

TALN 2009 – Session posters, Senlis, 24-26 juin 2009

Analyse automatique des noms déverbaux composés : pourquoi et comment faire interagir analogie et système de règles^(*)

Fiammetta Namer

UMR 7118 « ATILF » - Nancy Université
fiammetta.namer@univ-nancy2.fr

Résumé : Cet article aborde deux problèmes d'analyse morpho-sémantique du lexique : (1) attribuer automatiquement une définition à des noms et verbes morphologiquement construits inconnus des dictionnaires mais présents dans les textes ; (2) proposer une analyse combinant règles et analogie, deux techniques généralement contradictoires.

Les noms analysés sont apparemment suffixés et composés (HYDROMASSAGE). La plupart d'entre eux, massivement attestés dans les documents (journaux, Internet) sont absents des dictionnaires. Ils sont souvent reliés à des verbes (HYDROMASSER) également néologiques. Le nombre de ces noms et verbes est estimé à 5.400. L'analyse proposée leur attribue une définition par rapport à leur base, et enrichit un lexique de référence pour le TALN au moyen de cette base, si elle est néologique. L'implémentation des contraintes linguistiques qui régissent ces formations est reproductible dans d'autres langues européennes où sont rencontrés les mêmes types de données dont l'analyse reflète le même raisonnement que pour le français.

^(*) Cet article a été publié dans les actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles 2009. La publication originale est disponible sur le site www-lipn.univ-paris13.fr/TALN09/

Abstract: This paper addresses two morpho-semantic parsing issues: (1) to automatically provide morphologically complex unknown nouns and verbs with a definition; (2) to propose a methodology combining both rules and analogy, which are techniques usually seen as inconsistent with each other.

The analysed nouns look like both suffixed and compounded (HYDROMASSAGE). Most of them are not stored in dictionaries, although they are very frequent in newspapers or online documents. They are often related to verbs (HYDROMASSER), also lacking from dictionaries. The estimated amount of these nouns and verbs is 5,400. The proposed analysis assigns them a definition calculated according to their base meaning, and it increases the existing reference lexicon content with this base, from the moment that it is a new-coined form. The implementation of linguistic constraints which govern this word formations is reproducible in other West-European languages, where the same data type is found, subject to the same kind of analysis.

Mots-clés : Analyse morphologique, Annotation sémantique, Composition savante, Noms déverbaux, Règles, Analogie

Keywords: Morphological parsing, Semantic annotation, Neo-classical compounds, Deverbal nouns, Rules, Analogy

1 Introduction

Les noms composés et suffixés de la forme $YXSf_{X_N}$ (AEROCONTAMINATION, HYDROMASSAGE) forment un lexique de taille importante, estimé à 4.600 noms (cf. §.2.2). Dans ces noms, Y désigne une racine grecque ou latine (*aér(o)*, *hydr(o)*), X_V un thème verbal (*contaminat*, *mass*) et Sf_x est l'un des exposants de règles formatrices de noms déverbaux de procès, pour l'essentiel l'un des suffixes : *-ion*, *-age*, *-ment*, *-ance*. Nous posons ici la question de leur analyse automatique, et, corrélativement, de celle des verbes YX_V (AEROCONTAMINER, HYDROMASSER), également nombreux dans les documents (cf. §.2.2), auxquels ces noms sont apparentés.

Pour tout système d'analyse non- (ou semi-)supervisé, indépendant des connaissances linguistiques et visant l'identification des mots partageant une séquence commune, cette question ne se pose pas : n'importe quel découpage conduit à reconstituer la « famille » du nom. $YXSf_{X_N}$ va être indifféremment analysé sur XSf_{X_N} (CONTAMINATION, MASSAGE), ou YX_V (AEROCONTAMINER, HYDROMASSER), de manière à être relié, finalement, à X_V (CONTAMINER, MASSER).

Cependant, l'analyseur dont il est question ici (cf. §.3) se donne pour objectif la production d'un lexique annoté des traits syntaxico-sémantiques que la morphologie peut fournir, c'est-à-dire au moins la pseudo-définition du lexème analysé, calculée par rapport au sens de sa base. L'analyse des données est donc fondamentalement guidée par la conformité aux principes linguistiques. Or, ceux-ci préconisent qu'une seule analyse est possible (cf. §.2.2) : contrairement à ce qu'une décomposition linéaire laisserait croire, le verbe composé YX_V (HYDROMASSER) est dénominal, et sa « base » $YXSf_{X_N}$ (HYDROMASSAGE) est composée sur YSf_{X_N} (MASSAGE), ce nom étant lui-même dérivé de X_V (MASSER). L'obtention de résultats linguistiquement motivés se heurte, comme on le voit, à une difficulté méthodologique, puisque les verbes composés s'analysent sur une base linéairement plus complexe.

