SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

Constitution automatique d'une ressource morphologique : VerbAgent

Tribout Delphine***

Ligozat Anne-Laure**

Bernhard Delphine*

*Université de Strasbourg & LiLPa ; **ENSIIE & LIMSI ; ***Université Paris 8 & LLF
dbernhard@unistra.fr ; annlor@limsi.fr ; dtribout@linguist.jussieu.fr

1 Introduction

Les systèmes de traitement automatique des langues (TAL) intègrent souvent des connaissances de nature linguistique. Parmi ces types de connaissances, la morphologie est très souvent utilisée, en particulier dans le traitement des langues comme le français, à morphologie flexionnelle riche. Les informations morphologiques intégrées dans des systèmes de TAL concernent essentiellement la morphologie flexionnelle, c'est-à-dire la partie de la morphologie qui s'intéresse aux différentes formes que peut prendre un même lexème en fonction du contexte syntaxique.

Cependant, plusieurs travaux ont montré que l'intégration de la morphologie dérivationnelle, c'est-à-dire la partie de la morphologie qui s'intéresse aux relations entre plusieurs lexèmes, pouvait contribuer à améliorer les systèmes. Par exemple en terminologie, Jacquemin et al. (1997) ont montré que la morphologie dérivationnelle permet d'améliorer la reconnaissance de termes. En reconnaissance de la parole Creutz et al. (2007) ont montré que dans les langues à morphologie riche comme le Finnois, l'analyse en morphèmes rend le système plus robuste au problème posé par les mots hors vocabulaire. En traduction automatique, selon (Lee, 2004), l'analyse morphologique améliore les résultats lorsque les langues source et cible ont des structures morphologiques différentes. Enfin, des travaux tels que (de Loupy et al., 1998) ou (Moreau & Claveau, 2006), ont montré que la morphologie dérivationnelle peut améliorer la performance d'un système pour une tâche telle que la Recherche d'Information.

Si la morphologie dérivationnelle peut être utile aux systèmes de TAL, son intégration dans les systèmes peut se faire de deux façons : au moyen d'outils ou algorithmes, ou *via* des ressources dédiées. Si les ressources linguistiques posaient des problèmes de stockage il y a quelques décennies ce n'est plus le cas aujourd'hui, et on assiste au développement de ressources morphologiques. Pour le français il existe notamment *Morphalou*[1] et *Lefff*[2] (Sagot, 2010) qui traitent la flexion.*VerbAction*[3] (Hathout et al., 2002 ; Hathout & Tanguy, 2002) traite une partie de la dérivation, à savoir les noms d'action ou d'activité morphologiquement apparentés à des verbes. *Nomage* [4] (Balvet et al., 2010) est un lexique sémantique de noms déverbaux, appartenant aux classes aspectuelles des états, habitudes, activités, accomplissement et achèvements. Il existe également des ressources non spécifiques à la morphologie, mais qui intègrent néanmoins des informations propres à la morphologie dérivationnelle, comme *Prolexbase*[5] (Bouchou & Maurel, 2008 ; Tran & Maurel, 2006) ou *Dubois*[6], issue de (Dubois & Dubois-Charlier, 1997).

En matière de ressources dédiées à la morphologie dérivationnelle, Bernhard et al. (2011) ont souligné certains manques pour le français. Dans le cadre des systèmes de Question-Réponse, ils ont étudié les relations morphologiques présentes entre les mots de la question et les mots du passage contenant la bonne réponse. Ils ont ensuite évalué les ressources existantes qui couvrent les relations observées, et ont constaté, entre autres, qu'il n'existe pas de ressource spécifique pour les noms d'agent déverbaux, alors même que ce type de relation est présent entre les mots d'une question et ceux de la réponse. C'est pourquoi nous avons voulu constituer une ressource de noms d'agent déverbaux, afin de combler le

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100324

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

manque pointé par (Bernhard et al., 2011). Nous avons décidé d'appeler cette ressource VerbAgent, en référence à VerbAction de (Hathout et al., 2002).

Dans (Bernhard et al. 2011), les auteurs n'ont pas défini ce qu'ils considéraient comme un nom d'agent déverbal. C'est pourquoi nous allons dans un premier temps circonscrire les notions d'agent et de nom déverbal. Puis, nous présenterons la façon dont a été constituée la ressource, et les différentes méthodes de validation utilisées.

2 Noms d'agent déverbaux

Un nom d'agent déverbal est un nom morphologiquement dérivé d'un verbe, et dénotant un agent. Nous considérons comme dérivé d'un verbe un nom qui est morphologiquement analysable en synchronie, indépendamment de son étymologie. Ainsi un nom comme *directeur* est morphologiquement analysable comme dérivé du verbe *diriger*, même si, d'un point de vue étymologique, il vient du latin. Nous adoptons donc l'analyse des noms en *-eur* proposée par (Bonami, Boyé & Kerleroux, 2009), et nous plaçons de façon plus générale dans le cadre de la morphologie lexématique tel qu'il a été défini par (Matthews, 1972) et (Aronoff, 1994). Ce travail est ainsi mené dans la lignée des travaux menés en morphologie française, notamment (Fradin, 2003), (Fradin & Kerleroux, 2003), (Kerleroux, 2004), (Namer, 2009), (Villoing, 2009), (Dal & Namer, 2010).

