Cahiers de l'ILSL-45 (2016 : 123-160)

La question des données en morphologie

Nabil HATHOUT
Université de Toulouse & CNRS (F)
nabil.hathout@univ-tlse2.fr

1. Introduction

Je présente dans cet article un ensemble de réflexions qui ont été menées sous l'impulsion de Marc PLÉNAT, au sein du laboratoire ERSS puis de la composante CLLE/ERSS, sur les données qu'il convient d'utiliser en morphologie[1]. Cette question générale peut être déclinée selon différents points de vue relatifs à l'objet d'étude, les jeux de données, leur création, leur origine ou leur devenir :

- Quelle est la nature des données en morphologie ? Quels sont les objets auxquels s'intéresse la morphologie ?
- De quelles données a-t-on besoin pour réaliser des recherches en morphologie, pour l'analyse des phénomènes morphologiques ?
- Comment peut-on obtenir ces données ? Où est-il possible de les collecter ?
- Comment rentabiliser les efforts nécessaires à la constitution des jeux de données morphologiques[2] ?

---

[1] Les travaux et les réflexions présentés dans cet article ont été réalisés et menés en collaboration avec Marc PLÉNAT, Ludovic TANGUY, Fiammetta NAMER et Fabio MONTERMINI. Ils ont fait l'objet de plusieurs publications, notamment (HATHOUT, PLÉNAT & TANGUY 2003 ; HATHOUT, NAMER, PLÉNAT & TANGUY 2009 ; HATHOUT, MONTERMINI & TANGUY 2008 ; HATHOUT & NAMER, 2014a, 2014b). Mon texte reprend en grande partie les trois dernières références et la présentation de Démonette dans la quatrième. Le lecteur intéressé pourra également se reporter à PLÉNAT (2000), PLÉNAT *et al.* (2002) ou TANGUY (2012, 2013).

[2] Dans ce texte, j'utilise indifféremment les termes de « jeux de données morphologiques » et « collections d'exemples » pour désigner les données prêtes à l'emploi, directement exploitables dans le cadre de recherches en morphologie.

Les réponses que l'on peut apporter à ces questions dépendent du type d'analyse morphologique que l'on souhaite réaliser. En morphologie descriptive, les données jouent un rôle central dans les analyses, ce qui oblige les morphologues à consacrer à leur constitution une part importante de leur travail. La qualité des analyses descriptives dépend en effet directement de celle de données. En morphologie, comme dans les autres sciences, l'analyse vise à organiser les données en catégories et à identifier les propriétés qui caractérisent leurs éléments. Il faut donc que les données utilisées soient suffisamment variées pour contenir des représentants de chacune des catégories pertinentes pour le phénomène étudié. Par ailleurs, la **quantité** de ces données doit être suffisante pour permettre l'identification des régularités qui s'établissent dans ces catégories et qui interviennent dans la compréhension et l'explication du phénomène.

Les nouvelles technologies et le Web ont amélioré très significativement la quantité et la qualité des données rassemblées et utilisées par les morphologues du fait d'un accès plus facile à des ressources volumineuses et au développement d'outils permettant de les exploiter. La variété des données est également améliorée par la possibilité d'observer des productions langagières peu normées présentes dans les forums de discussion, les tweets, etc. Mais ces évolutions ont un cout qui n'est pas négligeable : les morphologues se trouvent en effet aujourd'hui dans l'obligation de prendre en compte des quantités énormes d'exemples dont le traitement est à la fois long et fastidieux ; les exemples collectés sont fortement bruités et nécessitent un nettoyage soigneux suivi d'une préparation préalable à toute utilisation (formatage et annotation). Une autre conséquence de l'augmentation de la quantité des données disponibles est un changement dans la nature même du travail des morphologues : la recherche

devient plus expérimentale ; une part du travail de plus en plus grande est consacrée à la collecte et au traitement des données ; les exemples décrits dans la littérature perdent de leur importance. Ce surcout soulève plusieurs questions :

- A-t-on réellement besoin d'autant de données ? Est-il rentable d'utiliser plus de données ? Les analyses morphologiques sont-elles significativement améliorées par la masse des exemples collectés ?
- Comment préserver les données collectées, nettoyées, préparées et analysées ? Cette question concerne notamment les formats, la complétude des jeux de données et les dépôts dans lesquels il faut les placer. Se pose en outre la question du cout de la mise au format et de la complétion des données.
- Quelles sont les utilisations possibles des données morphologiques une fois terminée l'analyse du phénomène étudié ? Une collection d'exemples peut-elle servir à d'autres études que celles pour laquelle elle a été créée ?
- Comment favoriser la constitution de jeux de données réutilisables ? Comment les diffuser ? Dans quelles conditions ? Sous quelle licence ?
- Peut-on mieux rentabiliser le temps et les efforts consacrés à la création des jeux de données ? Comment faire en sorte que la constitution d'une collection d'exemples ne soit pas seulement une sous-tâche préalable dans l'analyse d'un phénomène particulier ? Les jeux de données peuvent-ils avoir une valeur en eux-mêmes ? Comment mieux faire reconnaitre ce travail, notamment par les tutelles et les instances d'évaluation ?

Les réponses à cette dernière question passent par une meilleure considération du travail de constitution de ressources et de collections de données morphologiques. Actuellement, elles sont souvent considérées comme un

sous-produit de l'analyse morphologique sans réelle valeur. Le changement de conception passera probablement d'abord par l'usage : la mise à disposition systématique des données sur lesquelles reposent les analyses favorisera leur utilisation par d'autres chercheurs pour de nouvelles études. Un changement dans la politique éditoriale des journaux de morphologie et plus généralement de linguistique sera aussi nécessaire : il est essentiel de publier davantage d'articles consacrés à la description des jeux de données disponibles.

2. La nature des données utilisées en morphologie

Le point d'entrée pour toute recherche en morphologie – qu'elle soit morphématique ou lexématique – est le mot. Dans le premier cas, l'analyse morphologique vise à décomposer le mot en morphèmes et à organiser ces derniers dans une structure généralement arborescente comme en (1). Dans le second, l'analyse morphologie cherche à identifier les relations de forme et de sens qui s'établissent entre les mots. Ces relations décrivent notamment l'histoire dérivationnelle des lexèmes comme en (2).

(1) *décomposition* : [[dé- [composer]$_V$]$_V$ -ion]$_N$
(2) *décomposition* : décomposition$_N$ → décomposer$_V$ → composer$_V$

2.1 Les mots

Avoir les mots comme objet d'étude apporte à la morphologie un avantage considérable sur d'autres sous-disciplines de la linguistique car ces unités peuvent être facilement collectées dans les textes. Les mots sont notamment très faciles à identifier et à traiter. Leur identification repose sur une approximation de la réalisation du lexème par le mot graphique que l'on définit très simplement comme une chaine de caractères délimitée par des espaces ou des signes de ponctuation. Les mots graphiques sont par ailleurs faciles à manipuler. Une simple

substitution permet par exemple de transformer la chaine `décomposition` en `décomposer`. Un autre avantage des mots est la correspondance très régulière qui existe dans des langues comme le français entre les graphies et les formes phonémiques avec un corolaire intéressant : les mots construits par une règle de construction de lexèmes particulière peuvent être identifiés avec une grande précision à partir de leurs graphies. Par exemple, un mot qui se termine par la sous-chaine `ion` a de fortes chances d'être un nom déverbal construit par suffixation en *-ion*, comme c'est le cas de *décomposition*. Par ailleurs, la graphie d'une forme fléchie ne varie pas en fonction du contexte dans lequel elle apparait, contrairement par exemple aux structures syntaxiques. Au singulier, *décomposition* s'écrit toujours `décomposition`. À l'inverse, la construction *dire que* suivie d'une proposition peut se réaliser de façon contigüe ou non-contigüe comme en (3)

(3) a. Elle **dit qu'**il reviendra.
b. Elle ne **dit** pas **qu'**il reviendra.
c. Elle m'a **dit** à plusieurs reprises **qu'**il reviendra.