Cet article aborde en d'autres termes la résolution simultanée de deux problèmes : produire une analyse morpho-sémantique de données nominales et verbales, morphologiquement construites, quantitativement importantes, mais généralement absentes de dictionnaires ; mettre au point une approche analytique devant combiner deux techniques généralement contradictoires. L'article débute par l'exposition des données et des méthodes à mettre en oeuvre (§.2). Nous décrivons tout d'abord la manière dont le corpus a été constitué, et son intérêt en TALN (§.2.1) ; nous montrons au §.2.2 pourquoi l'analyse par règles de (dé)suffixation et (dé)composition des noms $YXSf_{X_N}$, et des verbes apparentés, est au mieux insuffisante, au pire impossible ; et quelle analyse, au contraire, apparaît à la fois linguistiquement plus motivée et algorithmiquement plus économique. Ensuite, le §.3 aborde la mise en oeuvre pratique de ces constats : tout d'abord (§.3.1), nous décrivons l'analyseur morphologique DériF, en rappelant les caractéristiques principales de son fonctionnement, et des informations produites lors de chaque analyse. Nous mettrons l'accent sur l'incompatibilité apparente entre l'approche de DériF et le raisonnement analogique nécessaire à la prise en compte des verbes YX_V . Enfin (§.3.2) nous montrons et illustrons par des exemples comment l'intégration de ce raisonnement dans DériF est rendu possible grâce aux spécificités des données à analyser. Avant de conclure (§.4), des perspectives d'évolution à court terme sont proposés (§.3.3).

2 Données et méthode

2.1 Collecte des données

Les noms de la forme $YXSf_{X_N}$ sont à la fois composés et suffixés. Le suffixe est pour l'essentiel l'une des séquences *-ion*, *-age*, *-ment* et *-ance*. Le constituant X_V est un verbe, et le constituant Y est une base nominale dite 'non autonome', car il s'agit d'une forme issue du grec ou du latin et ne constitue pas un atome pour la syntaxe. Les valeurs les plus fréquentes de Y_N dans les exemples de $YXSf_{X_N}$ observés dans le TLF_{nome} sont reproduites en (1). Le fait qu'elles aient un sens référentiel dénotatif est l'une des raisons qui conduisent à analyser les formes de (1) comme des bases supplétives de lexèmes, et non pas comme des préfixes : ces formes renvoient à des entités, ce qui justifie qu'on leur attribue la catégorie nominale.

- (1) *abdomino, aéro, bathy, bio, cardio, cryo, cuti, dermo, domo, électro, gastro, hémo, hydro, immuno, lacto, latéro, neuro, ophthalmo, osmo, oxydo, photo, radio, rhino, géo, pyro, thermo*

Les $YXSf_{X_N}$ ne constituent que 2% des noms morphologiquement construits présents parmi les nomenclatures du TLF_i ou TLF_{nome}¹. Ils sont plus fréquents dans les corpus journalistiques (d'après une recherche effectuée dans les recueils de textes électroniques issus de *Le Monde 1995 et 1999*, et de *L'Est Républicain 2002*) et leur présence est largement attestée dans les documents en ligne, comme en témoigne le résultat de l'estimation relatée dans ce qui suit. À partir des 2.200 noms déverbaux de la forme $Xion_N$ contenus dans le TLF_{nome}, nous avons automatiquement généré l'ensemble des séquences de la forme *hydroXion*, et absentes du TLF_{nome}. L'une après l'autre, ces séquences ont servi de requête via Yahoo au moyen du robot WaliM (Namer, 2003a). Après éviction manuelle des résultats erronés, cette recherche en ligne conduit à une liste de 237 formes différentes. La même méthode est répétée en remplaçant *hydro* par chacune des autres bases de (1). On obtient un total 3.400 noms, ce qui mène à la conclusion que les $YXion_N$ sont presque 27 fois plus nombreux que ce que le contenu du TLF_{nome} ne laisserait croire. De là, en examinant le nombre (moins important) des noms composés suffixés par *-age*, *-ment* et *-ance*, il est possible d'estimer à 4.600 le nombre total des $YXSf_{X_N}$ à analyser. La même succession automatique de requêtes via Yahoo est réalisée sur la base des candidats verbes YX_V générés au préalable à partir des 3.400 $YXion_N$. Seuls 800 d'entre eux renvoient des résultats non nuls². Les raisons qui expliquent la différence quantitative entre les YX_V et les $YXSf_{X_N}$ sont avancées au §.2.2.