La notion d'agent soulève quelques difficultés. Cette notion a été particulièrement développée dans le cadre d'études sur les rôles thématiques des arguments du verbe. C'est pourquoi nous allons dans un premier temps présenter les rôles thématiques, puis nous exposerons la définition d'un agent que nous avons retenue.

2.1 Les rôles thématiques

Les rôles thématiques ont été conçus depuis (Fillmore, 1968) comme une interface entre syntaxe et sémantique permettant de rendre compte de l'appariement entre les arguments sémantiques d'un verbe et ses dépendants syntaxiques. Le nombre et la caractérisation des rôles thématiques varient selon les approches et les auteurs. Pour ne présenter que quelques études menées dans des cadres théoriques très différents, il y a selon (Dowty, 1991) deux rôles thématiques uniquement : *agent* et *patient.* À l'inverse, Van Valin & LaPolla (1997) définissent, quant à eux, treize rôles thématiques : *agent, effectuateur, expérienceur, instrument, force, patient, thème, bénéficiaire, destinataire, but, source, localisation* et *chemin.* Enfin, pour Davis & Koenig (2000) il n'existe que cinq rôles : *agent, patient, état de chose, figure* et *site*.

Il est important de souligner que les rôles thématiques ont toujours été établis dans le but de catégoriser les arguments du verbe, c'est-à-dire toujours dans le contexte d'un énoncé, et jamais avec l'objectif de déterminer hors contexte la valeur sémantique d'un nom. Ainsi, dans une phrase comme *souris* sera considéré comme un agent dans les trois approches, tandis que dans la phrase *souris* sera considéré comme un patient. De la même façon, *balle* dans la phrase sera considéré comme un agent par Dowty et Davis et Koenig et comme un instrument par Van Valin et LaPolla, tandis que dans la phrase ce sera considéré comme un patient par Dowty et Davis et Koenig et comme un thème par Van Valin et LaPolla.

(1) La souris mange le fromage.

(2) Le chat mange la souris.

(3) La balle a cassé la vitre.

(4) Jean lance la balle.

Ces exemples illustrent le fait que l'affectation d'un rôle thématique à un nom est nécessairement lié à un énoncé particulier, et ne vaut pas hors contexte. En effet *souris* et *balle* ne peuvent, hors contexte, être définis à la fois comme des agents et des patients. D'autre part, d'un point de vue de sémantique lexicale,

il semble difficile de décrire ces noms comme des agents ou des patients, et on serait plutôt amenés à les définir comme un animé non humain pour *souris*, et comme un artefact pour *balle*.

Notre objectif étant de réaliser une ressource générique et utilisable dans de multiples applications de traitement automatique des langues, il semble que les critères d'ordre sémantique proposés dans le cadre d'analyses syntaxiques ne sont pas transposables hors contexte pour la morphologie. C'est pourquoi nous avons redéfini d'un point de vue morphosémantique ce que nous considérons comme un agent.

2.2 Définition retenue pour la constitution de la ressource

Dans un premier temps nous avons restreint la définition d'un agent à un individu animé humain. Une telle définition écarte donc les noms comme *balle*, qui ne sont pas des animés, mais également les noms comme *souris,* qui ne sont pas des humains. Distinguer les noms dénotant des animés humains des autres types de noms est relativement aisé. On peut s'appuyer pour cela sur les tests proposés par (Flaux & Van de Velde, 2000). Selon les auteurs, un nom dénotant un animé se distingue des autres noms, entre autres, parce qu'il autorise le pronom relatif *qui* précédé d'une préposition, alors que les autres types de noms ne le permettent pas, comme le montrent les exemples en .

(5) a. Le garçon à qui je parle.

b. Le garçon sur qui je compte.

c. *La chaise sur qui je suis assise.

Parmi les animés, la distinction entre humains et non humains peut se faire, selon Flaux et Van de Velde, grâce à un test supplémentaire : les noms dénotant un humain peuvent rentrer dans la structure exprimant la possession "N_1 est à N_2", tandis que les animés non humains ne le peuvent pas ou difficilement, comme le montrent les exemples (6a) et (6b), à moins que l'animé soit un animal domestique alors considéré comme quasi-humain (6c) :

(6) a. Cette maison est aux amis de mes parents.

b. *Ce trou est à la marmotte.

c. Cette balle est au chat du voisin.