Les principales conclusions que l'on peut tirer des remarques précédentes sont d'abord que les documents écrits sont particulièrement bien adaptés à la collecte de données pour la morphologie, y compris dans des textes qui ne peuvent pas être utilisés pour la recherche en phonologie ou en syntaxe. Dans le premier cas, les limitations sont principalement dues à la rareté des enregistrements oraux retranscrits phonologiquement. Dans le second, c'est l'impossibilité d'exploiter les très nombreux textes en français peu voire pas normé, disponibles dans les forums de discussion ou sur les plateformes de microblogage comme Twitter. Si la syntaxe de ces productions est souvent difficile à analyser du fait des interférences avec l'organisation discursive de ces textes, la graphie des mots est en général suffisamment standard pour permettre de les reconnaitre et

de les interpréter comme dans l'exemple (4a) où *gentillable* est clairement construit par suffixation en *-able* sur l'adjectif *gentil* dont il semble être ici un synonyme. Le second extrait (4b) met en évidence les connotations qui lui sont associées.

(4) a. c vré vré vré vrement tres **gentillable** de ta par...merci ma cherie..serieu ta du metre trop longtemps pour faire cet article... `http ://f-ceriiise-f.skyrock.com/...`

b. elle c une fille incroyable formidable **gentillable** et inoubliable (sa existe gentillable??)... `http ://latitmarionette.skyrock.com/1049941156-Lelex.html`

Sans prendre ici position sur la qualité (ou la recevabilité) de ces exemples (voir aussi section 3.2), il est indéniable que les extraits en (4), difficilement exploitables par la syntaxe et par la phonologie, sont susceptibles d'intéresser un morphologue qui envisage de réaliser une analyse sur les adjectifs en *-able*. Cet adjectif n'est pas présent dans les données utilisées par HATHOUT *et al.* (2003) ; aucune étude de cette suffixation ne prévoit ce type de dérivé ; le statut marginal de cet exemple peut être relativisé en exhibant d'autres adjectifs en *-able* construits sur des bases adjectivales comme *difficilable* (5a), *facilable* (5b), *seulable* (5c) ou *tristable* (5d) même s'ils sont relativement difficiles à interpréter.

(5) a. Xerath n'est pas non plus très **difficilable** à prendre en main `ffrlol.forumgratuit.org`

b. Je nsuis po... zune fille facilement **facilable**... `gniuuu.skyrock.com`

c. Dire qu'à une époque, les garçons faisaient la queue pour me bécoter l'oreille et que je me retrouve aujourd'hui plus **seulable** pour le restant des mes jours. `indra-nimportkoi.blogspot.fr`

d. ..Mdrrrrrrrrrr. ..vraiment c'est **tristable** hein.... En plus il te largue salement comme ça et tu oses encore le pleuré. ...otula oooh nini (tu galère niveau chopage de ... `/fr-fr.facebook.com/LaGossiPeuZe`

L'accumulation de tels exemples complique la tâche du morphologue, qui ne peut les ignorer en bloc en les déclarant simplement ininterprétables ou agrammaticaux (voir BAUER 2014 pour une discussion sur les conséquences de l'utilisation des grands corpus sur les notions de grammaticalité et de productivité).

2.2 L'évolution des ressources et du nombre des exemples

Les exemples précédents sont des exemplaires caractéristiques du type de données que l'on peut actuellement collecter pour l'étude d'un phénomène morphologique. Auparavant, c'est-à-dire avant la création du Web, les recherches en morphologie s'appuyaient sur des relevés effectués dans des ouvrages imprimés et des dictionnaires, à un cout prohibitif en temps puisqu'il fallait lire dans leur intégralité la totalité des livres du corpus pour compiler les listes de mots sur lesquels portaient les études. Ce cout était d'autant plus élevé que le nombre des exemples intéressants contenus dans les livres est généralement très faible. En contrepartie, ces exemples étaient irréprochables car provenant de textes édités qui ont subi des révisions nombreuses. À titre d'exemple, on trouve **183** adjectifs ayant la finale `able` dans un corpus d'environ 800 000 mots composé des huit romans de la base Frantext-démonstration parus entre 1803 et 1908. La collecte d'exemples dans les dictionnaires exige un effort moindre car la lecture se limite à la seule nomenclature. Ainsi, la nomenclature de la 8^{e} édition du dictionnaire de l'Académie (31 934 entrées) contient **444** adjectifs se terminant par `able`. Un premier changement d'échelle a eu lieu à partir des années 1990, lorsque les morphologues ont eu accès à des corpus électroniques comme Frantext, mis en ligne en 1992, et à des dictionnaires informatisés comme le *Trésor de la Langue Française informatisé* (TLFi), à partir de 2000. La première évolution concerne le temps nécessaire à la collecte des exemples, qui

est réduite à quelques minutes, voire quelques secondes. La seconde est l'augmentation du nombre d'exemples et leur plus grande diversité. Les données extraites des dictionnaires électroniques sont les mêmes que celles qui se trouvent dans les versions imprimées. On peut par exemple extraire du TLFi **1034** adjectifs en `able` que l'on peut utiliser sans révision ni correction. Ce n'est pas le cas des données provenant d'autres types de corpus, notamment journalistiques qui s'avèrent fortement bruités et qui imposent une vérification et un nettoyage de chacun des exemples qui en sont extraits. Signalons que le rendement de ces corpus n'est pas toujours très élevé : 5 années d'archives du quotidien *Libération* couvrant la période 1995-1999 (87 millions de mots) contiennent **767** adjectifs finissant en `able` tandis que 4610 documents (227 millions de mots) de la base Frantext-intégral (consultée en mars 2015) en fournissent près de **2900**[3]. Ces quelques exemples de ressources imprimées et électroniques donnent une idée de l'évolution de leur taille et de leur capacité à fournir aux morphologues un nombre croissant d'exemples.

À partir de la fin des années 1990, les morphologues ont eu accès à ce que l'on peut considérer comme la ressource ultime : le Web. Cette source d'exemples innombrables présente des caractéristiques uniques qui la rendent incontournable, facilement accessible mais relativement difficile à utiliser et à exploiter (GREFENSTETTE 1999 ; KILGARRIFF & GREFENSTETTE 2003 ; HATHOUT *et al.* 2009a ; TANGUY 2013).

[3] Le nombre exact de ces adjectifs ne peut pas être calculé informatiquement car l'information catégorielle n'est pas disponible pour les listes de mots.

Ses principaux avantages sont :

> 1/ sa taille exceptionnelle. Le Web est sans nul doute la plus grande collection de documents disponibles même si l'on ne peut accéder qu'à une fraction des pages existantes. En effet, l'accès à cette ressource ne peut se faire que par l'intermédiaire des moteurs de recherche qui n'enregistrent dans leurs index qu'un petit sous-ensemble du Web (TANGUY 2012). Cette fraction reste néanmoins d'une taille qui dépasse celle de tout autre corpus. Or la taille d'une ressource détermine sa richesse, tant en nombre de lexèmes différents, d'emplois différents pour ces lexèmes et par suite de sens différents.
>
> 2/ la diversité des documents sur les plans diatopique (influence des substrats régionaux), diastratique (sociolectes) et diaphasique (styles et registres de langue). On trouve également sur le Web énormément de documents techniques qui relèvent d'un grand nombre de domaines de spécialité, même si certains comme l'informatique sont mieux représentés que d'autres. Le Web enregistre notamment un nombre exceptionnel de discussions informelles sur les forums, les blogs, les plateformes de microblogage, etc. qui donnent accès à une langue très spontanée, souvent peu normée, dans laquelle la créativité lexicale est forte. Ces types de textes ne sont (et n'ont jamais été) disponibles dans aucune autre ressource.
>
> 3/ le Web fournit un accès rapide voire immédiat aux évolutions des langues, par exemple à l'explosion des constructions en *-itude* en 2007 et à l'extension des conditions d'emploi de cette suffixation qui l'a accompagnée (KOEHL 2012a ; KOEHL & LIGNON, 2014).

Il faut néanmoins garder à l'esprit que le Web n'est pas un corpus. Nul ne connait sa composition, ne dispose d'un inventaire des documents qu'il contient, ni de leurs caractéristiques fondamentales comme leur date et lieu de publication, le nom, l'âge, la nationalité ou le sexe de leurs auteurs, la langue dans laquelle ils sont rédigés, le niveau de maitrise de cette langue par le ou les auteurs, l'utilisation

éventuelle d'outils de traduction automatique, etc. Il n'est pas équilibré comme l'est par exemple le British National Corpus. Il n'est représentatif de rien, sinon de lui-même. La taille du Web n'est pas connue et ne peut être mesurée. Du fait de sa constante évolution, les expériences réalisées sur le Web ne sont généralement pas reproductibles. Il est impossible d'obtenir la liste des mots utilisés dans le Web. Les moteurs de recherche ne donnent accès qu'à une sélection des documents qu'ils ont indexés et l'interrogation ne peut s'effectuer qu'au moyen de formes (mots simples ou de séquences de mots). On ne peut pas obtenir l'ensemble des pages indexées par un moteur de recherche qui contiennent un mot donné. Google annonce par exemple que son index contient plus de 4,3 millions de documents comprenant le mot *décomposition* mais limite à 406 la liste des résultats affichables. Malgré ces limitations, le Web reste une mine d'exemples inégalable qui nous a notamment permis, en 2002, de collecter près de **4000** adjectifs en *-able* ne figurant pas dans le TLFi.