Comme le montre le **Tab.1**, le nom $YXSf_{X_N}$ est analysable comme composé savant³ (1a, 2a), tout en étant apparenté à YX_V (1b); la fréquence d'attestation de YX_V dans les textes varie

¹ Les nomenclatures du TLF_i qui nous servent de référentiel pour la langue générale comportent plus de 99.000 lemmes catégorisés NOM, VERBE, ADVERBE et ADJECTIF.

² La requête a porté sur la disjonction de l'infinitif, des participes et des formes sigulier et 3^e pp des temps de l'indicatif et du présent du subjonctif.

³ Contrairement à la suffixation et à la préfixation, qui consistent en l'application d'une règle à une base, la composition est un procédé de formation lexicale qui fait intervenir deux unités possédant un sens référentiel. Traditionnellement, un composé savant ou néoclassique YX comporte au moins un constituant d'origine grecque ou latine ; dans les cas qui nous intéressent, le composé YX s'interprète comme un hyponyme de X .

d'un verbe à l'autre (1b versus 2b) ; X_V est, lui, un verbe enregistré dans le TLF. Les nombres entre parenthèses quantifient les occurrences recueillies sur La Toile lors de la collecte automatique relatée *supra*. Le cas échéant, l'attestation de $YXSf_{X_N}$ dans le TLF est indiquée.

	$YXSf_{X_N}$ (# occ) / XSf_{X_N}	YX_V (# occ) / X_V
1	(a) hydromassage (148000) / massage thermorégulation (TLF) / régulation	(b) hydromasser (3) / masser thermoréguler (9340) / réguler
2	(a) hémocoagulation (#) / coagulation	(b) *hémocoaguler (0) / coaguler

Tab. 1 - Noms en $YXSf_x$, et verbes apparentés

2.2 Analyse "classique" ?

Au vu des données exposées ci-dessus, le choix à opérer réside dans la manière dont l'analyseur doit rattacher $YXSf_{X_N}$, YX_V et le verbe X_V . Comme cela a été évoqué au §.1, cette décision est sans objet pour toute approche dont l'objectif est de réunir des formes supposées appartenir à une même famille pour améliorer des systèmes de recherche d'information. C'est le cas des raciniseurs (Porter, 1980, Savoy, 1999), ou des systèmes basés sur l'apprentissage (Bernhard, 2007, Claveau et L'Homme, 2005, Gaussier et al., 2000, Hathout, 2003, Moreau et Claveau, 2006, Zweigenbaum et al., 2003). Mais l'analyse présentée ici a pour finalité l'annotation sémantique des lexèmes construits, et en particulier la prédiction de la relation sémantique que tisse le lexème analysé avec sa base. Pour cette raison, l'identification d'une orientation linguistiquement motivée est nécessaire pour l'analyse de $YXSf_{X_N}$ et YX_V . Suivant cette perspective, deux obstacles s'opposent au schéma (2), pourtant intuitif.

$$(2) \quad *YXSf_{X_N} < YX_V < X_V$$

Tout d'abord, si un nom $YXSf_{X_N}$ (HYDRODISTILLATION) était dérivé d'un verbe composé YX_V , alors l'existence de $YXSf_{X_N}$ impliquerait nécessairement celle de YX_V . Or la comparaison des fréquences des YX_V avec celles des $YXSf_{X_N}$ est très en défaveur des verbes, ce qui contredit cette hypothèse. En réalité, certaines contraintes, portant sur la relation entre Y_N et X_V , dictent les conditions de bonne formation de YX_V , alors qu'elles n'affectent pas la formation de $YXSf_{X_N}$ (cf. 2a versus 2b, Tab.1). L'observation de ces contraintes, qui dépassent le cadre de cet article et dont il ne sera pas question ici, suffit néanmoins à démontrer que $YXSf_{X_N}$ (e.g. CARDIORESISTANCE, CUTIREACTION), dont l'existence et la bonne formation ne sont pas tributaires de celles de YX_V (*THERMORÉSISTER, *CUTIREAGIR), n'est pas dérivé de YX_V . Ajoutons que les YX_V sont systématiquement absents du TLF (alors que les $YXSf_{X_N}$ peuvent y figurer), et qu'ils sont défectifs : les seules formes rencontrées sur la Toile sont l'infinitif, les 3èmes personnes singulier et pluriel du présent, et les participes.

