueur de tennis étant appelé *tennis player*. TENNISMAN semble donc être construit en français à partir d'un élément de formation appartenant à une autre langue. À ce titre, il ressemble aux composés néoclassiques comme BIOGRAPHE ou CHRONOMETRE construits à partir d'éléments de formation appartenant au grec ou au latin, et est donc traité de la même façon qu'eux, c'est-à-dire considéré comme construit. Au total, l'application de ces deux critères (intégration au lexique et non-construction en français) nous a permis de considérer 600 noms empruntés à d'autres langues comme des noms simples appartenant au lexique français.

2.2.4 Antonomases

Enfin, le dernier problème rencontré est celui posé par les antonomases telles que POUBELLE, WATT ou MADRAS. Si ces noms peuvent sembler intégrés au lexique français au point de faire oublier à certains locuteurs leur statut originel de nom propre, d'autres noms propres devenus plus récemment des noms communs peuvent conserver pour certains locuteurs une trace de leur origine. Ce peut être le cas par exemple de KLEENEX, RIMMEL, MARTINI ou ROQUEFORT. Afin de déterminer si un nom propre est intégré dans le lexique des noms communs du français, nous avons appliqué la même méthodologie que pour les emprunts et avons vérifié leur présence ou absence dans le *Petit Robert*. Les noms absents du dictionnaire ont été exclus de notre corpus, tandis que les autres ont été intégrés à la liste des noms simples. Au total, 78 noms issus de noms propres et trouvés dans le *Petit Robert* ont été intégrés au corpus.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

2.3 Conclusion sur la constitution du corpus

La détection automatique de noms simples et la validation manuelle ont donc permis d'établir une liste de noms simples à annoter sémantiquement. Cependant, le corpus obtenu étant encore trop conséquent pour une annotation manuelle de qualité, nous avons appliqué un seuil de fréquence et n'avons retenu que les noms ayant une fréquence supérieure ou égale à 0,3 million d'occurrences dans *Lexique 3*. Ce seuil fixé à 0,3 nous a semblé le meilleur compromis pour avoir à la fois un corpus suffisamment important pour être représentatif du lexique simple et une quantité raisonnable de données à annoter pour les annotateurs tout en ne risquant pas de passer à côté de noms importants dans le lexique français. Une fois appliqué le filtre de la fréquence, nous avons obtenu une liste de 3 489 noms morphologiquement simples que nous avons annotés sémantiquement.

3 Annotation sémantique

Ayant établi un lexique des noms simples du français, nous en proposons une analyse sémantique, afin de vérifier l'hypothèse de Croft selon laquelle les noms simples dénotent principalement des objets (Croft 1991, à paraître). Cette section est consacrée aux questions relatives à l'annotation des noms. Nous commencerons par définir les trois catégories Objet, Action et Propriété, ainsi que les tests permettant d'en identifier les éléments. Nous indiquerons ensuite comment sont traités les noms appartenant à plusieurs catégories, autrement dit les noms qui présentent des sens multiples.

3.1 Tests pour la classification des noms simples

Croft (1991, à paraître) indique peu de critères d'identification formelle des catégories sémantiques qu'il manipule, privilégiant la caractérisation ontologique : les objets sont définis comme des êtres ou des choses, les actions comme des situations dynamiques et transitoires qui mobilisent des participants, et les propriétés comme des grandeurs ou des qualités qui s'appliquent à des entités. Pour déterminer en français la catégorisation comme N d'objet, d'action ou de propriété, nous nous fondons sur un ensemble de tests recueillis dans les travaux de sémantique nominale, notamment Godard & Jayez (1996), Kleiber *et al.* (2012) pour les noms d'objet, Gross & Kiefer (1995), Godard & Jayez (1996), Haas *et al.* (2008), Arnulphy *et al.* (2011) pour les noms d'action, et Van de Velde (1995), Flaux & Van de Velde (2000), Beauseroy (2009), Goossens (2011) pour les noms de propriété. Les tests retenus présentent le double avantage de faire l'objet d'un large consensus et de correspondre, nous semble-t-il, à la catégorisation que propose Croft.