3. « More data is better data »

Il apparait de façon implicite dans les chiffres présentés ci-dessus que la taille des corpus et le nombre des exemples sont des facteurs importants pour les études en morphologie et que l'on pourrait reprendre un slogan bien connu de la linguistique de corpus : *More data is better data* (CHURCH & MERCER 1993) ou sa version français « gros c'est beau » (PÉRY-WOODLEY 1995). Plus précisément, l'approche extensive en morphologie (PLÉNAT 2000 ; HATHOUT *et al.* 2003) consiste à fonder les analyses morphologiques sur le plus grand nombre possible d'exemples. Elle considère en effet que la quantité d'exemples pris en compte détermine directement la qualité des analyses et que ces derniers doivent être collectés de manière systématique. De leur nombre dépend la bonne compréhension des procédés et des phénomènes étudiés.

Les analyses des phénomènes qui sont fondées sur de grands nombres d'exemples sont plus fines et rendent mieux compte des données moins centrales, plus « exceptionnelles ». Les recherches menées à l'ERSS sur les adjectifs en *-esque* et en *-able* illustrent parfaitement les progrès qui ont été rendus possibles par la morphologie extensive.

3.1 Les voyelles moyennes devant *-esque*

La suffixation en *-esque* construit des adjectifs dont les bases peuvent être des noms communs (6a) et des noms propres (6b). Elle a fait l'objet de plusieurs études menées à l'ERSS pendant une dizaine d'années par PLÉNAT et ses collaborateurs. Ces dérivés constituent en effet un matériau adapté à l'étude des contraintes dissimilatives, qui constitue l'objet réel des recherches de PLÉNAT (2011). Ces contraintes pénalisent l'apparition à faible distance de phonèmes identiques ou similaires.

(6) a. sultan → sultanesque
b. Molière → moliéresque

Dans ces adjectifs, PLÉNAT s'intéresse principalement au comportement des voyelles moyennes antérieures (/e, ɛ, ø, œ/) qui se trouvent à la fin des radicaux des dérivés en *-esque* et qui sont suivies d'une consonne fixe (i.e. non-latente) comme en (7) :

(7) a. Cervantes → cervantesque
b. enchanteur → enchanteuresque

Ces exemples montrent que dans certains dérivés, la rime tombe mais pas dans d'autres. Quels sont les conditions du maintien ou de la chute de la rime ?

Une consultation du TLFi permet de collecter 104 adjectifs dérivés en *-esque* qui ne posent aucun problème car ils sont formés par simple concaténation du thème de la base et de l'exposant du suffixe, comme en (8).

(8) a. Molière → moliéresque
b. Raphaël → raphaélesque

PLÉNAT a entrepris avec l'aide de SERNA une collecte systématique de dérivés en *-esque*, notamment dans les romans de San Antonio. En 1997, 800 dérivés ont ainsi été rassemblés, qui font apparaitre que les rimes en /ɛ/ suivies d'une consonne fixe peuvent tomber lorsque la base comporte au moins quatre syllabes (9). Les bases de trois syllabes qui finissent en /s/ sont normalement raccourcies (10). PLÉNAT est cependant intrigué par l'exemple (11) dans lequel la finale *-eur* est supprimée alors que la base ne comporte que trois syllabes.

(9) a. Pantagruel → pantagruesque
b. consommateur → consommatesque
(10) a. Cervantes → cervantesque
b. cosinus → cosinesque
(11) tirailleur → tiraillesque

La collecte s'est poursuivie pendant encore quelques années. En 2001, PLÉNAT dispose de plus de 3000 dérivés qui confirment que les rimes composées d'une voyelle moyenne antérieure et d'une consonne fixe tombent dans les radicaux de quatre syllabes ou plus. De même, les rimes dont la consonne finale est identique à l'une des consonnes du suffixe (/s/ ou /k/) tombent dans les radicaux de deux et trois syllabes (12). Mais ces données ont également permis à PLÉNAT de mettre au jour une régularité inédite et bien plus surprenante : la rime tombe aussi lorsque la consonne finale du radical est répétée (PLÉNAT & ROCHÉ 2003), c'est-à-dire lorsqu'elle y apparait une seconde fois comme en (13).

(12) (Louis de) Funès → funesque

(13)

	consonne répétée
a. co**l**one**l** → colonesque	/l/
b. Be**n** Lade**n** → benladesque	/n/
c. ti**r**ailleu**r** → tiraillesque	/r/
d. In**t**erne**t** → internesque	/t/

Cette étude montre qu'il est possible d'aborder la morphologie comme une science d'observation (HATHOUT *et al.* 2009a). L'augmentation du nombre des exemples pris en compte par les analyses morphologiques a un effet similaire à l'introduction du microscope dans les sciences naturelles. Quand l'observation d'une centaine de dérivés ne permet pas de dégager de généralité intéressante, celle de 3000 exemples fait apparaitre des régularités inédites qui conduisent à de nouvelles conclusions.

Ces régularités concernent notamment des configurations qui, dans la collection réduite, paraissent exceptionnelles. Mais dès lors qu'elles sont mieux représentées dans les collections étendues il devient possible d'identifier les facteurs qui expliquent leur fonctionnement et de proposer une analyse du phénomène capable de les intégrer au cas général.

3.2 La sémantique des dérivés en *-able*

Les avancées rendues possibles par l'approche extensive en morphologie ne concernent pas seulement la dimension morphophonologique. Elles peuvent également être significatives sur le plan sémantique comme l'illustre l'étude de la suffixation en *-able* de HATHOUT *et al.* (2003). Plusieurs études de cette suffixation (DUBOIS 1969 ; PLÉNAT 1988 ; LEEMAN & MELEUC 1990 ; LEEMAN 1992 ; ANSCOMBRE & LEEMAN 1994 ; FRADIN 2003) avaient été réalisées antérieurement à partir de collections dont la taille n'excède pas les 1400 adjectifs ; ce qui correspond approximativement au nombre des dérivés en *-able* des grands dictionnaires de langue comme le TLF ou le GRLF (*Grand Robert de la Langue Française*). Sur le plan sémantique, la suffixation en *-able* était analysée comme ayant un « sens passif ». Les dérivés en *-able* sont en effet principalement construits sur des verbes et sont utilisés pour modifier des noms qui, dans l'évènement dénoté par la base verbale, ont un rôle de patient (14). Dans

une étude plus ancienne, GAWELKO (1977) a identifié trois petites séries de dérivés construits sur des bases nominales, en l'occurrence des noms de taxes (corvée, taille, gabelle, etc.), de véhicules et de titres (15).

(14) réparer → réparable
'on peut réparer le téléphone'
= 'le téléphone peut être réparé'
= 'le téléphone est réparable'

(15) a. corvée → corvéable
b. cycle → cyclable
c. président → présidentiable

Les données disponibles au début des années 1990 comportaient cependant des dérivés dont l'analyse est problématique. On trouve d'une part des dérivés qui se comportent différemment des passifs de leurs verbes de base (16). Plus gênants sont les dérivés pour lesquels le nom recteur ne peut pas être analysé comme correspondant à un argument du verbe de base (17).

(16) a. Marie répare le téléphone
Le téléphone est réparable
b. Cette robe coute 100 euros
* 100 euros sont coutables
c. Un terrain atterrissable
* L'avion atterrit le terrain

(17) Une robe à un prix abordable

En 2003, HATHOUT, PLÉNAT & TANGUY ont collecté et analysé 5286 dérivés. Ce nombre est plus de trois fois supérieur à celui des exemples considérés dans les études antérieures. Les auteurs constatent tout d'abord que le sens de la grande majorité des adjectifs en *-able* peut être décrit comme « passif ». Ils observent cependant que les noms recteurs des adjectifs en *-able* peuvent représenter une grande part des participants à l'évènement dénoté par le

verbe de base. On trouve par exemple des sujets comme en (18) et des compléments indirects comme en (19).