Une fois les animés humains distingués, faire la distinction entre différents types sémantiques de noms est en revanche beaucoup moins aisé. On peut par exemple se demander s'il faudrait distinguer sémantiquement des noms comme *enfant, père*... de noms comme *chanteur, président*... Toutefois, dans la mesure où notre ressource est une ressource morphologique nous avons laissé de côté cette question, et nous avons conservé comme seul critère discriminant le fait que le nom soit dérivé d'un verbe ou non. Cependant, le fait que le nom désigne un humain et soit dérivé d'un verbe n'est pas suffisant. En effet, il nous semblait important de distinguer un nom comme *destinataire* d'un nom comme *contestataire,* le premier désignant la personne à qui est destiné quelque chose, tandis que le second désigne la personne qui conteste quelque chose. Pour distinguer ces deux types de noms nous avons utilisé les critères de (Dowty, 1991) distinguant les proto-agents des proto-patients, qui sont rappelés dans les Tableau 1. Selon Dowty, pour considérer un argument comme proto-agentif il n'est pas nécessaire que l'ensemble des propriétés des proto-agents s'appliquent, mais l'argument doit avoir plus de propriétés propres aux proto-agents qu'aux proto-patients. Comme pour les autres rôles thématiques proposés dans d'autres études, ces critères ont été établis dans le cadre d'une analyse syntaxique. Cependant, ils nous ont semblé utiles à notre tâche dans la mesure où ils permettent de distinguer *destinataire* et *contestataire.* En effet, selon les critères de Dowty, *contestataire* est bien un proto-agent, tandis que *destinataire* est un proto-patient. En appliquant les critères de Dowty des noms tels que *ronfleur* ou *connaisseur* ont été considérés comme des agents, alors qu'ils sont traités comme des expérienceurs par Van Valin et LaPolla.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

Propriétés des proto-agents	Propriétés des proto-patients
est volitionnellement impliqué dans un événement ou un état	subit un changement d'état
sait ou perçoit	est un thème incrémental
cause un événement ou le changement d'état d'un autre participant	est affecté causalement par un autre participant
se déplace par rapport à un autre participant	est statique par rapport au mouvement d'un autre participant
existe indépendamment de l'événement dénoté par le verbe	n'existe pas indépendamment de l'événement dénoté par le verbe

Tableau 1: Propriétés des rôles proto-agent et proto-patient d'après (Dowty, 1991)

Par ailleurs, nous avons utilisé un autre critère qui confirme l'utilisation des critères proto-agentifs de (Dowty, 1991) : les noms que nous avons considérés comme des agents peuvent toujours être le sujet du verbe dont ils dérivent, contrairement aux autres types de noms, ainsi que le montrent les exemples en . Ce test, en plus des critères de Dowty, nous a donc conduites à conserver comme noms d'agent des noms tels que *dormeur, ronfleur* ou *connaisseur*.

(7) a. Le signataire signe la pétition.

b. *Le destinataire destine la lettre.

c. Le ronfleur ronfle fort.

d. Le connaisseur connaît bien ce vin.

Ainsi, nous avons considéré comme des noms d'agent déverbaux tous les noms :

- dérivant morphologiquement d'un verbe ;
- dénotant un humain ;
- correspondant aux critères proto-agentifs de (Dowty, 1991) ;
- pouvant être le sujet du verbe dont ils dérivent.

3 Constitution automatique de la ressource

La ressource VerbAgent a été constituée au moyen de deux types de méthodes qui permettent de récupérer de façon automatique des couples verbe-nom, dont le nom peut être considéré comme un nom d'agent déverbal tel que cela a été défini dans la section précédente. La première méthode est exclusivement basée sur les propriétés formelles des noms, tandis que la seconde se fonde sur leurs propriétés sémantiques, via les définitions fournies par le dictionnaire *Littré*. Ces deux types d'approches ont été combinés afin de minimiser les problèmes inhérents à chacune. En effet, exploiter les propriétés formelles des noms ne garantit pas la relation sémantique du nom avec le verbe. À l'inverse, une relation sémantique adéquate entre un verbe et un nom ne garantit pas que le second dérive morphologiquement du premier. Ces deux méthodes sont présentées ci-dessous.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

3.1 Heuristiques basées sur les propriétés formelles des noms

La première méthode de détection automatique de noms d'agent déverbaux repose exclusivement sur les propriétés formelles des noms. En français, on peut en effet identifier certains suffixes qui semblent corrélés à la formation de noms d'agent déverbaux, par exemple le suffixe *-eur*, comme dans *danseur* dérivé du verbe *danser*, ou le suffixe *-ant,* comme dans *dirigeant* dérivé du verbe *diriger*. Nous avons manuellement identifié neuf suffixes liés à des règles de formation de noms d'agent déverbaux :

(8) a. *-eur* (*danser* > *danseur*)

b. *-euse* (*chanter* > *chanteuse*)

c. *-rice* (*inspecter* > *inspectrice*)

d. *-eresse* (*défendre* > *défenderesse*)

e. *-aire* (*contester* > *contestataire*)

f. *-ant* (*attaquer* > *attaquant*)

g. *-ante* (*diriger* > *dirigeante*)

h. *-ent* (*adhérer* > *adhérent*)

i. *-ente* (*présider* > *présidente*)