Noms d'objet – Les noms d'objet dénotent ici des entités localisées dans l'espace et dotées d'une étendue spatiale, qu'elles soient animées ou non animées, discrètes ou continues. Pour être étiqueté « objet », un nom doit ainsi passer le test de localisation spatiale (1) et / ou les tests de mise en évidence des propriétés physiques de l'entité décrite — matière (2), couleur (3), dimensions (4).

(1) N *se trouver* + complément de localisation spatiale (e.g. *sur la table, à côté du sac, dans le jardin, à Paris,* etc.) : *Mon frère / le cadeau / le sac se trouve dans la voiture.*

(2) N + complément de constitution matérielle (e.g. *en coton, en chêne massif, en terre-cuite,* etc.) : *un manteau en laine / une statue en marbre / un panier en osier.*

(3) N + adjectif de couleur (e.g. *violet, beige, jaune d'or*, etc.) : *une soie jaune d'or / un hippopotame turquoise / un vélo violet.*

(4) N + complément de dimension : taille (*de x mètres de large, de x m^2, de x m^3, de x hectares*, etc.), poids (*de x grammes, de x kilos*, etc.) : *une tente de 2 mètres de haut / un cigare de 15 grammes / un vignoble de plusieurs hectares.*

Dans le test (1), les sites de localisation qui dénotent des objets comprenant une facette informationnelle (e.g. *dans ce livre / dans cette photo / dans ce discours*) ne doivent pas être utilisés. Ces compléments de localisation permettent en effet de mettre en évidence les noms d'entité à contenu conceptuel (cf. section

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

4.2). De même, les compléments de propriétés physiques dans les tests (2), (3) et (4) ne doivent pas présenter une acception métaphorique, sous peine de fausser l'application du test : les locutions telles que *en or*, les adjectifs comme *rouge* et *vert*, ou *grand* et *gros* sont donc exclus des tests.

On notera que les noms d'objet identifiés d'après (1)-(4) ne se réduisent pas aux noms comptables. En effet, les noms de matière, qui sont massifs, peuvent valider le test (1) (*Le sable se trouve dans le hangar*).

Noms d'action – Les noms d'action englobent ici l'ensemble des noms dénotant des situations dynamiques, que celles-ci soient duratives ou ponctuelles, téliques ou atéliques. Les noms d'action doivent vérifier au moins un des tests de dynamicité ci-dessous.

(5) N *avoir lieu à tel moment (à tel endroit)* : *Le crime / le festival / le striptease a eu lieu hier sur les Champs Elysées.*

(6) N *se produire à tel moment (à tel endroit)* : *Le clash / le tsunami / l'exode s'est produit en 2006.*

(7) *Effectuer* N : *effectuer un travelling / un safari / un cursus.*

(8) *Procéder à* N : *procéder à un raid / à un baptême / à un référendum.*

Ces tests ne dépendent pas de la présence ou non d'arguments syntaxiques associés au nom : ceux-ci sont compatibles (*Une manifestation des agriculteurs a eu lieu à Paris ce matin*) mais non requis pour la validation des tests (5)-(8). L'existence de structures argumentales nominales est abondamment commentée dans les travaux sur les nominalisations, dont beaucoup sont corrélées morphologiquement à des prédicats verbaux dynamiques (cf. Grimshaw, 1990 ; Alexiadou 2001 ; Borer 2003). L'existence d'une structure argumentale ne saurait toutefois constituer un test d'identification des noms d'action. D'une part, certains noms statifs admettent également des arguments (*la méfiance de Pierre*). D'autre part, certains noms dynamiques, fussent-ils déverbaux, ne sont pas dotés d'une structure argumentale (**le jardinage de Pierre*). La question de savoir si les noms simples qui dénotent des actions peuvent ou non avoir une structure argumentale est ouverte. Il est néanmoins manifeste que certains d'entre eux admettent un complément en *de* renvoyant à des participants de l'action (*le crime de Pierre*, *la grève des ouvriers*, *le match de l'équipe de France*).