(18) D'une manière générale, la sensibilité au gel d'une pâte de ciment est étroitement liée à la quantité d'eau "**gelable**".

(19) Le premier PC «**parlable**». On pourra maintenant causer à son ordinateur.

Ces deux exemples soulèvent une question essentielle, relative à la nature et à l'acceptabilité des données prises en compte dans les analyses (voir section 2.1). Les dérivés *gelable* en (18) et *parlable* en (19) sont-ils des lexèmes du français ? Sont-ils des lexèmes acceptables ? Peut-on (ou doit-on) les utiliser comme des instances de dérivés en *-able* ? La réponse de HATHOUT *et al.* (2003) est clairement affirmative. Ils considèrent en effet qu'en l'absence d'indices clairs permettant d'affirmer qu'un énoncé *n'*a *pas* été produit par un locuteur ayant une bonne maitrise du français, et si cet énoncé ne comporte ni erreur ni dysfluence, alors il doit être intégré à la collection des exemples à analyser. On ne peut en effet en aucun cas se limiter au français fortement normé que l'on trouve dans les grands dictionnaires de langue. Le rôle du linguiste est d'expliquer le fonctionnement de la langue parlée par les locuteurs et non de celle qui est idéalisée par les instances de normalisation institutionnelles. La lecture des documents dont sont extraits (18) et (19) montrent clairement que les locuteurs qui les ont produits ont une maitrise parfaite de la langue, qu'ils connaissent les normes comme le montre l'emploi des guillemets, qu'ils ont considéré que la création et l'utilisation de ces dérivés est légitime et qu'elles ne posent pas de problème de compréhension ou d'interprétation aux lecteurs. Dans ces conditions, rien ne justifie l'exclusion de ces exemples.

La question de l'acceptabilité des exemples collectés est permanente. Elle se pose pour chaque dérivé présent dans les

corpus ou sur le Web et doit faire l'objet d'une réponse au cas par cas dont l'une des conséquences est le cout prohibitif de la constitution des collections d'exemples. Signalons que la collecte intensive d'exemples a elle-même une influence sur les jugements d'acceptabilité dans la mesure où il est parfois difficile d'estimer la qualité d'un dérivé non-standard isolé, mais son interprétation peut devenir plus facile lorsqu'il est rapproché d'autres mots similaires. Remarquons par ailleurs que les exemples illustrant les emplois des dérivés ne sont pas édités : les fautes d'orthographe et de typographie sont conservées.

La diversité des participants à l'évènement dénoté par un verbe de base que son dérivé en *-able* permet de modifier peut être illustrée par les exemples suivants de l'adjectif *pêchable* (20-29). La plupart proviennent de blogs ou de groupes de discussion. Certains sont plus marqués que d'autres.

(20) Poisson Avec ce concept révolutionnaire, enfin les gros **poissons** sont **pêchables** au coup !

(21) Taille des poissons La sur-pêche et le non respect de la **taille pêchable** en Guadeloupe a entraîné une forte régression de la population.

(22) Lieu de pêche […] 3 km de **rives pêchables**, bien aménagées pour le lancer […]

(23) Longueur du lieu de pêche La **longueur pêchable** sur les 2 berges est de 2 025 mètres.

(24) Étendue d'eau La **rivière** reste **pêchable** en été , […]

(25) Jour 31 Aout Eau très haute (9,7 m3/s) et froide (9°C), premier **jour pêchable** depuis le 15 Aout. Quelques gobages, surtout des petits poissons, …

(26) Saison de pêche C'est vrai, la carte de pêche complète à 75€, rapportée aux nombres de **jours pêchables**, et même si ça augmente chaque année, ce n'est pas hors de prix. …

(27) Vent Jusqu'à 14 ça va, au delà je sors pas car le **vent** devient trop gênant voir **impêchable**. …

(28) Conditions météorologiques Si le vent monte trop et que les **conditions** ne deviennent plus **pêchables,** plusieurs solutions s'offrent à vous :. - tout plier et attendre une accalmie ...

(29) Fil de pêche je remarque aprés quelques lancers (je peche generalement a 40 metres en etang) que **mon nylon** se met a vriller et devient **impechable**. ...

La plasticité sémantique de ces adjectifs n'était pas signalées dans les études de la suffixation en *-able* publiées avant 2003. Elle permet notamment de modifier les objets, les lieux où se trouvent les proies et ceux où se postent les pêcheurs, les instruments, etc. mais aussi les propriétés de ces participants, notamment leur dimension ou leur force (pour le vent). On peut rattacher à ce dernier cas et plus précisément à (21), l'exemple en (16) où le prix est une propriété de l'un des participants, en l'occurrence la robe. Dans le cas de *pêchable*, il semble en effet que tout participant à l'évènement dénoté par le verbe *pêcher* ainsi que toutes ses propriétés puissent être modifiés par l'adjectif, à l'exception de l'agent, le pêcheur. Cette conclusion peut être reformulée comme suit :

> X peut être dit *pêchable* si
> - X a une propriété qui favorise l'évènement dénoté par le verbe *pêcher* ;
> - X intervient dans l'évènement mais ne peut pas être l'agent.

Dans HATHOUT (2009), j'ai proposé une analyse plus générale en termes de dynamique des forces (TALMY 2000) qui ne nécessite pas d'exclure explicitement les agents :

> X peut être dit *pêchable* si X est susceptible d'exercer une force antagoniste qui s'oppose à la réussite de l'évènement dénoté par le verbe *pêcher*, mais qui n'est pas suffisante pour l'empêcher.

Par exemple, une berge est *pêchable* si les éventuelles difficultés liées à son accès et à son utilisation n'empêchent pas que l'on puisse y pêcher du poisson avec succès. Notons que cette analyse peut aussi rendre compte de certains dérivés désadjectivaux comme *seulable* : l'auteure a du succès auprès des garçons ; ce succès exerce une force qui s'oppose à la réalisation (l'avenance) de l'état de solitude ; cette force n'est pas suffisante.

Ces avancées dans la description du sens des adjectifs en *-able* dépendent directement de la taille de la collection d'exemples réunis pour cette étude. Les analyses qui en découlent permettent de réintégrer des dérivés jusque-là considérés comme exceptionnels comme *abordable* et d'expliquer que *coutable* ne soit pas attesté : *couter* décrit une propriété qui n'implique pas de succès ni d'échec.

L'étude de HATHOUT *et al.* (2003) a aussi permis de compléter celle de GAWELKO (1977) en mettant au jour quatre nouvelles séries d'adjectifs dénominaux dont les bases dénotent des noms de construction (30), de lieu (31), de peine (32) et de finalité (33).

(30) Terrain 1200m M2 arboré et **piscinable**.
(31) L'objet **muséable** est à votre image [...]
(32) le simple fait de prier un dieu, ou même de prêcher le Juge, était un fait grave, **peinable de mort**.
(33) L'évolution du prix de la commission pour les Bintje **fritables** est présentée à la figure 4.

Par ailleurs, HATHOUT *et al.* (2003) ont montré que les dérivés dénominaux font partie des mêmes séries que les dérivés déverbaux et que leur création s'explique d'abord par l'absence de verbe permettant de dénoter l'évènement évoqué par le nom de base.

4. Les hauts et les bas dans la collecte de données extensives pour la morphologie

L'utilisation de données réelles en grande quantité se généralise. En témoignent les collections réunies dans le cadre des thèses en morphologie soutenues ces dernières années comme celles de TRIBOUT (2010) ou KOEHL (2012b). Les données utilisées dans KOEHL (2012b) proviennent en grande partie du Web, qui tend à devenir la source quasi-universelle de tous les exemples utilisés dans les recherches morphologiques en synchronie. Je rappelle qu'en réalité, nous n'avons accès qu'à une petite partie des pages Web : seules les occurrences de mots qui figurent dans les pages indexées par les moteurs de recherches peuvent être retrouvées. Idéalement, les morphologues pourraient constituer leurs collections directement à partir de ces index qui normalement contiennent l'ensemble des mots qui apparaissent dans les pages référencées. Le résultat serait quasi-parfait : le nombre des exemples maximal, la précision très élevée, et la collection peu biaisée car ne reflétant pas les intuitions des linguistes qui les ont constituées, etc. Cet idéal est malheureusement totalement inaccessible car la qualité des moteurs de recherche dépend crucialement de celle de leurs index ; leur valeur est colossale et leur protection maximale.