Le deuxième obstacle concerne la formation de YX_V . Ce verbe semble résulter d'un procédé de composition à partir du verbe X_V par incorporation du nom Y (HYDRO(=eau)+DISTILLER > HYDRODISTILLER). Or, le procédé dit d'incorporation (Baker, 1988, Mithun, 1984), n'est pas un moyen de construction disponible dans les langues européennes (Creissels, 2005, Haspelmath, 2002) : par exemple, le verbe HYDROMASSER ne s'analyse pas comme le résultat de l'incorporation de la base supplétive nominale °hydro=EAU au verbe MASSER.

En termes d'analyse automatique, ces deux remarques impliquent que l'enchaînement (2) est illégitime puisque l'étape $YXSf_{X_N} < YX_V$ est contredite par les données, et que l'étape $YX_V < X_V$ est illégale dans le système morphologique du français. Ces deux problèmes disparaissent

quand on examine c'est l'analyse inverse (3). Elle suppose que YX_V est construit à partir de $YXSf_{X_N}$ par rétroformation : le construit est une séquence plus « courte » que la base (Becker, 1993, Nagano, 2007). Le nom « de base » $YXSf_{X_N}$ s'analyse, lui, sur XSf_{X_N} par composition savante. XSf_{X_N} est naturellement formé sur base verbale X_V :

$$(3) \quad YX_V < YXSf_{X_N} < XSf_{X_N} < X_V$$

En synthèse, l'application des règles classiques de (dés)affixation/(dé)composition pour l'analyse des noms $YXSf_{X_N}$ de notre corpus est inadaptée, puisque ces règles ne sont pas en mesure d'analyser YX_V , qui n'est ni le produit d'une affixation, ni celle d'une composition. En revanche, un autre raisonnement, utilisé en TALN (il constitue le principe fondamental des systèmes basés sur l'apprentissage de règles, cf. *supra*), constitue un moyen élégant de prendre en compte la rétroformation : il s'agit de l'analogie proportionnelle. Ce raisonnement stipule que YX_V est à $YXSf_{X_N}$ ce que X_V est à XSf_{X_N} (sur l'analogie en morphologie, voir (Dal, 2008, Skousen et al., 2002)).

Il faut donc mettre au point d'une méthodologie d'analyse, qui active un raisonnement analogique pour analyser les verbes de la forme YX_V , et exclusivement dans ce cas. Ce module particulier doit être intégré dans un système de règles, donc apparemment algorithmiquement incompatible. Dans la suite, après un bref rappel des propriétés de l'analyseur utilisé, nous examinons la manière dont un module d'analyse par analogie a néanmoins pu être intégré au système.

3 Mise en oeuvre de la méthode

3.1 Analyseur DériF

DériF (Namer, 2002, Namer, 2003b) est un analyseur morphologique du français basé sur l'application de règles et s'inspirant du courant lexématique de la morphologie (Fradin, 2003). En cela, il se sert de connaissances externes et s'oppose aux systèmes non-supervisés comme ceux de (Gaussier et al., 2000, Goldsmith, 2001, Moreau et Claveau, 2006) par exemple. Les règles sont motivées linguistiquement et diffèrent en cela des raciniseurs/désaffixeurs qui ont surtout été développés pour l'anglais (Hull, 1996). L'application de ces règles garantit la prédiction systématique de relations sémantiques entre les lexèmes apparentés lors d'une analyse, et de l'attribution de traits syntaxiques et sémantiques sur ces mêmes lexèmes.

L'entrée d'une règle est une unité lexicale non fléchie munie d'une catégorie grammaticale (un lexème). La sortie consiste en l'analyse complète de ce lexème jusqu'à obtention d'une nouvelle unité non décomposable (le simple). Cette analyse, dont les étapes sont retracées sous forme d'historique (4a), consiste en la construction de la famille morphologique reliant le lexème analysé au simple (4b), en la définition du lexème analysé en fonction de la valeur du lexème de base (i.e. celui avec lequel il est directement relié morphologiquement) et des propriétés de la règle qui relie ces deux unités (4c). Enfin, l'analyse conduit également à attribuer automatiquement à chaque élément de la famille morphologique l'ensemble des traits qui reflètent les contraintes imposées aux lexèmes par chaque règle morphologique reconnue lors de l'analyse (la technique utilisée est décrite dans (Namer, 2002)). En (4d), la règle reliant *STIMULATEUR* à *STIMULER* dit du premier qu'il désigne un instrument ou un agent, et du second qu'il instancie un prédicat dynamique et agentif.