Noms de propriété – Les noms de propriété dénotent des situations d'aspect statif, i.e. non dynamique. Nous regroupons parmi ces noms aussi bien les noms d'état que les noms de qualité, qui dénotent respectivement des propriétés transitoires ou structurelles. Ainsi définis, les noms de propriété doivent pouvoir s'employer dans au moins une des constructions suivantes.

(9) *Un état de* N : *un état de misère / d'ascèse / de liesse*

(10) *Ressentir / éprouver du* N : *ressentir de la haine / éprouver de la honte*

(11) *Faire preuve de* N : *faire preuve de flegme / de fougue / d'audace*

(12) *Être d'un grand* N, à condition que la construction soit sémantiquement équivalente à *avoir du* N : *être d'une grande intelligence (≈ avoir de l'intelligence)* vs **être d'une grande famille (≠ avoir de la famille)*

Nous avons appliqué l'ensemble de ces tests aux noms figurant dans notre lexique, et annoté ceux-ci en conséquence. Les substantifs ne vérifiant aucun des tests retenus ont été considérés à ce stade comme sémantiquement indéterminés.

3.2 Traitement des N simples à sens multiples

Toute description sémantique se heurte au problème de la délimitation des sens lexicaux. L'annotation sémantique de nos 3 489 noms simples comme noms d'objet (O), d'action (A) ou de propriété (P), de fait, a requis d'identifier au préalable leur éventuelle multiplicité de sens. Toutefois la quantité des données à annoter nous a contraints à ne pas entrer dans des distinctions de sens fines et à nous concentrer sur

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

l'opposition entre les acceptions d'objet et les autres. Nous indiquons ici les choix relatifs à l'identification des unités à annoter compte tenu du jeu d'étiquettes adopté.

Critères de distinction des sens – Dans cette étude, les noms sont considérés comme dotés de sens multiples uniquement s'ils passent des tests de plus d'une des trois catégories définies (13-15). Les noms à sens multiples, mais dont les acceptions appartiennent toutes à une même catégorie, sont traités comme monosémiques : le nom BUREAU en est un exemple car il dénote un objet dans toutes ses acceptions — dont deux sont illustrées en (16).

(13) REPAS : A/O
Le repas a lieu tous les jours à 13 heures. → Action
Le repas se trouve sur la table. → Objet

(14) INITIATIVE : A/P
Cette initiative a eu lieu en décembre de cette année. → Action
Elle fait preuve d'initiative. → Propriété

(15) AMOUR : O/P
Son amour se trouve dans le hall. → Objet
Ressentir de l'amour → Propriété

(16) BUREAU : O
Un bureau en bois clair → Objet
Le bureau se trouve au premier étage, porte de droite. → Objet

Ce traitement des sens multiples présente l'inconvénient de ne pas mesurer précisément le nombre de sens Objet, Action ou Propriété dans le lexique constitué, mais uniquement le nombre de noms ayant (au moins) un de ces sens. Le nom BUREAU ne sera par exemple comptabilisé qu'une fois comme nom d'objet alors qu'il a *a priori* plusieurs acceptions d'objet distinctes. De plus, notre étude ne portant pas sur la polysémie des noms simples en tant que telle mais sur leur propension à dénoter des objets ou d'autres éléments, nous avons pris le parti de limiter les sources d'ambiguïté en optant pour un jeu d'étiquettes très simple et de nous reposer sur des critères clairs de délimitation des sens, fondés sur l'application des tests.

Relations entre sens – Une fois identifiés les différents sens d'un lemme, la question peut se poser de savoir s'il existe ou non un lien entre ces sens, autrement dit s'ils relèvent de l'homonymie ou de la polysémie (Bréal, 1904 ; Kleiber, 1999), voire de différents types sémantiques pouvant coexister, ce dernier phénomène ayant été décrit dans la littérature en termes de multitypage (Pustejovsky, 1995 ; Godard & Jayez, 1996 ; Jacquey 2006), de facettes (Cruse, 1995), de métonymie intégrée (Kleiber, 1999) ou d'ambivalence sémantique (Milićević & Polguère, 2010). Nous avons à ce stade opté pour un traitement uniforme des sens multiples. Ceux-ci sont ainsi décrits au sein d'une même entrée, qu'il s'agisse de cas d'homonymie sans changement de genre (17) ou de polysémie (18). Seuls les homonymes de genres différents ont été considérés comme des lexèmes distincts (19-20), 22 paires de ce type étant concernées dans notre lexique.