4.1 Interrogation automatisée des moteurs de recherche

La protection des index n'a pas toujours été aussi forte qu'elle l'est aujourd'hui et les possibilités d'interrogation offertes par les premiers moteurs étaient plus variées que celles que nous connaissons actuellement. Au début des années 2000, les morphologues avaient la possibilité de soumettre aux moteurs de recherche comme AltaVista ou Yahoo! des quantités importantes de requêtes automatiques. Ils pouvaient ainsi récupérer en peu de temps des nombres élevés de candidats dérivés. Le moteur AltaVista acceptait en

outre des requêtes par patron dans lesquelles une partie des mots pouvait être remplacée par un joker. Par exemple, une requête `pro*able` permettaient de récupérer des pages contenant des mots comme (34).

(34) probabilisable, professorable, promenable, promotable, promouvable, prononciable, protégeable

Ces requêtes permettaient en théorie de récupérer dans l'index du moteur tous les mots susceptibles d'avoir été construits par une règle spécifique de construction de lexèmes. Leur intérêt le plus important résidait dans le fait qu'elles fournissaient des mots que le morphologue n'avait pas prédits, parce qu'ils n'entraient pas dans sa conception du phénomène. Un tel apport pouvait se révéler décisif pour l'analyse.

La possibilité de soumettre des requêtes automatiques aux moteurs de recherche a été exploitée par plusieurs outils comme Webaffix (TANGUY & HATHOUT 2002 ; HATHOUT & TANGUY 2003) ou WaliM (NAMER 2003, 2009). Webaffix est une boite à outils d'acquisition lexicale à partir du Web qui dispose de plusieurs modules. Le premier permet de construire un ensemble de requêtes incluant des jokers permettant de récupérer les mots connus du moteur et qui contiennent un affixe donné. Le second crée des requêtes en prédisant des formes possibles à partir de schémas d'affixation appris sur un lexique flexionnel comme TLFnome[4]. Le troisième permet de soumettre ces deux types de requêtes à un moteur de recherche et de réaliser diverses opérations de nettoyage des résultats en éliminant les pages qui ne contiennent pas le ou les mots recherchés, celles qui ne sont pas rédigées dans la langue souhaitée, celles où le

[4] TLFnome est un lexique flexionnel du français construit à l'INaLF/ATILF à partir de la nomenclature du Trésor de la Langue Française. Morphalou, la version XML de cette ressource est distribuée par le CNRTL à l'adresse suivante : www.cnrtl.fr/lexiques/morphalou/

mot se trouve dans une liste, etc. Webaffix a été utilisé pour constituer les collections d'exemples de plusieurs études en morphologie extensive menées à l'ERSS et a fortement contribué à la démonstration de la supériorité de l'approche extensive en morphologie.

Plusieurs outils similaires ont été développés à la même époque, notamment le méta-moteur WaliM réalisé par NAMER pour vérifier si des mots prédits sont attestés sur le Web. Les formes des mots dérivées sont générées à partir de TLFnome au moyen de GéDériF (NAMER & DAL 2000), un générateur morphologique qui implémente des règles de construction de lexèmes conçues et mises au point par des linguistes. Ces formes font ensuite l'objet de requêtes soumises au moteur Yahoo! dont les résultats sont filtrés pour ne conserver que les mots ayant au moins une attestation.

4.2 Les gros corpus de page Web

À partir de 2003, les possibilités d'interroger les moteurs de recherche au moyen de robots sont petit à petit devenues plus limitées jusqu'à disparaitre complètement. Aujourd'hui, plus aucun moteur ne les accepte. Seule l'interrogation manuelle au moyen d'un navigateur est autorisée. Ces restrictions constituent un retour en arrière dont les conséquences sur la morphologie extensive sont importantes. Les linguistes doivent se contenter de vérifier l'attestation des dérivés les plus probables qui sont généralement les moins intéressants, dans la mesure où ce sont les moins susceptibles de faire progresser les descriptions. La taille des collections d'exemples sur lesquelles ils fondent leurs études se trouve réduite de façon significative, conduisant à un moins grand nombre de généralisations et à des généralisations plus grossières. Enfin, ces collections sont plus biaisées que par le passé car elles dépendent de l'intuition de celui qui prédit les formes

dérivées, qui, involontairement tend à favoriser les formes compatibles avec sa théorie et à pénaliser celles dont elles seraient des contre-exemples.

Ces restrictions ont également conduit les morphologues à se tourner vers un « succédané » du Web, à savoir les gros corpus de pages Web comme frWaC (BARONI *et al.* 2009). Ce corpus français de 1,6 milliard de mots, librement disponible à des fins de recherche, a été constitué par BARONI et son équipe pour réaliser des études de sémantiques distributionnelles. D'autres corpus, plus gros encore, ont été compilés dans le cadre de programmes de recherche comme Quaero. Exalead, le maitre d'œuvre, a ainsi crée un corpus de plus 2,5 millions de pages contenant plus de 3,3 milliards de mots. SAJOUS, TANGUY et moi avons réalisé une étude d'acquisition morphologique sur le corpus Exalead, afin notamment de comparer les exemples collectés avec ceux que l'on pouvait obtenir en utilisant Webaffix (HATHOUT *et al.* 2009b). L'étude portait sur les suffixations déverbales en *-age*, *-ment* et *-ion* pour lesquelles nous disposons de collections réunies dans le cadre du projet WesConVa (DAL *et al.* 2004). L'un des enseignements de cette étude est qu'on trouve en moyenne un nouveau déverbal (i.e. absent du TLF) dans 2000 pages Web, soit environ 1200 dérivés dans les 2,5 millions de pages. À titre de comparaison, le lexique Verbaction, créé à partir du TLF contient 3800 déverbaux en *-age*, *-ment* et *-ion*. Le corpus Exalead permet donc d'augmenter la collection des exemples disponibles de l'ordre de 30%, là où Webaffix permettait des progressions allant de 300% à 3000% ! Malgré sa taille exceptionnelle, ce corpus est de fait tout petit. Il impose de plus au morphologue d'avoir des compétences minimales en traitement automatique des langues, du type de celles présentées dans TANGUY & HATHOUT (2007) car il ne dispose pas d'une interface d'interrogation.

Cette baisse considérable dans les capacités de découverte de nouveaux dérivés a de nombreuses conséquences : le cout de création des jeux de données augmente dans la mesure où elle nécessite une plus grande intervention des morphologues, qui doivent notamment soumettre leurs requêtes manuellement ; les jeux de données sont plus petits ; il faut plus de temps pour trouver des exemples intéressants, non-prévus par les théories et les analyses actuelles.

Je voudrais enfin signaler que la constitution d'une collection d'exemples pour une étude morphologique comporte aussi des « frais cachés » souvent assez lourds. Les attestations récupérées sur le Web sont fortement bruitées : chaque exemple doit être soigneusement examiné. Or une campagne peut ramener plusieurs dizaines de milliers de candidats dont il faut vérifier l'acceptabilité. Les faux positifs sont en effet très nombreux. Une forme peut être le résultat d'une faute de frappe, d'une faute d'orthographe, d'une erreur de découpage dues à une césure ou à l'omission d'un espace ; elle peut appartenir à une partie du discours autre que celle qui est visée, être un nom propre (par exemple, un identifiant dans un blog) ou un mot d'une autre langue ; elle peut apparaitre dans du code informatique, dans une adresse mail ou une URL ; elle peut être produite par un traducteur automatique ou par une personne qui n'a clairement pas une maitrise suffisante de la langue ; etc. À cela s'ajoutent les questions d'ambigüité, comme par exemple les noms en *-eur* en français qui peuvent être des noms d'agent (*tailleur*) et des noms de propriétés (*longueur*), même si cette seconde dérivation tend à être aujourd'hui remplacée par la suffixation en *-ité* (KOEHL 2012b). J'ajoute enfin que ce travail philologique doit être répété pour chaque nouvelle collecte.