Cependant, certains de ces procédés méritent une discussion. Le suffixe *-aire* par exemple permet de former des adjectifs à partir de noms, comme *planétaire* 'relatif aux planètes', mais il peut également construire des noms, soit à partir de noms, comme *pétitionnaire* 'personne qui signe une pétition', soit à partir de verbes, comme *contestataire* 'personne qui conteste'. Dans certains cas le nom dérivé est ambigu et peut être analysé à la fois comme dérivé d'un nom et d'un verbe, comme *démissionnaire*, qui peut être analysé comme dérivé du verbe *démissionner* avec le sens 'personne qui démissionne' et comme dérivé du nom *démission* avec le sens 'personne qui donne sa démission'. L'analyse des noms suffixés en *-aire* n'est donc pas toujours évidente, mais nous avons souhaité prendre en compte ce suffixe dans la formation de noms d'agents déverbaux parce qu'il nous a semblé qu'un certain nombre de relations entre un verbe et un nom en *-aire* pouvaient être pertinentes dans certaines tâches de traitement automatique des langues. Par exemple dans le cadre de la recherche d'information ou de la tâche question-réponse une relation entre *signer* et *signataire* peut être utile, car elle permet de faire le lien entre une question comme "Qui a signé l'accord de Maastricht ?" et un document-réponse comme "L'Allemagne, la France, la Belgique... sont les principaux signataires de l'accord de Maastricht". C'est pourquoi nous avons inclus le suffixe *-aire* dans la liste des suffixes permettant de construire des noms d'agent déverbaux. De la même façon, nous avons inclus les suffixes *-ent* et *-ente* même si ces procédés ne sont plus productifs actuellement, parce qu'ils constituent un patron régulier jouant le même rôle sémantique que la suffixation en *-eur*, et permettent d'établir la relation entre nom et verbe dans un certain nombre de cas, comme par exemple pour *adhérer-adhérent* ou *présider-président*. En effet, la relation *présider-président* peut permettre, à partir d'une question comme "Qui a présidé le dernier conseil de l'Europe", de récupérer un document-réponse comme "M. *X*, le président du dernier conseil de l'Europe...". C'est la raison pour laquelle nous avons également pris en compte les suffixes *-ent*/*-ente*.

Pour récupérer les noms d'agent sur la base des propriétés formelles des noms nous avons utilisé le lexique *Morphalou.* Celui-ci est un lexique librement accessible de formes fléchies du français, constitué automatiquement à partir de la nomenclature du TLF. Il contient 539 413 formes fléchies correspondant à 68 075 lemmes. La liste de noms d'agent déverbaux a été constituée en deux temps. Nous avons tout d'abord récupéré tous les noms de *Morphalou* se terminant par l'un des neuf suffixes présentés en . Puis nous avons vérifié, pour chaque nom, qu'un verbe formellement proche existait dans le lexique *Morphalou*. La vérification a été effectuée au moyen d'heuristiques basées sur la forme des noms et des verbes. Par exemple lorsque le nom se termine par *-eur* la règle la plus générale permettant d'obtenir le verbe dont il dérive est la suivante :

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

(9) supprimer le suffixe *-eur* puis ajouter *-er*

Cette règle permet par exemple de récupérer le verbe *chanter* à partir du nom *chanteur*. D'autres règles sont nécessaires pour rendre compte de relations formellement plus complexes entre le nom et le verbe, comme pour le nom *formateur* et le verbe *former,* ou le nom *finisseur* et le verbe *finir*. La relation entre *formateur* et *former* est gérée par la règle , et la relation entre *finisseur* et *finir* par la règle .

(10) supprimer le suffixe *-ateur* puis ajouter *-er*

(11) supprimer le suffixe *-isseur* puis ajouter *-ir*

Au total une vingtaine de règles ont été établies, grâce auxquelles 4 067 paires nom-verbe dont le nom se termine par l'un des suffixes mentionnés ci-dessus ont été récupérées. Comme cela a été mentionné plus haut, cette méthode de récupération des noms d'agent déverbaux pose quelques problèmes. En effet, une ressemblance formelle entre un nom et un verbe ne garantit pas que les deux sont morphologiquement reliés. Par exemple la paire *accentuer – accentueur* 'oiseau du genre passereau' est récupérée grâce à la règle alors que le nom *accentueur* n'est pas morphologiquement lié au verbe *accentuer*, mais dérive du latin *accentor*. Dans d'autres cas le nom et le verbe appartiennent bien à la même famille dérivationnelle, mais le nom n'est pas dérivé du verbe. C'est le cas par exemple de la paire *rougir – rougeur* qui est récupérée par l'une des heuristiques établies. Dans ce cas, le nom et le verbe sont bien morphologiquement liés, mais ils ne le sont pas directement : ils dérivent tous deux de l'adjectif *rouge*. À l'issue de cette étape de constitution de la ressource, une validation des paires récupérées est donc nécessaire. La validation sera présentée dans la section .

3.2 Patrons de définition du Littré

Pour compenser les problèmes inhérents à la première méthode de constitution de la ressource, nous avons défini une seconde méthode, fondée sur les définitions des noms fournies par le dictionnaire *Littré*. Pour cela nous avons utilisé le *XMLittré*, une version électronique du *Littré* présentée dans un format XML. Cette ressource contient les données du dictionnaire de la langue française d'Emile Littré, qui comprend 78 423 entrées, et, pour chacune, différentes informations comme la prononciation, la nature, et plusieurs définitions (appelées variantes). Cette ressource ayant été constituée à partir d'un dictionnaire publié à la fin du XIX^{e} siècle ne reflète donc pas l'usage actuel de la langue, et peut contenir des emplois vieillis. Cela ne constitue pas nécessairement un problème pour une analyse morphologique. En revanche, de façon plus problématique, il est certain que nombre de mots sont susceptibles d'être absents du dictionnaire parce que trop récents. Nous avons néanmoins utilisé cette ressource parce qu'elle est libre et diffusée dans un format xml facilement exploitable.