(17) GREVE : A/O
La grève aura lieu demain. → Action
Une grève de plusieurs kilomètres → Objet

(18) CIRQUE : A/O
Tout ce cirque a eu lieu la semaine dernière. → Action
Le cirque se trouve sur la place du village. → Objet

(19) MANCHE (masc.) : O
Le manche de la pelle se trouve dans le garage. → Objet

(20) MANCHE (fém.) : A/O
La deuxième manche a eu lieu après une heure de jeu. → Action
La manche beige du pull. → Objet

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

La distinction de traitement entre homonymes de même genre et homonymes de genres distincts, discutable *a priori*, n'a été retenue que parce qu'elle repose sur un critère purement morphologique. L'identification des homonymes de même genre parmi l'ensemble des noms simples étudiés nécessiterait que l'on dispose de critères précis pour les distinguer des cas de polysémie. Or, on sait que la frontière entre polysémie et homonymie est poreuse (Bréal 1904, Kleiber 1999). Se lancer dans cette entreprise de description impliquerait du reste d'identifier l'ensemble des sens des 3 489 noms simples.

Au total, environ 200 des 3 489 noms du lexique sont « multisémiques » au sens défini plus haut. Les noms présentant des sens Action et Objet sont les plus fréquents parmi eux.

4 Résultats et analyse

Nous donnons, dans cette dernière section, les résultats de l'annotation sémantique des noms simples et soulignons les limites de la tripartition objet, action, propriété en listant les cas de noms non réductibles à l'une de ces trois classes.

4.1 Répartition des N simples

La classification sémantique des 3 489 noms simples en noms d'objet, d'action ou de propriété confirme l'intuition exprimée par Croft (Croft 1991, à paraître), mais nous invite à nuancer son propos : les noms d'objet sont certes très majoritaires au sein du lexique considéré, cependant la proportion des noms pouvant dénoter autre chose qu'un objet (près d'un quart) ne permet pas de les considérer comme de simples exceptions à la règle. Comme le montre le tableau ci-dessous, on compte parmi les noms simples un nombre important de noms d'action, quelques noms de propriété, et près de 15% de noms qui ne peuvent être classés dans aucune de nos trois catégories de départ.

Noms ayant un sens Objet	Noms ayant un sens Action	Noms ayant un sens Propriété	Noms ayant un autre sens
2 807	300	77	512
75,9%	8,1%	2,1%	13,9%

Tableau 3 : Répartition sémantique des N simples

4.2 Les N simples non classés

De nombreux noms simples dans notre liste ne valident aucun des tests identificatoires proposés et paraissent ainsi échapper à la catégorisation comme nom d'objet, d'action ou de propriété. Nous pouvons apparier certains de ces noms à des types nominaux plus spécifiques, régulièrement rencontrés dans notre corpus.

Des noms comme MOT, EXERGUE, SARCASME, MYTHE, SCOOP, PONCIF, DILEMME dénotent des objets informationnels. Ces noms sont définis récursivement par Godard & Jayez (1996) comme pouvant se construire avec *se trouver dans GN* et *être contenu dans GN*, si le GN localisateur a lui-même pour noyau un nom d'objet informationnel (e.g. *Ce mythe se trouve dans l'Iliade*). Selon Flaux & Stosic (2012), les noms en question dénotent des idéalités, i.e. des entités dotées d'un contenu conceptuel interprétable, et ils sont souvent associés à un référent sensible, la catégorie se subdivisant à son tour en idéalités langagières, géométriques, musicales, iconiques, etc. De fait, de nombreux objets informationnels sont liés à des objets (e.g. PORTRAIT, LOI, POEME), voire à des actions (e.g. CAS, MENSONGE, ANECDOTE), ce qui confère à cette catégorie une représentation lexicale assez large : 137 noms de notre corpus présentent au moins une acception d'objet informationnel.