5. La préservation des collections de données morphologiques

Les données, devenues plus couteuses à obtenir que par le passé, n'en demeurent pas moins indispensables à toute recherche en morphologie. L'histoire esquissée dans les sections précédentes est celle d'une évolution, dont le point de départ a été l'utilisation de collections limitées car difficiles à constituer. À ces époques, les philologues produisaient des ouvrages dans lesquels ils compilaient à la main le vocabulaire. Les ressources sont ensuite devenues plus accessibles, puis de plus en plus conséquentes jusqu'à l'abondance amenée par le Web et les premiers moteurs de recherche. Cette mutation a permis de problématiser la place des données réelles dans le dispositif de recherche en morphologie et d'établir définitivement (1) qu'elles sont indispensables et (2) que la qualité des analyses dépend directement de la quantité des données utilisées. Aujourd'hui la situation redevient plus complexe, et il faut inventer de nouvelles manières de travailler afin de s'adapter aux restrictions imposées par les moteurs de recherche. L'une d'elles est d'utiliser des ressources où l'on sait que la créativité morphologique est peu contrainte par les normes institutionnelles comme les tweets de la plateforme Twitter où les enfants se déclarent in-dormables, où les restaurants sont étoilables, etc. Si le travail de nettoyage et de vérification des exemples constitue une part importante dans le cout de création d'une collection d'exemples, les morphologues ont généralement peu conscience de sa valeur et rares sont ceux qui finalisent leurs jeux de données et les mettent à la disposition de la communauté. C'est à cet aspect qu'est consacrée cette dernière section : la préservation des collections qui va de pair avec une meilleure reconnaissance du travail investi dans leur constitution.

5.1 La conservation et la dissémination des données en morphologie

Dans de nombreuses sous-disciplines de la linguistique, notamment en socio- ou en psycholinguistique, le partage et la réutilisation de jeux de données existants est une pratique bien établie ; ce n'est pas encore le cas en morphologie. En psycholinguistique, par exemple, des ressources comme CELEX (pour l'anglais, l'allemand ou le néerlandais ; BAAYEN *et al.* 1995) ou Lexique (pour le français ; NEW 2006) servent à constituer du matériel expérimental pour les études sur le lexique mental et sur les traitements morphologiques. D'autres ressources comme CHILDES (MACWHINNEY 2000) sont utilisées dans de très nombreuses recherches sur l'acquisition du langage. Il n'existe, en revanche, rien de comparable pour la description morphologique. Chaque étude débute par la constitution d'une nouvelle collection d'exemples dont le point de départ est généralement un dictionnaire électronique comme le TLFi. Le recours au Wiktionnaire est plus rare car il n'existe pas pour l'instant d'outil d'interrogation adapté à la recherche morphologique[5]. Suit une pêche aux exemples sur le Web plus ou moins fastidieuse, plus ou moins fructueuse. Les dérivés sont ensuite décrits dans des tables ou une base de données dont le contenu dépend essentiellement des facteurs qui interviennent dans les analyses prévues. Il arrive souvent par ailleurs, que certaines parties de la collection soient traitées de manière plus approfondie que d'autres. Le format de ces jeux de données est également *ad hoc* et parfois hétérogène, notamment lorsqu'ils se composent de plusieurs fichiers. Ce « manque de considération » envers les données s'explique

[5] Cette situation devrait néanmoins évoluer grâce à GLÀFFOLI (http://redac.univ-tlse2.fr/glaffoli/), l'interface d'interrogation du lexique GLÀFF (SAJOUS *et al.*, 2013 ; HATHOUT *et al.*, 2014), qui permet à l'utilisateur d'extraire les entrées du Wiktionnaire en combinant différents critères catégoriels et de forme.

d'abord par le fait que ces collections sont créées spécifiquement pour une étude particulière et que leur dissémination ne fait pas partie des objectifs du morphologue qui les constitue : l'utilisation de données déjà analysées ne fait pas (encore) partie des pratiques de la communauté ; la finalisation et la dissémination des données comportent par ailleurs un surcout qui ne saurait se justifier que si ce travail était considéré comme ayant une valeur en soi, s'il pouvait être valorisé par des publications et obtenir une reconnaissance suffisante des instances d'évaluation.

Notons qu'il existe de très nombreux travaux sur les formats de données lexicales, conçus et développés essentiellement en traitement automatique des langues, comme LMF (Lexical Markup Format ; FRANCOPOULOS 2006). Ces formats se caractérisent par une grande généricité mais ils ne sont pas utilisés pour la description des données en morphologie parce que la plupart des morphologues ne les connaissent pas et qu'ils ne sont pas suffisamment adaptés aux besoins de ces derniers.

Pour amener les morphologues à partager et réutiliser plus systématiquement les collections et les analyses associées, il faudrait concevoir une ressource à large couverture, disposant d'une interface d'interrogation et d'outils de gestion intuitifs. Une telle ressource doit disposer d'une architecture qui accepte une grande variété de descriptions morphologiques. Elle doit être aussi œcuménique que possible sur le plan théorique et ne doit pas être fondée sur des hypothèses ou des présupposés sur le contenu ou la forme des analyses morphologiques. Afin de permettre l'alignement des différentes descriptions, les informations doivent être suffisamment décomposées. L'information y sera distribuée du fait de la décomposition des informations, et redondante – une même information pouvant apparaitre dans plusieurs éléments. La redondance viendra également des différences de granularité dans la

ressource d'éléments d'information qu'elle réunit. Cette ressource devra permettre que certaines informations soient manquantes ou incomplètes. Enfin, elle indiquera explicitement l'origine de chacune des informations qu'elle contient pour reconnaitre le crédit de leurs auteurs, permettre la citation de leur travaux et éventuellement la sélection ou le masquage de certaines des descriptions. L'objectif premier est donc de créer une ressource où les morphologues puissent intégrer leurs collections d'exemples, obtenir des données pour de nouvelles études et à terme l'utiliser comme un outil intégré de constitution et de stockage des jeux de données morphologiques.

5.2 Le réseau morphologique *Démonette*

Une ressource qui satisfait en partie aux spécifications présentées ci-dessus est en cours de développement dans le cadre d'une collaboration avec NAMER (HATHOUT & NAMER 2014a, 2014b). Ce nouveau réseau lexical du français, baptisé *Démonette*, se caractérise par la variété des relations morphologiques qu'il décrit, à la fois directes (entre ascendants et descendants) et indirectes (au sein de la même famille dérivationnelle), simples et complexes, et par le nombre des traits morphologiques, phonologiques, catégoriels et sémantiques dont sont munis les sommets, qui représentent les lexèmes, et les arcs, qui représentent les relations dérivationnelles. Démonette est conçu pour articuler des informations provenant de deux systèmes fondés sur des principes totalement opposés. Le premier est DériF (NAMER 2009, 2013), un analyseur morphologique dérivationnel qui implémente une vingtaine de règles de construction de lexèmes définies et mises au point par des linguistes comme la suffixation en *-age*, la préfixation en *dé-* ou la composition savante. Ce système prend en entrée des formes de citation de lexèmes (construits) munies de catégories grammaticales. Pour chaque lexème construit,

DériF calcule un lexème de base, le procédé dérivationnel utilisé pour le construire, la liste de ses antécédents dérivationnels et une glose de son sens construit (35). Les analyses sont réalisées en appliquant récursivement les règles implémentées. DériF dispose en outre de listes d'exceptions qui permettent de prendre en compte les irrégularités lexicales.

(35) enneigement/NOM ==> [[en [neige NOM] VERBE] ment NOM] (enneigement/NOM, enneiger/VERBE, neige/NOM) "(Action - résultat de l'action) de enneiger"

Démonette contient également des analyses provenant de Morphonette (HATHOUT 2011) un réseau lexical du français basé sur une conception relationnelle et paradigmatique de la morphologie. Dans ce lexique, les propriétés morphologiques sont décrites par la position des lexèmes dans le réseau, position identifiée par les paradigmes qui les contiennent. Par exemple, la position d'un dérivé comme *modifiable* est décrite par sa famille dérivationnelle qui rassemble les lexèmes *modifier*, *modification*, *modificateur*, *modificatif*, *modifiant*, *modifieur, immodifiable*, etc. et par sa série qui contient l'ensemble des dérivés en *-able* : *agaçable*, *agitable*, *chevauchable*, *définissable*, *différenciable*, *rechargeable*, *réconciliable*, *soutenable*, etc. Morphonette est composé de filaments, c'est-à-dire de triplets (m, p, $s_p(m)$) où m est une entrée, p est un membre de la famille dérivationnelle de m et $s_p(m)$ est la sous-série dérivationnelle de m relativement à p. $s_p(m)$ est l'ensemble des mots du lexique qui se trouvent dans une relation similaire à celle que m entretient avec p. En d'autres termes, un mot u appartient à $s_p(m)$ s'il existe un mot v tel que m:p=u:v (i.e. tel que m, p, u, v forment une analogie). L'exemple (36) présente le filament de l'adjectif m = *modifiable* pour p = *modificateur*. Ce filament illustre l'une des caractéristiques originales de Morphonette,

à savoir qu'il décrit à la fois des relations directes et des relations indirectes, comme ici entre deux dérivés du verbe *modifier*.