Analyse automatique de noms déverbaux composés

- (4) a cardio-stimulateur/N => [[card N*] [[stimuler V] eur N] N]
b (cardio-stimulateur/N , [card,N*]:stimulateur/N , stimuler/V)
c " Type particulier de stimulateur en rapport avec le coeur "
d stimuler/V : [aspect=dynamique, sous_cat = <NPagent, ...>]
stimulateur/N : [concret=oui, comptable=oui]

Le principe fondamental de DériF (qui illustre d'une certaine manière celui de tous les analyseurs à base de règles) est que l'activation d'une règle d'analyse dépend de la détection formelle ou catégorielle d'un lexème morphologiquement construit. La démarche générale de détection consiste pour l'essentiel en trois types de tâches, qui sont activées récursivement.

Type1 : repérage d'une séquence identifiable comme un suffixe (e.g. *-able*), sur un lexème muni de la catégorie appropriée à ce suffixe (DETECTABLE_A), et appel du module d'analyse du lexème (DETECTABLE_A < DETECTER_V); chaque module prévoit, lorsque cela est linguistiquement motivé, le déclenchement des procédés d'analyse de la préfixation (INDETECTABLE_A < DETECTABLE mais RECONSTRUCTEUR_N \uparrow ⁴ CONSTRUCTEUR_N) ou de la composition (THERMODETECTEUR_N < DETECTEUR_N < DETECTER_V mais PHOTOSENSIBILISER_V \uparrow SENSIBILISER_V).

Type2 : repérage d'un suffixe, mais sur un lexème appartenant à une catégorie incompatible (PORTABLE_N), et appel du module d'analyse par conversion produisant la catégorie appropriée (PORTABLE_N < PORTABLE_A). De là, les tâches de Type1 peuvent être réenvisagées.

Type3 : En l'absence de suffixe, identification d'un procédé de préfixation (DESHERBER_V < HERBEN) ou de composition (HYDROCEPHALE_A < °céphale=TETEN), à condition évidemment que cette analyse formelle soit conforme au rapport catégoriel attendu par la règle.

L'existence de chaque résultat calculé est vérifiée dans un lexique de référence, qui regroupe l'ensemble des nomenclatures du TLF, et qu'enrichit un ensemble de lexèmes collectés dans divers corpus (au total : 99.093 noms, adjectifs et verbes). Ce lexique est complété par une table des éléments de formation, qui réunit la liste des 1.370 radicaux supplétifs gréco-latins nominaux ou verbaux auxquels la construction du vocabulaire savant sait souvent appel. Quel que soit le type du module d'analyse mis à contribution, il obéit à deux principes :

- | |
|--|
| (5) a le résultat doit être attesté dans le lexique de référence |
| b l'activation d'une analyse se traduit par une « désaffixation » ou une « décomposition » (Type 1,3) ou un changement de catégorie (Type2). |

L'analyse par rétroformation des verbes composés YX_V est paralysée par (5b) qui reflète les fondements théoriques de l'analyseur. De surcroît, (5a) bloque la reconnaissance de leur « base » YXSfx_N, à chaque fois que le nom est absent du dictionnaire (ce qui souvent le cas, cf. §.2.1).

La technique exposée au §.3.2 est mise en œuvre pour améliorer le système d'origine et prendre en compte ces données. L'intégration du module d'analyse par analogie conduit à

⁴ CONSTRUIT \uparrow BASE signifie que CONSTRUIT ne s'analyse pas comme dérivé de BASE (cf. note 5).

contourner (5a) et (5b), mais dans des circonstances contrôlées par les propriétés formelles des unités analysées.

3.2 Résultats : Analyse de $YXSf_{X_N}$ et de YX_V

Les $YXSf_{X_N}$ pris en compte par DériF sont suffixés par *-ion*. En effet, ce sont ces noms de base qui servent le plus souvent à construire les verbes YX_V . Ce choix garantit donc la manipulation d'un corpus important pour la réalisation de l'expérience. Pour ces YX_{ion_N} , la seule analyse ($Y_N + XSf_{X_N}$) se déroule comme décrit au §.3.1 (Type 1), puisque $YXSf_{X_N}$ ne s'interprète que comme composé. Ensuite, DériF se réapplique sur le résultat obtenu : le constituant tête XSf_{X_N} est reconnu comme dérivé de X_V . Ci-dessous, l'exemple (6) résume la suite d'étapes d'analyse reliant $YXSf_{X_N}$ à X_V .

Les verbes en YX_V issus de la collecte décrite au §.2.1 constituent, quant à eux, des instances de formation régressive, que DériF va donc traiter au moyen d'un raisonnement analogique. Étant donné l'approche générale de DériF, ce mécanisme n'est envisageable ici que grâce aux particularités des composants Y et X, qui facilitent les tâches d'identification (et d'analyse) des composés à traiter par analogie. La technique consiste en la succession de trois tâches :

- (T1) la reconnaissance d'un verbe de la forme YX , c'est-à-dire commençant par une séquence appartenant à la liste des éléments de formation de catégorie nominale réunis en (1), cf. §.2.1 ;
- (T2) l'isolation de Y_N et X_V ;
- (T3) la reconstitution de XSf_{X_N} .