En marge des noms d'action, il existe des noms comme SPORT, YOGA, KARTING, SOLFEGE, PANTOMIME, SALSA, SHOPPING, qui dénotent des activités. Ces noms ne décrivent pas des entités concrètes, et on peut

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

penser qu'ils sont d'aspect dynamique. Ils ont toutefois pour particularité de ne pas dénoter d'occurrences d'actions, et de ne vérifier aucun de nos tests de catégorisation comme nom d'action. Non comptables, ces noms se distinguent par leur prédilection pour l'emploi générique au singulier (e.g. *Le yoga est souvent une pratique quotidienne*) et dans la tournure *faire du N* (e.g. *Pierre fait du judo*). Leur absence de structure argumentale est caractéristique, et inattendue pour des noms décrivant des situations qui impliquent des participants (**le handball de Pierre*).

Les noms temporels, comme MOMENT, PERIODE, JOUR, HEURE, SIECLE, OCTOBRE, SAMEDI, constituent une autre catégorie particulière. Ces noms dénotent des parties du temps, et ils s'emploient régulièrement pour désigner des repères temporels, que ce soit avec des prépositions temporelles (*pendant cette période*, *après ce jour*, *en octobre*) ou en emploi absolu (*samedi dernier*). Les noms temporels n'ont pas de trait aspectuel de dynamicité ou de stativité ; ils ne comportent qu'une indication de durée et de délimitation temporelle, et possèdent une certaine autonomie référentielle. Berthonneau (1989) distingue les noms de temps selon qu'ils sont ou non étalonnés (e.g. HEURE *vs* MOMENT), les premiers pouvant constituer des unités de mesure temporelle. Par ailleurs, certains noms temporels, comme les noms de jour (JEUDI), de mois (MARS) ou de saison (HIVER), spécifient le repérage temporel en indiquant une position donnée dans un cycle préétabli.

Les autres noms simples qui dénotent des unités de mesure, comme TONNE, LIEUE, MILE, ONCE, LUMEN, PHOT, AMPERE, échappent également à la catégorisation comme nom d'objet, d'action ou de propriété. Ces noms s'emploient dans la quantification de certaines propriétés, qui ont leur propre dénomination (e.g. TAILLE, POIDS, LUMINOSITE), mais ils ne sont pas eux-mêmes des noms de propriété. Ils se caractérisent par leur rôle prédicatif et ont notamment la particularité, lorsqu'ils sont précédés d'un numéral, d'entrer dans la construction de syntagmes binominaux quantificateurs (e.g. *deux tonnes de N*) (cf. Benninger, 1999).

D'autres catégories encore peuvent être mentionnées. Par exemple, les noms de monnaie (e.g. DOLLAR, EURO, PESO, YUAN, PIASTRE) ont une acception d'objet, mais ils dénotent également une abstraction économique, référentiellement autonome et ne se définissant pas comme une propriété. Les noms de maladie (e.g. ASTHME, ANGINE, TETANOS, RHUME, SCORBUT) ne s'emploient ni dans les constructions des noms d'action, ni dans celles des noms de propriétés. Il paraît difficile de leur attribuer des propriétés aspectuelles. Cette catégorie se scinde entre noms comptables et non comptables (e.g. *trois angines* vs *#trois asthmes*), les premiers pouvant s'associer référentiellement à des périodes (*une angine de quatre jours*, *pendant son rhume*), quand les seconds se limitent essentiellement à des emplois prédicatifs ou définis singuliers (*un cas de scorbut*, *le vaccin contre le tétanos*). Enfin, les noms de phénomènes (e.g. LUEUR, ECHO, RELENT, BRISE, MIASME) apparaissent comme une catégorie intermédiaire entre les noms d'objet et les noms d'action. Ils décrivent des entités non matérielles, mais accessibles aux sens et localisées dans l'espace-temps (e.g. *Il y avait une lueur / de l'écho / des relents de cuisine dans le hall ce matin*), bien qu'ils s'emploient difficilement à la fois avec *avoir lieu* et *se trouver*. Ils constituent en outre une classe hétérogène, selon qu'ils dénotent des phénomènes lumineux, sonores, olfactifs (cf. Kleiber & Vuillaume, 2011), climatiques, etc., et privilégient alors la description spatiale ou temporelle.