L'extraction de noms d'agent déverbaux à partir des définitions du *Littré* s'est faite en deux étapes. Dans un premier temps nous nous sommes basées uniquement sur la sémantique des définitions. Puis, nous avons ajouté une contrainte morphologique au patron de définition des noms d'agent, afin d'être sûres de ne récupérer que les noms d'agents déverbaux.

Pour extraire de façon automatique les noms d'agent d'après leurs définitions, nous avons, lors de la première étape, uniquement pris en compte la sémantique des noms d'agent. Pour cela nous avons tout d'abord dû repérer la façon dont sont généralement définis les noms d'agent dans le dictionnaire. Nous avons donc étudié les définitions de noms d'agent prototypiques, comme *chanteur, danseur, président, dirigeant*... ce qui nous a permis de repérer deux patrons de définition des noms d'agent : "Celui, celle qui" ou "Celui qui" suivi généralement du verbe base. Ainsi, pour le nom d'agent *chanteur*, l'une des définitions est : "Celui, celle qui chante, qui fait métier de chanter". Grâce à ces patrons de définition nous avons extrait 2 944 noms. Cependant, comme cela a été mentionné plus haut, le patron de définition des des noms d'agent ne garantit pas que le nom est réellement dérivé du verbe qui suit "Celui, celle qui" dans la définition. Par exemple cette méthode d'extraction a retourné des noms d'humains qui ne sont pas dérivés de verbes mais de noms, comme *académicien,* dérivé de *académie,* dont l'une des définitions commence par "Celui qui fait partie d'une société de gens de lettres", ou encore *pianiste,* dérivé de *piano*, et défini comme "Celui, celle qui joue du piano".

C'est pourquoi, nous avons ensuite restreint les noms extraits lors de la première étape, en ajoutant une contrainte morphologique entre le verbe suivant *qui* dans la définition et le nom vedette. En réalité, cette contrainte était formelle plus que morphologique, car elle exigeait simplement que les deux premiers caractères du nom et du verbe soient identiques. Cette seconde étape nous a permis de rejeter les noms comme *académicien* et *pianiste*, dont le verbe suivant *qui* dans la définition ne commence pas par les deux mêmes caractères que le nom, respectivement *ac* et *pi*, mais par *fa* et *jo*. Cette seconde extraction nous a permis de recueillir 1 121 noms.

Certes, cette liste de noms d'agents obtenue après la seconde étape est plus restreinte, et comporte nécessairement des manques. Ainsi, le nom *agresseur* défini comme "Celui qui attaque le premier" n'est pas récupéré parce que sa définition ne correspond pas à la contrainte formelle rajoutée lors de la deuxième étape, alors qu'il s'agit bien d'un nom d'agent dérivé du verbe *agresser*. Mais on peut supposer qu'elle sera plus précise, ce que nous confirmerons par comparaison avec une partie validée manuellement de notre ressource.

Cette liste extraite du Littré devrait nous permettre à la fois de valider les paires verbe-nom établies avec la première méthode, et de les compléter éventuellement avec des noms d'agents qui ne correspondraient pas aux heuristiques ayant permis de récupérer les paires.

4 Validation de la ressource

Pour valider notre ressource, nous avons utilisé plusieurs méthodes, en visant ainsi la meilleure validation possible. Nous avons tout d'abord fait une validation manuelle, puis nous avons vérifié cette validation grâce aux définitions du Littré. Nous avons ensuite utilisé d'une part un réseau de cooccurrences lexicales construit à partir du journal *Le Monde*, et d'autre part les Google Books N-grams. La validation de la ressource n'étant pas encore achevée, nous présentons les méthodes utilisées et les résultats obtenus sur un échantillon de 364 paires verbe-nom, représentant environ 9% de la ressource totale.

4.1 Validation manuelle

En un premier temps nous avons vérifié manuellement que le nom était effectivement dérivé du verbe et qu'il désignait bien un agent tel que nous l'avons défini en section 2. La vérification du lien sémantique entre le nom et le verbe a été réalisée grâce au *TLFi* lorsque le nom était trop rare ou inconnu de nous, par exemple pour *amodiateur* "propriétaire qui cède une terre, une exploitation rurale par amodiation", dérivé du verbe *amodier* "donner à ferme un bien foncier, une exploitation rurale". La validation manuelle de l'échantillon a révélé que 76% des paires de VerbAgent étaient correctes, c'est-à-dire qu'elles étaient bien constituées d'un verbe et d'un nom d'agent dérivé. 24% des paires étaient en revanche incorrectes.

Parmi les erreurs, il est notable que la moitié est constituée de noms en *-ant* ou en *-aire*, qui sont bien dérivés du verbe, mais qui ne dénotent pas un agent, comme *adoucissant* ou *aliénataire*, dérivés respectivement de *adoucir* et *aliéner*. Il est possible que les heuristiques de récupération des noms d'agent incluant ces deux suffixes ne soient pas assez contraignantes d'un point de vue sémantique. Nous les avions pourtant incluses afin de ne pas perdre des noms d'agent comme *dirigeant* ou *signataire*. Cependant il est évident que l'inclusion de ces suffixes engendre du bruit, que nous espérons toutefois éliminer grâce aux autres méthodes de validation. Quant aux autres paires erronées, il s'agit dans 19% des cas de noms en *-eur* qui sont bien déverbaux mais qui dénotent un instrument, comme *accélérateur* ou *aspirateur*. Enfin, les 31% restants sont des erreurs d'analyse comme *actionner – actionnaire* ou *aigrir – aigreur*.