(36) (modifiable, modificateur, {amplifiable, glorifiable, identifiable, justifiable, clarifiable, mystifiable, rectifiable, sanctifiable, simplifiable, spécifiable, unifiable, vérifiable})

Les informations issues de DériF et de Morphonette ont été décomposées et intégrées dans le réseau Démonette, qui décrit les relations dérivationnelles entre des couples de mots. Ces relations sont caractérisées par cinq types de propriétés : caractéristiques des entrées (graphies, catégories et types sémantiques) ; description « topologique » de la relation (orientation et complexité) ; description constructionnelle de la relation (types et exposants des procédés dérivationnels et thèmes dérivationnels) ; gloses concrètes et abstraites ; descriptions phonologiques (transcriptions API). Ces informations sont illustrées dans le tableau 1 (page suivante) pour la relation entre AMORTIR et le nom de résultat AMORTISSEMENT (2e colonne ; cette relation est issue de DériF) et entre le nom d'agent MODIFICATEUR et le nom d'action MODIFICATION (3e colonne ; cette relation est issue de Morphonette). Dans les entrées, la 1re information est la graphie, la 2e, l'étiquette morphosyntaxique au format Grace et la 3e, le type morpho-sémantique (@ = prédicat, @AGM = agent masculin, @RES = résultat, @ACT = nom d'action).

	amortir ← amortissement	modificateur ← modification
Entrée 1	amortir/Vmn----/@	modificateur/Ncms/@AGM
Entrée 2	amortissement/Ncms/@RES	modification/Ncfs/@ACT
Relation 1 ← 2	descendant/simple	transversale/simple
Construction 1		suf/eur/modificat
Construction 2	suf/ment/amortiss	suf/ion/modificat
Glose concrète	réaliser l'action dont le résultat est un amortissement	(agent masculin OR instrument) de la modification
Glose abstraite	réaliser l'action dont le résultat est @RES	(agent masculin OR instrument) de @ACT
Phono 1	amɔʁtiʁ	mɔdifikatœʁ
Phono 2	amɔʁtisəmɑ̃	mɔdifikasjɔ̃

Tableau 1 – Descriptions des relations morphologiques dans Démonette

Dans le tableau 1 (page précédente), les relations décrivent l'entrée 1 relativement à l'entrée 2. On observe que Démonette contient des relations descendantes qui, comme dans la 2e colonne connectent une base à l'un de ses dérivés, mais aussi des relations transversales entre des mots qui appartiennent à la même famille dérivationnelle. Les autres particularités sont la présence de descriptions morphosémantiques (types), la redondance des informations (chaque dérivation donne lieu à deux relations ; un lexème a autant de descriptions qu'il y a de relations dans lesquelles il apparait), la conception cumulative du sens (chaque mot a autant de gloses sémantiques concrètes et abstraites qu'il a de relations dérivationnelles) et l'indication de l'origine de chacune des informations (omises dans le tableau 1)[6]. Par ailleurs, certaines informations peuvent ne pas être renseignées si elles ne sont pas fournies par la ressource originale. Grâce à son architecture « ouverte » Démonette peut être complétée par une variété de ressources morphologiques dérivationnelles. La liste des traits morphologiques, phonologiques, catégoriels et sémantiques peut être étendue pour inclure celles d'une nouvelle ressource que l'on souhaiterait y ajouter. L'enrichissement de Démonette se fait au moyen des programmes de transfert spécifiques à chaque ressource. Démonette est distribuée sous une licence du domaine publique classique (Creative Commons).

Nous envisageons dans un proche avenir de construire une interface d'interrogation et d'extraction de collections

[6] Les graphies, les étiquettes morphosyntaxiques et les transcriptions phonologiques sont celles de la forme de citation du lexème car c'est son identifiant standard. Dans une version ultérieure, Démonette sera couplée à un lexique flexionnel et phonologique similaire à GLÀFF qui liste l'ensemble des graphies/formes phonologiques / étiquettes morpho-syntaxiques de chaque lexème.

d'exemples adaptée aux pratiques des morphologues. À terme, Démonette sera intégrée dans une application Web plus ambitieuse permettant de réaliser une grande partie des descriptions morphologiques : collecte des exemples, stockage, nettoyage, annotation et analyse.

6. Conclusion

Cette courte histoire de la morphologie extensive présente les évolutions de cette nouvelle approche, des questions qu'elle pose et des perspectives qu'elle ouvre. La question principale est assurément celle de la nature des données que la morphologie doit décrire. Quelle est la place des constructions spontanées en français non-normé, qui aujourd'hui constituent la plus grande partie des productions écrites : la multitude des communications personnelles comme les SMS, tweets, messages instantanés sur Facebook, posts sur des blogs, etc. dépassent de très loin la quantité des textes professionnels, commerciaux, journalistiques, etc. rédigés dans un français plus conforme aux normes institutionnelles. L'irruption de ces nouvelles formes de communication et de la langue qui y est utilisée est généralement totalement ignorée par les linguistes et notamment les morphologues, du fait de leur formation académique mais aussi parce qu'ils sont souvent totalement démunis face à la créativité de ces nouveaux locuteurs-rédacteurs. La seconde évolution qu'apporte la morphologie extensive concerne la place des données et l'importance de la constitution de collections étendues dans le travail des morphologues. Ces derniers doivent apprendre à manipuler et exploiter des données en grand nombre et acquérir de nouvelles méthodes de travail, plus expérimentales et probablement, à terme, plus quantitatives. Ces évolutions auront également une conséquence sur la nature des études qui deviendront plus techniques, plus longues, plus couteuses et qui imposeront de travailler en équipe.

La morphologie extensive n'est naturellement pas une évolution isolée en linguistique. La place des données change aussi dans d'autres sous-disciplines avec un recours plus important aux corpus annotés au niveau syntaxique comme le Penn TreeBank (MARCUS *et al.* 1993) ou le French TreeBank (ABEILLÉ *et al.* 2003) ou discursif comme le Penn Discourse TreeBank (PRASAD *et al.* 2008) ou le corpus Annodis (PÉRY-WOODLEY *et al.* 2009). Ces sources de données riches ont fait l'objet de nombreuses publications et peuvent sur certains aspects nous servir d'exemples pour une meilleure rentabilité de la création de collections d'exemples en morphologie. Une autre direction dans laquelle une solution serait la bienvenue est la création d'outils de moissonnage lexical du Web capables de créer des collections d'attestations génériques. La mise en place de dispositifs de ce type ne peut cependant être réalisée qu'avec un soutien important des agences nationales et européennes de financement de la recherche.

Références

ABEILLÉ Anne, CLÉMENT Lionel & TOUSSENEL François. (2003). Building a Treebank for French. In *Treebanks*, 165-187. Dordrecht: Kluwer.

ANSCOMBRE Jean-Claude & LEEMAN Danielle. (1994). La Dérivation des Adjectifs en *-ble* : Morphologie ou Sémantique ? *Langue Française 103*, 32-44.

BAAYEN R. Harald, PIEPENBROCK Richard & GULIKERS Leon. (1995). *The CELEX lexical database (release 2). CD-ROM*. Philadelphia: Linguistic Data Consortium.

BARONI Marco, BERNARDINI Silvia, FERRARESI Adriano & ZANCHETTA Eros. (2009). The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43-3*, 209-226.

BAUER Laurie. (2014). Grammaticality, acceptability, possible words and large corpora. *Morphology 24-2*, 83-103.

CHURCH Kenneth W. & MERCER Robert L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational linguistics 19-1*, 1-24.

DAL Georgette, LIGNON Stéphanie, NAMER Fiammetta & TANGUY Ludovic. (2004). Toile contre Dictionnaires : Analyse Morphologique en Corpus de Noms Déverbaux Concurrents. In *Colloque international sur les noms déverbaux*, Villeneuve-d'Ascq.

DUBOIS Jean. (1969). *Grammaire Structurale du Français : La Phrase et les Transformations*. Paris : Larousse.

FRADIN Bernard. (2003). *Nouvelles Approches en Morphologie*. Paris : Presses Universitaires de France.

FRANCOPOULO Gil, MONTE George, CALZOLARI Nicoletta, MONACHINI Monica, BEL Nuria, PET Mandy & SORIA Claudia. (2006). Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC (2006), Genova, Italy.