L'ensemble de ces cas, listés sans exhaustivité, indiquent que la tripartition sémantique en objets, actions et propriétés ne saurait épuiser la dénotation des noms simples en français. Certaines catégories nominales, bien que statistiquement marginales, ont des particularités descriptives et distributionnelles qui les distinguent clairement des noms d'objet, d'action et de propriété, et qui justifient leur prise en compte dans une classification complète des noms simples. L'analyse sémantique du lexique des noms simples montre donc à la fois l'importance de certaines macro-catégories sémantiques (objet, action, propriété), mais aussi la nécessité de prendre en considération, pour établir une typologie nominale exhaustive, l'existence de micro-catégories ne relevant pas des premières.

5 Conclusion

Du point de vue sémantique, cette étude a permis de confirmer et de quantifier l'intuition croftienne selon laquelle les noms simples dénotent des objets, puisque trois quarts des noms de notre corpus ont au moins

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

une acception objectuelle. Le quart restant est constitué de noms dénotant des actions, des propriétés ou d'autres choses. Il y a donc des noms simples qui, à l'instar des noms construits, peuvent dénoter des actions (catégorie prioritairement réservée aux verbes) ou des propriétés (catégorie prioritairement réservée aux adjectifs). Les 15% de noms non annotés sémantiquement peuvent être catégorisés comme « noms de maladies », « noms de temps », « noms de phénomènes », etc. Ils mettent à mal la tripartition de Croft, qui se révèle insuffisante pour classifier sémantiquement l'ensemble des noms simples du français.

Cette première annotation laisse entrevoir plusieurs pistes de travail. D'une part, le traitement des noms à sens multiples demande de plus amples développements. La distinction entre les différentes acceptions et la prise en compte du degré de disjonction sémantique, de l'homonymie au multitypage lexical, pourront être affinées pour préciser l'étiquetage nominal. D'autre part, l'annotation des noms simples qui ne dénotent pas des objets, des actions ou des propriétés est à poursuivre. Elle soulève la question de la définition et des critères d'identification des autres types nominaux, à peine abordée ici. On pourrait également se demander, en inversant la problématique de Croft, si certaines catégories résiduelles, comme celles des noms de temps ou des noms de mesure, ne sont pas constituées essentiellement de noms simples.

Du point de vue morphologique, notre travail a permis de préciser plusieurs points relatifs à l'identification des noms simples. Définis par défaut comme non construits, ces noms sont rarement étudiés pour eux-mêmes, et leur identification ne semble pas *a priori* soulever de difficultés. Pourtant, tenter de constituer une liste de noms simples a révélé des problèmes empiriques et méthodologiques de reconnaissance de ces noms. Certains cas, par exemple les antonomases ou les apocopes, pourraient être traités différemment, notamment en recourant à des expériences psycholinguistiques, ou, dans le cas des apocopes, en élaborant un calcul fréquentiel satisfaisant pour déterminer quelle forme peut être considérée comme simple, d'après l'usage. Un approfondissement des cas problématiques sera nécessaire afin d'aboutir à la liste la plus exhaustive possible des noms simples usuels du français.

Références bibliographiques

Alexiadou, A. (2001). *Functional Structure in Nominals: Nominalization and Ergativity*. Amsterdam: John Benjamins.

Arnulphy, B., Tannier, X. & Vilnat, A. (2011). Un lexique pondéré des noms d'événements en français. In *Actes de TALN 2011*, 51-56.

Beauseroy, D. (2009). *Syntaxe et sémantique des noms abstraits statifs : Des propriétés verbales et adjectivales aux propriétés nominales*. Thèse de doctorat, Nancy-Université.

Benninger, C. (1999). *De la quantité aux substantifs quantificateurs*. Metz : Université de Metz, coll. Recherches Linguistiques 23.

Berthonneau, A.-M. (1989). *Composantes linguistiques de la référence temporelle : Les compléments de temps, du lexique à l'énoncé*. Thèse d'état, Paris Diderot-Paris 7.

Borer, H. (2003). Exo-skeletal vs. endo-skeletal explanations: syntactic projections and the lexicon. In Moore, J. & Polinski, M. (éds), *The Nature of Explanation in Linguistic Theory*, Stanford : CSLI Publications, 31-67.