Cette validation manuelle est relativement fiable mais nécessiterait le travail de plusieurs personnes et la confrontation de leurs différentes validations, afin de minimiser au maximum les erreurs de jugement personnel. Cependant une telle validation serait très coûteuse. C'est pourquoi, sur la base de la partie validée manuellement, nous avons essayé de mettre au point une méthode de validation automatique qui nous permettrait de limiter de manière automatique les erreurs engendrées par les heuristiques formelles,

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

et de réduire ainsi la validation manuelle qui restera certainement à faire. Pour cela nous avons comparé les paires verbe-nom créées de manière automatique et validées manuellement avec d'autres ressources.

4.2 Confrontation des méthodes de constitution de la ressource

Tout d'abord, nous avons comparé les paires créées par heuristiques et validées manuellement avec les noms d'agents extraits du *Littré* lors de la première étape, c'est-à-dire sans la contrainte formelle. Cela a fait ressortir 92 noms communs aux deux méthodes de construction de la ressource. Sur ces 92 noms, 87 sont des noms ayant été considérés, lors de la validation manuelle, comme des noms d'agent déverbaux.

Nous avons ensuite comparé les paires créées par heuristiques avec les noms d'agents extraits du *Littré* lors de la seconde étape, c'est-à-dire avec la contrainte formelle entre le nom et le verbe. Nous avons alors obtenu 60 noms communs aux deux méthodes de constitution de la ressource. Mais ces 60 noms étaient tous des noms validés comme corrects lors de la validation manuelle. Les données de ces deux comparaisons sont résumées dans le tableau 2.

	Patron de définition "Celui qui, celle qui" ou "Celui qui" uniquement	Ajout de la contrainte formelle entre le verbe et le nom
Nombre de noms en commun	92	60
Nombre de noms d'agents déverbaux en commun	87	60

Tableau 2: Comparaison des résultats obtenus par l'extraction du Littré avec la méthode à base d'heuristiques

Si l'on compare ces résultats avec la validation manuelle de l'échantillon de VerbAgent, qui comporte 275 couples verbe-nom corrects, cette validation automatique par comparaison avec les données extraites des définitions du Littré ne présente donc pas un très bon rappel. En effet, celui-ci est d'environ 22% pour le second patron. En revanche cette validation est très précise. Le faible rappel s'explique par le fait que certaines définitions de noms d'agents déverbaux ne suivent pas les patrons que nous avons spécifiés, comme *agresseur* par exemple. Mais il s'explique aussi grandement par le fait que certains noms d'agents sont absents du *Littré*, comme *avaliseur*.

4.3 Cooccurrents de *Le Monde*

Une autre ressource qu'il nous a semblé intéressant d'exploiter, et qui était à notre disposition pour le français, est un réseau de cooccurrences lexicales construits à partir de corpus du journal *Le Monde* (Ferret, 1998 :281-288). Ce réseau a été construit sur un corpus de 24 mois du *Monde*, en utilisant une fenêtre de 20 mots, et en ne tenant pas compte de l'ordre au sein des cooccurrences. Seules les cooccurrences de fréquence supérieure à 5 ont été conservées, de sorte que le réseau contient 31 000 mots. Une mesure de cohésion entre deux mots est calculée par estimation de l'information mutuelle. Les cooccurrents d'un mot sont ensuite classés par ordre décroissant de leur valeur de cohésion. Notre hypothèse est que si une paire verbe-nom possède des cooccurrents communs, elle sera reliée sémantiquement, et sera donc plus susceptible d'être issue d'une dérivation.

Nous avons donc extrait, pour chaque paire verbe-nom, leurs cooccurrents les plus proches, et considéré qu'une paire était reliée si elle avait au moins un cooccurrent en commun. Le tableau 3 présente les 10 premiers cooccurrents du nom *chanteur* et du verbe *chanter* dans cette ressource, ainsi que leur valeur de cohésion associée. Cette paire nom-verbe ne présente qu'un cooccurrent commun dans les dix premiers, le nom *crooner*.

10 premiers cooccurrents de *chanteur*		10 premiers cooccurrents de *chanter*	
parolier	0,325	colorature	0,336
raï	0,322	gospel	0,321
crooner	0,320	baudet	0,319
zeppelin	0,318	baryton	0,317
guitariste	0,318	diction	0,316
folk	0,317	**crooner**	0,316
percussionniste	0,316	parlé	0,315
choriste	0,313	psaume	0,314
kabyle	0,313	soprano	0,314
bassiste	0,312	piaf	0,313

Tableau 3: Premiers cooccurrents de la paire *chanteur-chanter*.

Le principal inconvénient de cette méthode est que la taille du corpus est limitée, et de nombreux mots sont absents du réseau. Ainsi, sur l'ensemble de la ressource, seules 571 paires sont retrouvées dans le réseau, c'est-à-dire qu'il n'y a que 571 paires pour lesquelles à la fois le verbe et le nom apparaissent dans le corpus.