Si la suite (T1)-(T3) conduit à un résultat non nul, alors la dernière tâche consiste à concaténer Y_N , isolé en (T2), à XSf_{X_N} , reconstitué en (T3), pour former la « base » du composé verbal YX_V , c'est-à-dire le nom $YXSf_{X_N}$. Ce nom est considéré comme toujours possible : s'il n'est pas enregistré dans le lexique de référence de DériF, le système l'ajoute à une liste complémentaire. Comme l'illustre (7), l'analyse s'accompagne de la glose « Procéder à $YXSf_{X_N}$ ». Finalement, le déroulement complet de l'analyse de $HYDRODISTILLER_V$ donné en (8) est la succession des étapes (7) puis (6) :

- (6) hydrodistillation/N => (hydrodistillation/N , [hydr,N*]:distillation/N , distiller/V) , " Type particulier de distillation en rapport avec le(s) eau "
- (7) hydrodistiller/V=> (hydrodistiller/V, hydrodistillation/N) " Procéder à le—la hydrodistillation "
- (8) hydrodistiller/V=> (hydrodistiller/V, hydrodistillation/N, [hydr,N*]:distillation/N, distiller/V) " Procéder à le—la hydrodistillation "

Dans la suite (T1)-(T3), une seule tâche semble présenter quelque difficulté, il s'agit de la reconstitution de XSf_{X_N} à partir de X_V . En effet, le thème (ou radical) verbal que l'on retrouve dans un nom événementiel diffère du thème verbal de l'infinitif, comme en témoignent les exemples dans le **Tab. 2**, où le nom est suffixé par *-ion*. Dans ce cas (Bonami et Boyé, 2003, Bonami et al., 2009), le thème d'un verbe X_V sélectionné par la RCL-*ion*, dit thème caché, correspond par défaut au radical du supin du verbe latin dont est issu X_V , et n'est jamais sélectionné en flexion.

Thème du verbe à l'infinitif	Thème utilisé par la RCL-ion
distill (DISTILLER)	distillat (DISTILLATION)
réduit (REDUIRE)	réduct (REDUCTION)
fléchi (FLECHIR)	flex (FLEXION)

Tab. 2 : Correspondance entre le thème de X à l'infinitif et son thème caché

La résolution de ce problème d'appariement est facilité par le fait que DériF dispose déjà d'une table de correspondances, dont le **Tab. 2** constitue un échantillon. Cette table, inspirée de la liste Verbaction (Hathout, 2001, Hathout, 2003) comporte environ 2.200 paires formées d'un verbe à l'infinitif et de son thème caché, et est utilisée pour l'analyse des noms en *-ion* et en *-eur*, comme en témoignent, respectivement, la relation formelle entre $STIMULATEUR_N$ et $STIMULER_V$ en (4), et celle qu'établissent $DISTILLATION_N$ et $DISTILLER_V$ en (8). La seule opération à effectuer pour réaliser (T3) est par conséquent la consultation de la table de correspondances inversée.

En somme, l'analyse des YX_V contrevient aux principes de DériF édictés en (5) : les noms de base en $YXSf_{X_N}$ absents du référentiel servent à incrémenter celui-ci (ce qui s'oppose à 5a), et l'activation de l'analyse consiste en la reconstitution du déverbal suffixé $YXSf_{X_N}$, par analogie avec le rapport identifiable entre X_V et XSf_{X_N} (ce qui contourne 5b). Le déclenchement du module d'analyse par analogie étant maîtrisé par l'identification formelle d'un élément de formation Y_N sur une forme YX_V , les risques du surgénération (ou de boucle infinie) sont écartés.