Bréal, M. (1904). *Essai de sémantique*. Paris : Hachette.

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*. Tübingen : Niemeyer.

Corbin, D. (1992). Hypothèse sur les frontières de la composition nominale. *Cahiers de grammaire, 17*, 25-55.

Croft, W. (1991). *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University Press of Chicago.

Croft, W. (à paraître). *Morphosyntax: constructions of the world's languages. Chap. 1 : Grammatical categories, semantic classes and information*. Non publié.

Cruse, D.A. (1995). Polysemy and related phenomena from a cognitive linguistic viewpoint. In St Dizier, P. &

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801182

Viegas, E. (éds), *Computational Lexical Semantics,* Cambridge: Cambridge University Press, 33-39.

Flaux, N. & Stosic, D. (2012). Les noms d'idéalités sont-ils polysémiques. In Saussure, L. & Rihs, A. (éds), *Études de sémantique et de pragmatique françaises*, Bern : Peter Lang, 167-190.

Flaux, N. & Van de Velde, D. (2000). *Les noms en français : Esquisse de classement*. Paris : Ophrys.

Fradin, B., Montermini, F. & Plénat, M. (2009). Morphologie grammaticale et extragrammaticale. In Fradin, B., Kerleroux, F. & Plénat, M. (éds), *Aperçus de morphologie du français,* Saint-Denis : Presses Universitaires de Vincennes, 21-45.

Godard, D. & Jayez, J. (1996). Types nominaux et anaphores : le cas des objets et des événements. In De Mulder, W., Tasmowski-De Ryck, L. & Vetters, C. (éds), *Anaphores temporelles et (in-)coherence, Cahiers Chronos, 1*, Amsterdam : Rodopi, 41-58.

Goossens, V. (2011). *Propositions pour le traitement de la polysémie régulière des noms d'affect*. Thèse de doctorat, Université Grenoble 3.

Grimshaw, J. (1990). *Argument Structure*. Cambridge Mass.: The MIT Press.

Gross, G. & Kiefer, F. (1995). La structure événementielle des substantifs. *Folia Linguistica, 29*, 43-65.

Haas, P., Huyghe, R. & Marín, R. (2008). Du verbe au nom : calques et décalages aspectuels. In Durand, J., Habert, B. & Laks, B. (éds), *Congrès Mondial de Linguistique Française 2008*, Paris : Institut de Linguistique Française, 2039-2053.

Haspelmath, M. (2006). Against Markedness (And What to Replace It With). *Journal of Linguistics, 42-1,* 25-70.

Jacquey, E. (2006). Un cas de « polysémie logique » : modélisation de noms d'action en français ambigus entre processus et artefact. *TAL, 47/1*, 137-166.

Kerleroux, F. (2012). Il y a nominalisation et nominalisation. *Lexique, 20*, 157-172.

Kleiber, G. (1999). *Problèmes de sémantique : La polysémie en questions*. Villeneuve d'Ascq : Presses Universitaires du Septentrion.

Kleiber, G., Benninger, C., Biermann-Fischer, M., Gerhard-Krait, F., Lammert, M., Theissen, A. et Vassiliadou, H. (2012). Typologie des noms : le critère *se trouver* + SP loc. *Scolia, 26*, 105-130.

Kleiber, G. & Vuillaume, M. (2011). Sémantique des odeurs. *Langages, 181*, 17-36.

Milićević, J. & Polguère, A. (2010). Ambivalence sémantique des noms de communication langagière du français. In Neveu, F., Muni Toke, V., Durand, J., Klingler, T., Mondada, L. & Prévost, S. (éds), *Congrès Mondial de Linguistique Française 2010*, Paris : Institut de Linguistique Française, 1029–1050.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge Mass.: The MIT Press.

Tribout, D. (2010). *Les conversions de nom à verbe et de verbe à nom en français.* Thèse de doctorat, Université Paris Diderot-Paris 7.

Van de Velde, D. (1995). *Le spectre nominal : Des noms de matières aux noms d'abstractions*. Louvain : Peeters.