Afin d'évaluer la pertinence des cooccurrences, nous avons comparé les paires présentant au moins un cooccurrent commun avec la partie validée de VerbAgent. 39 paires ont été trouvées dans le réseau de cooccurrents, parmi lesquelles 22 ont un cooccurrent commun. Sur ces 22 paires, 20 ont effectivement été validées comme correctes dans VerbAgent, et 2 n'ont pas été validées : *accablant-accabler* et *amusant-amuser*. On peut noter que ces deux paires sont bien reliées sémantiquement et morphologiquement, mais que les noms ne correspondent pas à des noms d'agents. Cette méthode semble donc donner un indice sur la relation entre le nom et le verbe, mais nécessiterait un corpus de plus grande taille pour fournir des résultats plus complets.

Nous avons également commencé à étudier la possibilité de valider des paires en comparant les termes comprenant le verbe et le nom d'agent dans un corpus : ainsi, pour la paire *chanter-chanteur*, les termes *chanter un opéra* et *chanteur d'opéra* fournissent également une indication sur la relation entre le nom et le verbe. Toutefois cette étude est encore en cours, et il est encore trop tôt pour fournir des résultats.

4.4 N-grammes de mots

Enfin, nous avons également utilisé des n-grammes de mots, c'est-à-dire des suites de mots contigus, pour déterminer s'ils pouvaient permettre de valider les paires verbe-nom constituées de manière automatique par les heuristiques. Le corpus utilisé pour cette étude est issu des Google Books Ngrams[7], qui comprend des n-grammes de mots extraits de la numérisation de livres.

Pour réaliser la validation, nous avons dans un premier temps constitué des n-grammes de tous les noms et verbes de VerbAgent, puis nous avons comparé les mots apparaissant dans les n-grammes des nom et verbe constituant une paire. Tout d'abord nous avons extrait, pour les noms, tous les trigrammes

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

constitués d'un nom de VerbAgent, suivi du déterminant *du*, *des* ou *de*, et d'un autre mot. Le trigramme a ainsi la forme "nom+du/des/de+mot", par exemple "dirigeant+de+entreprise". Pour les verbes, nous avons extrait tous les trigrammes constitués d'un verbe de VerbAgent, suivi du déterminant *un/une/le/les/des/son/ses,* et d'un autre mot, de sorte que le trigramme a la forme "verbe+un/une/le/les/des/son/ses+mot", par exemple "diriger+une+entreprise". Puis, pour chaque paire verbe-nom de VerbAgent, par exemple pour la paire *diriger-dirigeant*, nous avons compté le nombre de trigrammes étant des variantes, c'est-à-dire dont le troisième mot est identique, dans l'exemple ci-dessus *entreprise*. Le tableau 4 montre les variantes retrouvées pour la paire *utilisateur-utiliser*; on peut cependant constater que la dernière ligne ne correspond pas réellement à une variante.

trigrammes contenant *utilisateur*	trigrammes associés contenant *utiliser*
utilisateur de services	utiliser ses services
utilisateur de logiciels	utiliser des logiciels
utilisateur de logiciels	utiliser les logiciels
utilisateur de systèmes	utiliser des systèmes
utilisateur de base	utiliser une base

Tableau 4: Trigrammes contenant les mots *utilisateur* ou *utiliser*.

Nous avons ainsi extrait 1 795 variantes de termes, correspondant à 231 paires. L'évaluation sur l'échantillon validé manuellement de VerbAgent montre que 19 paires sont trouvées grâce à cette méthode, dont une seule n'est pas une paire validée.

La méthode semble donc précise, mais elle manque réellement de couverture, et ne semble pas, de ce fait, constituer un bon moyen de valider automatique une ressource constituée de manière automatique.

5 Conclusion et perspectives

Dans cet article, nous avons présenté la constitution automatique d'une ressource morphologique de noms d'agent déverbaux. Puis, à partir d'un échantillon validé manuellement, nous avons présenté différentes pistes envisagées pour mettre au point une méthode de validation automatique qui permettrait de réduire la validation manuelle.

La méthode de constitution de la ressource par heuristiques fondées sur les propriétés formelles des noms et des verbes s'est révélée intéressante dans la mesure où elle semble posséder une bonne couverture du phénomène de formation de noms d'agent déverbaux. Cependant elle engendre également presque 25% de bruit, d'après la validation manuelle de l'échantillon, et nécessite de ce fait une réelle validation, qu'elle soit manuelle, automatique ou semi-automatique.

Les différentes études de validation automatique de la ressource montrent des résultats décevants. Toutefois, ces résultats ne remettent pas en cause les méthodes essayées, mais semblent davantage révéler la difficulté à trouver des méthodes adaptées pour les mots peu fréquents.

Dans l'avenir, nous prévoyons de poursuivre la validation de VerbAgent. Pour cela nous envisageons d'améliorer les méthodes de validation automatique mises en œuvre, notamment par l'augmentation de la taille des corpus considérés. Nous envisageons également de combiner les différentes méthodes utilisées. En effet, dans la mesure où les paires validées par chacune des méthodes sont différentes, une combinaison des méthodes pourrait peut-être permettre d'obtenir un meilleur rappel de validation.