GAWELKO Marek. (1977). *Évolution des Suffixes Adjectivaux en Français*. Wrocław : Zakład Narodowy im. Ossolińskich.

GREFENSTETTE Gregory. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. In *Proceedings of the 21st ASLIB Conference on Translating and the Computer*.

HATHOUT Nabil. (2009). *Contributions à la Description de la Structure Morphologique du Lexique et à l'Approche Extensive en Morphologie*. Toulouse : Université Toulouse le Mirail - Toulouse II. Habilitation à diriger des recherches.

HATHOUT Nabil. (2011). Morphonette: a Paradigm-Based Morphological Network. *Lingue e Linguaggio 10-2*, 245-264.

HATHOUT Nabil, PLÉNAT Marc & TANGUY Ludovic. (2003). Enquête sur les Dérivés en *-able*. *Cahiers de Grammaire 28*, 49-90.

HATHOUT Nabil & TANGUY Ludovic. (2003). Webaffix : Une Boite à Outils d'Acquisition Lexicale à Partir du Web. *Revue québécoise de linguistique 32-1*, 61-84.

HATHOUT Nabil, MONTERMINI Fabio & TANGUY Ludovic. (2008). Extensive Data for Morphology: Using the World Wide Web. *Journal of French Language Studies 18-1*, 67-85.

HATHOUT Nabil, NAMER Fiammetta, PLÉNAT Marc & TANGUY Ludovic. (2009a). La Collecte et l'Utilisation des Données en Morphologie. In FRADIN Bernard, KERLEROUX Françoise & PLÉNAT Marc (Eds), *Aperçus de Morphologie du Français*. Saint-Denis : Presses Universitaires de Vincennes, 267-287.

HATHOUT Nabil, SAJOUS Franck & TANGUY Ludovic. (2009b). Looking for French Deverbal Nouns in an Evolving Web (a Short History of WAC). In *Proceedings of Web as Corpus (2009) (WAC5)*, San Sebastián.

HATHOUT Nabil & NAMER Fiammetta. (2014a). Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology 11-5*, 125-168.

HATHOUT Nabil & NAMER Fiammetta. (2014b). La Base Lexicale Démonette : Entre Sémantique Constructionnelle et Morphologie Dérivationnelle. In *Actes de la 21^{e} Conférence sur le Traitement Automatique des Langues Naturelles TALN (2014)*, Marseille, 208-219.

HATHOUT Nabil, SAJOUS Franck & CALDERONE Basilio. (2014). GLÀFF, a large versatile French lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA).

KILGARRIFF Adam & GREFENSTETTE Gregory. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational linguistics 29-3*, 333-347.

KOEHL Aurore. (2012a). *Altitude*, *négritude*, *bravitude* ou la Résurgence d'une Suffixation. In *Actes du 3e Congrès Mondial de Linguistique Française (CMLF 2012), vol. 1*, 1307-1323.

KOEHL Aurore. (2012b). *La Construction Morphologique des Noms Désadjectivaux Suffixés en Français*. Nancy : Université de Lorraine Thèse de doctorat.

KOEHL Aurore & LIGNON Stéphanie. (2014). Property Nouns with *-ité* and *-itude* : Formal Alternation and Morphopragmatics or the sad-itude of the Aité$_{N}$. *Morphology 24-4*. 351-376.

LEEMAN Danielle. (1992). Deux Classes d'Adjectifs en *-ble*. *Langue Française 96*, 44-64.

LEEMAN Danielle & MELEUC Serge. (1990). Verbes en tables et Adjectifs en *-able*. *Langue Française 87*, 30-51.

MACWHINNEY Brian. (2000). *The CHILDES project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum.

MARCUS Mitchell P., MARCINKIEWICZ Mary Ann & SANTORINI Beatrice. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics 19-2*, 313-330.

NAMER Fiammetta. (2003). WaliM : Valider les Unités Morphologiques Complexes par le Web. In *Les unités morphologiques. Actes du 3e Forum de morphologie (Silexicales 3)*. Villeneuve d'Ascq : Presses Universitaires de Lille, 142-150.

NAMER Fiammetta. (2009). *Morphologie, Lexique et Traitement Automatique des Langues : L'Analyseur DériF*. Paris : Hermès Science-Lavoisier.

NAMER Fiammetta. (2013). A Rule-Based Morphosemantic Analyzer for French for a Fine-Grained Semantic Annotation of Texts. In MAHLOW Cerstin & PIOTROWSKI Michael (Eds), *SFCM 2013 CCIS 380*. Heidelberg : Springer, 93-115

NAMER Fiammetta & DAL Georgette. (2000). GÉDÉRIF: Automatic Generation and Analysis of Morphologically Constructed Lexical Resources. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens: ELRA.

NEW Boris. (2006). Lexique 3 : Une Nouvelle Base de Données Lexicales. In *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*, Louvain-la-Neuve : Presses Universitaires de Louvain.

PÉRY-WOODLEY Marie-Paule. (1995). Quels Corpus pour quels Traitements Automatiques ? *Traitement Automatique Des Langues 36-1/2*, 213-232.

PÉRY-WOODLEY Marie-Paule, ASHER Nicholas, ENJALBERT Patrice, BENAMARA Farah, BRAS Myriam, FABRE Cécile, FERRARI Stéphane, HO-DAC Lydia-Mai, LE DRAOULEC Anne, MATHET Yann, MULLER Philippe, PRÉVOT Laurent, REBEYROLLE Josette, TANGUY Ludovic, VERGEZ-COURET Marianne, VIEU Laure & WIDLÖCHER ANTOINE. (2009). ANNODIS : une Approche Outillée de l'Annotation de Structures Discursives. In *Actes de la 16e conférence sur le*

traitement automatique des langues naturelles (TALN-2009), Senlis, France.

PLÉNAT Marc. (1988). Morphologie des Adjectifs en *-able*. *Cahiers de Grammaire 13*, 101-132.

PLÉNAT Marc. (2000). Quelques Thèmes de Recherche Actuels en Morphophonologie Française. *Cahiers de Lexicologie 77*, 27-62.

PLÉNAT Marc. (2011). Enquête sur Divers Effets des Contraintes Dissimilatives en Français. In ROCHÉ Michel, BOYÉ Gilles, HATHOUT Nabil, LIGNON Stéphanie & PLÉNAT Marc (Eds), *Des Unités Morphologiques au Lexique*, Paris : Hermès, 145-190.

PLÉNAT Marc, TANGUY Ludovic, LIGNON Stéphanie & SERNA Nicole. (2002). La Conjecture de Pichon. *Corpus 1*, 105-150.

PLÉNAT Marc & ROCHÉ Michel. (2003). Prosodic Constraints on Suffixation in French. In BOOIJ Geert E., DECESARIS Janet, RALLI Angela & SCALISE Sergio (Eds), *Topics in Morphology. Selected Papers from the third Mediterranean Morphology Meeting*. Barcelone : Universitat Pompeu Fabra, 285-299.

PRASAD Rashmi, DINESH Nikhil, LEE Alan, MILTSAKAKI Eleni, ROBALDO Livio, JOSHI Aravind K. & WEBBER Bonnie L. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the sixth international conference on Language Resources and Evaluation* (LREC 2008), Marrakech.

SAJOUS Franck, HATHOUT Nabil & CALDERONE Basilio. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN-2013)*. Les Sables d'Olonne, France, 285-298.

TALMY Leonard. (2000). *Toward a Cognitive Semantics*. Cambridge: MIT Press.

TANGUY Ludovic. (2012). *Complexification des Données et des Techniques en Linguistique : Contributions du TAL aux Solutions et aux Problèmes*. Toulouse : Université de Toulouse 2 - Le Mirail, Habilitation à diriger des recherches.

TANGUY Ludovic. (2013). La Ruée Linguistique vers le Web. Texto! *Textes et Cultures 18-4*.

TANGUY Ludovic & HATHOUT Nabil. (2002). Webaffix : Un Outil d'Acquisition Morphologique Dérivationnelle à partir du Web. In

PIERREL Jean-Marie (Ed.), *Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, Nancy : ATALA, 245-254.

TANGUY Ludovic & HATHOUT Nabil. (2007). *Perl pour les Linguistes. Programmes en Perl pour Exploiter les Données Langagières*. Paris : Hermès Science-Lavoisier.

TRIBOUT Delphine. (2010). *Les Conversions de Nom à Verbe et de Verbe à Nom en Français*. Université Paris 7. Thèse de doctorat.