3.3 Bilan, Perspectives

Pour valider intégralement cette étude, dont les résultats n'ont été obtenus que pour le corpus composé des 3.400 noms en *-ion* et des 800 verbes apparentés, il est nécessaire de réitérer le processus de collecte pour les noms en $YXage_N$, $YXment_N$, $YXance_N$ de manière à vérifier que le nombre de ces données corrobore l'estimation proposée au §.2.1. Il apparaît, lors de premières expériences, que les noms dérivés de verbes statifs (*ABONDANCE*, *EXISTANCE*) sont des composants le plus souvent impropres à la formation de $YXSf_{X_N}$ (*GEOAPPARTENANCE* constitue l'une des rares exceptions). On constate également que d'autres classes sémantiques de prédicats, comme les verbes d'apparition ou d'émission de substance, cf. (Levin, 1995) sont des bases possibles pour les noms en $YXSf_{X_N}$ (*ELECTROEMISSION*) mais le verbe YX_V apparenté n'existe pas. Ces observations nous portent à formuler des hypothèses sur les conditions de formation de YX_V , qui, si elles sont avérées, conduisent en retour à prédire les propriétés aspectuelles du verbe X_V qui en est à l'origine. Ainsi, on peut d'ores et déjà inférer la dynamicité du prédicat incarné par X_V de l'existence de $YXSf_{X_N}$. Une étude plus complète, portant sur l'ensemble des YX_V et des $YXSf_{X_N}$ répertoriés sur la Toile, doit permettre de prédire, suivant l'attestation ou non du verbe et/ou du nom composé, des valeurs plus précises portant sur l'aspect et la transitivité pour le verbe X_V dont dérivent ces composés. Une autre source d'information est apportée par les relations actantielles que tisse Y_N avec X_V , et la fréquence avec laquelle chaque relation est observée : Y_N joue-t-il majoritairement le rôle du patient de X_V (*CUTI-REACTION*, *HYDROPOMPAGE*, *THERMOREGULER*) ? Dans ce cas, on peut présager que X_V est soit transitif, soit inaccusatif. Y_N fonctionne-t-il au contraire essentiellement comme ajout instrumental de X_V (*THERMOCOMPRIMER*, *ELECTRONETTOYAGE*) ? Cela pourrait signifier que X_V désigne un accomplissement (la finalité du prédicat nécessitant l'intervention de cet instrument). En somme, le croisement des

informations apportées par l'existence et la fréquence des YX_V et des $YXSfx_N$, avec celles déduites de la structure interne de ces composés va servir à enrichir le potentiel d'annotations syntaxico-sémantiques sur X_V , suivant le mécanisme d'affectation de traits appliqué lors de chaque analyse par DériF sur le construit et sa base, comme cela est illustré au §.3.1, exemple (4d).

4 Conclusion

Nous avons proposé une expérience d'analyse de noms et verbes morphologiquement complexes, le plus souvent néologiques, formant un corpus estimé à 5.400 unités. En marge de l'analyse de ces lexèmes, qui combine décomposition morphologique et annotation sémantique, le système réalise en outre l'enrichissement automatique d'un lexique de référence au moyen des bases calculées par l'analyseur, quand elles sont néologiques. Cet ajout automatique reflète la disponibilité systématique de la base, en tant que lexème que le locuteur peut vouloir créer à tout moment. L'originalité de l'approche présentée réside dans le fait qu'analogie et règles collaborent, ce, grâce à l'identification de marqueurs formels caractérisant les données à traiter. L'objectif est de fournir à ces données une analyse linguistiquement motivée accompagnée d'une définition calculée à partir du sens de leur base morphologique. Les méthodes et résultats présentés dans cet article nous ont permis de confirmer une fois de plus le rôle de la Toile dans la découverte de données lexicales massives (Hathout et al., 2009)⁵, données qui pourtant illustrent des phénomènes peu pris en compte en traitement automatique, et considérés comme périphériques par les théories morphologiques. La coopération, même limitée, entre analogie et règles indique comment un système peut mettre en œuvre un raisonnement psycholinguistiquement plausible, tout en privilégiant autant que possible les solutions d'analyse les plus efficaces.

Enfin, il est intéressant de remarquer que les noms et verbes composés dont il a été question ici partagent deux propriétés qui se recoupent : (1) ils appartiennent le plus souvent, en tant que composés néoclassiques, à des vocabulaires spécialisés variés (médecine, chimie, botanique, aéronautique, etc...) ; (2) ils existent dans plusieurs langues romanes (voire de toute l'Europe de l'ouest), avec la même structure, la même fréquence d'apparition, les mêmes conditions de formation, les mêmes spécificités morphosémantiques. De ce fait, le raisonnement analogique mis en pratique pour le français devrait être aisément généralisable.

Références

- BAKER M.C. (1988). *Incorporation A theory of Grammatical Function Changing*. Chicago: UCP.
- BECKER T. (1993). Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology. *Yearbook of Morphology* 1992:1-27.
- BERNHARD D. (2007). Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique. *TALN-2007*, Toulouse:367-376.

⁵ Au sujet les discussions portant sur l'usage de la Toile comme corpus et les précautions à prendre, voir entre autres (Grefenstette, 1999, Lüdeling et al., 2007, Resnik et Elkiss, 2005). Sur l'usage de la Toile dans une perspective d'acquisition morphologique, se référer par exemple à (Namer, 2003a) ou (Hathout et Tanguy, 2005).