Enfin, nous envisageons également, à plus long terme, de diffuser VerbAgent une fois la validation terminée. Pour cela nous avons commencé une réflexion sur le format de diffusion de la ressource, qui sera certainement dans un format XML et qui respectera au mieux les standards s'appliquant aux ressources langagières, tels que TEI ou LMF par exemple. Mais cette réflexion doit encore être approfondie.

Références bibliographiques

Aronoff, M. (1994). Morphology by Itself. Cambridge : The MIT Press.

Balvet, A., L. Barque et R. Marín (2010). Building a Lexicon of French Deverbal Nouns from a Semantically Annotated Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 1408–1413.

Bernhard, D., B. Cartoni et D. Tribout (2011). A Task-based Evaluation of French Morphological Resources and Tools. *Linguistic Issues in Language Technology, 5(2),* 1–41.

Bonami, O., G. Boyé et F. Kerleroux (2009). L'allomorphie radicale et la relation flexion-construction. In Fradin, B., F. Kerleroux et M. Plénat (éds), *Aperçus de Morphologie*, Saint-Denis : Presses Universitaires de Vincennes, 103–125.

Bouchou, B., et D. Maurel (2008). Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres. *Traitement Automatique des Langues, 49(1)*, 61–88.

Creutz, M., T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, et A. Stolcke (2007). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP), 5(1),* 1–29.

Dal, G. et F. Namer (2010). Les noms en *-ance*/*-ence* du français : quel(s) patron(s) constructionnel(s) ? In Neveu, F., V. Muni Toke, T. Klinger, J. Durand, L. Mondada et S. Prévost (éds), *Actes du CMLF 2010 - 2e Congrès Mondial de Linguistique Française*, EDP Sciences, 893–907.

Davis, A. et J.-P. Koenig (2000). Linking as constraints on word classes in a hierarchical lexicon. *Language 76,* 56–91.

de Loupy, C., P. Bellot, M. El Bèze et P.-F. Marteau (1999). Query expansion and classification of retrieved documents. In *Proceedings of the 7th Text REtrieval Conference (TREC-7),* 443–450.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67(3)*, 547–619.

Ferret, O. (1998). ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage. Thèse de doctorat, Université Paris-Sud.

Fillmore, C. (1968). The Case for Case. In Bach, E. et R. Harms (éds), *Universals in Linguistic Theory,* New-York : Holt Rinehart & Winston, 1–88.

Flaux, N. et D. Van de Velde (2000). *Les noms en français : esquisse de traitement.* Paris : Ophrys.

Fradin, B. (2003). *Nouvelles approches en morpologie*, Paris : PUF.

Fradin, B. et F. Kerleroux (2003). Quelles bases pour les procédés de la morphologie constructionnelle ? In Fradin, B., G. Dal, N. Hathout, F. Kerleroux, M. Plénat et M. Roché (éds), *Les unités morphologiques. Actes du 3e forum de morphologie,* Lille : Presses Universitaire du Septentrion, 76–84.

Hathout, N., et L. Tanguy (2002). Webaffix: Discovering Morphological Links on the WWW. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, 1799–1804.

Hathout, N., F. Namer, et G. Dal (2002). An Experimental Constructional Database : The MorTAL Project. In Boucher, P. (éd), *Many Morphologies*, Somerville : Cascadilla Press, 178–209.

Jacquemin, C., J. Klavans et E. Tzoukermann (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational*

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100324

Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97), 24–31.

Kerleroux, F. (2004). Sur quels objets portent les opérations morphologiques de construction. *Lexique, 16*, 85–123.

Lee, Y.-S (2004). Morphological Analysis for Statistical Machine Translation. In Dumais, S. et S. Roukos (éds), *Proceedings of HLT- NAACL 2004*, 57–60.

Matthews, P.H. (1972). Inflectional Morphology. Cambridge : Cambridge University Press.

Moreau, F. et V. Claveau (2006). Extension de requêtes par relations morphologiques acquises automatiquement. *Information – Interaction – Intelligence, 6(2)*, 31–50.

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues : l'analyseur DériF*. Paris : Hermès-Lavoisier.

Sagot, B (2010). The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Online Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10),* 2744–2751.

Tran, M. et D. Maurel (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues, 47(1),* 115–139.

Van Valin R., et R. LaPolla (1997). *Syntax. Structure, meaning and function.* Cambrige : Cambrige University Press.

Villoing, F. (2009). Les composés V-N. In Fradin, B., F. Kerleroux et M. Plénat (éds), *Aperçus de Morphologie*, Saint-Denis : Presses Universitaires de Vincennes, 175–197.

---

1 Disponible à l'adresse http://www.cnrtl.fr/lexiques/morphalou/

2 Disponible à l'adresse http://alpage.inria.fr/~sagot/lefff.html

3 Disponible à l'adresse http://redac.univ-tlse2.fr/lexiques/verbaction.html

4 Disponible à l'adresse http://sites.google.com/site/nomagesite/

5 Disponible à l'adresse http://www.cnrtl.fr/lexiques/prolex/

6 Disponible à l'adresse http://rali.iro.umontreal.ca/Dubois/

7 Disponible à l'adresse http://books.google.com/ngrams