Analyse automatique de noms déverbaux composés

- BONAMI O., BOYE G. (2003). Supplétion et classes flexionnelles. *Langages* 152:103-126.
- BONAMI O., BOYE G., KERLEROUX F. (2009). L'allomorphie radicale et la relation flexion-construction. In *Aperçus de Morphologie du français*, eds. B. Fradin, F. Kerleroux & M. Plénat, Paris: Presses Universitaires de Vincennes, 103-126.
- CLAVEAU V., L'HOMME M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie - Utilisation comparée de ressources endogènes et exogènes. *TIA*, Rouen.
- CREISSELS D. (2005). *Syntaxe Générale, une introduction typologique 2 : la phrase*: Langues et syntaxe. Paris: Hermès-Lavoisier.
- DAL G. (2008). L'analogie dans le domaine du lexique construit : un retour ? *1er Colloque Mondial de Linguistique Française*, Paris:1575-1587.
- FRADIN B. (2003). *Nouvelles approches en morphologie*. Paris: PUF.
- GAUSSIER E., GREFENSTETTE G., HULL A. D., ROUX C. (2000). Recherche d'information et traitement automatique des langues. *Traitement Automatique des Langues* 41:473-493.
- GOLDSMITH J. (2001). Unsupervised Learning of Morphology of a Natural Language. *Computational Linguistics* 27:153-198.
- GREFENSTETTE G. (1999). The WWW as a Resource for Example-Based MT Tasks. *ASLIB 'Translating and the Computer' Conference*, London.
- HASPELMATH M. (2002). *Understanding Morphology*: Understanding Language. London: Arnold.
- HATHOUT N. (2001). Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. *TALN-2001*, Tours.
- HATHOUT N. (2003). L'analogie, un moyen de croiser les contraintes et les paradigmes. *Revue d'Intelligence Artificielle*:923-934.
- HATHOUT N., TANGUY L. (2005). Webaffix : une boîte à outils d'acquisition lexicale à partir du Web. *Revue Québécoise de Linguistique* 32:61-84.
- HATHOUT N., NAMER F., PLENAT M., TANGUY L. (2009). La collecte et l'utilisation des données en morphologie. In *Aperçus de Morphologie du français*, eds. B. Fradin, F. Kerleroux & M. Plénat, Paris: Presses Universitaires de Vincennes, 267-287.
- HULL A. D. (1996). Stemming Algorithms - A case study for detailed evaluation. *Journal of the American Society of Information Science* 47:70-84.
- LEVIN B., RAPPAPORT HOVAV M. (1995). *Unaccusativity*. Cambridge, MA: MIT Press.
- LÜDELING A., EVERT S., BARONI M. (2007). Using Web Data for Linguistic Purposes. In *Corpus Linguistics and the Web*, eds. M. Hundt, N. Nesselhauf & C. Biewer, Amsterdam: Rodopi, 7-24.
- MITHUN M. (1984). The evolution of noun incorporation. *Language* 60:847-894.
- MOREAU F., CLAVEAU V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. *Revue I3 (Information - Interaction - Intelligence)* 6:31-50.
- NAGANO A. (2007). Marchand's analysis of back-formation revisited: back-formation as a type of conversion. *Acta Linguistica Hungarica* 54:33-72.
- NAMER F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. *TALN-2002*, Nancy, France:235-244.
- NAMER F. (2003a). WaliM : valider les unités morphologiquement complexes par le Web. In *Silicales 3 : les unités morphologiques*, eds. B. Fradin et al., Villeneuve d'Ascq: Presses Universitaires du Septentrion, 142-150.
- NAMER F. (2003b). Automatiser l'analyse morpho-sémantique non affixale : le système DériF In *Cahiers de Grammaire*, eds. N. Hathout, M. Roché & N. Serna, Toulouse: ERSS, 31-48.
- PORTER M. F. (1980). An algorithm for suffix stripping. *Program* 14:130-137.

- RESNIK P., ELKISS A. (2005). The linguist's search engine: an overview. *43rd Association for Computational Linguistics (ACL)*, Ann Arbor, MI:33-36.
- SAVOY J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science* 50:944 - 952.
- SKOUSEN R., LONSDALE D., PARKINSON D. B. eds. (2002). *Analogical Modeling*. Amsterdam/Philadelphia: Benjamins.
- ZWEIGENBAUM P., HADOUCHE F., GRABAR N. (2003). Apprentissage de relations morphologiques en corpus. *TALN-2003*, Batz-sur-Mer:285-294.