REMARQUES SUR L'USAGE DES CORPUS EN MORPHOLOGIE

Bernard Fradin, UMR 7110 LLF, CNRS & U de Paris 7
Georgette Dal, U Lille 3, UMR 8163 STL, CNRS & U de Lille 3 et Lille 1
Natalia Grabar, U de Paris 5, U729 INSERM, SPIM; Foundation Health on the Net
Fiammetta Namer, UMR 7118 ATILF, CNRS & Nancy-Université
Stéphanie Lignon, UMR 7118 ATILF, CNRS & Nancy-Université
Delphine Tribout, UMR 7110 LLF, CNRS & U de Paris 7
Pierre Zweigenbaum, UPR 3251 LIMSI CNRS & INALCO Paris

0. Résumé

L'objet de cet article est triple. Il s'agira tout d'abord de rappeler en quoi l'accès à des ressources numérisées importantes a changé qualitativement la donne en morphologie. Il s'agira ensuite de montrer que ces ressources nécessitent un gros travail de mise en forme avant de pouvoir être exploitées et que les procédures de mise en forme demandent à être consignées avec soin car elles conditionnent les résultats des interrogations portant sur les données triées. Il s'agira enfin de constater que l'utilisation en masse de ressources numérisées n'abolit pas le recours au jugement des locuteurs, qui demeure indispensable quand il faut déterminer l'acceptabilité des formes, mais qu'elle le met en perspective par le jeu de l'attesté.

1. Introduction

La possibilité de recourir à de grandes masses de données numérisées et de développer des outils d'interrogation et de mise en forme de ces données a changé la perception de certains phénomènes linguistiques et, partant, leur description. Elle a également permis d'envisager des recherches impossibles à mener sinon, comme, par exemple, l'élaboration d'une grammaire de la variation terminologique (Jacquemin 1999). Ce changement est particulièrement net en sémantique (Condamines 2005b) et, plus encore peut-être, en morphologie, dans la mesure où celle-ci entretient une relation privilégiée avec le lexique.

Cette arrivée massive de données numérisées détermine, en retour, de la part des linguistes des stratégies visant à rechercher l'apport optimal que peuvent offrir ces données. Ceux-ci se trouvent souvent placés devant une alternative identique à celle qui s'offre aux créateurs de corpus :

— explorer un ensemble de documents aussi grand que possible. L'hypothèse implicite est que plus nombreuses sont les données, meilleures elles sont (« *more data is better data* » (Habert, Nazarenko & Salem 1997 : 146)), car on peut observer plus de phénomènes linguistiques et chaque phénomène de manière plus complète ;

— étudier des corpus échantillonnés autour d'un ou plusieurs critères. L'hypothèse implicite est que l'échantillonnage permet une appréhension plus précise du phénomène linguistique étudié.

Ces deux solutions sont difficilement conciliables, mais elles ont chacune leur raison d'être.

Nous nous limiterons ici aux travaux menés dans le champ de la morphologie constructionnelle, c'est-à-dire celui de la construction de lexèmes. Il y aurait aussi des études intéressantes à mener dans le domaine de la morphologie flexionnelle, par exemple la recherche sur la Toile de formes qui s'écartent des formes reçues (*vivèrent* / *vécurent*) ou bien encore la mise en évidence des stratégies d'évitement utilisées par les locuteurs : *faisait frire*, *était en train de frire* au lieu de *friait*, forme défective de FRIRE (Apothéloz & Boyé (sous presse)).

Afin de bien situer le débat, il nous semble nécessaire de proposer au préalable une typologie des ressources numérisées qui seront utilisées de manière différenciée en morphologie. La première opposition concerne les objets que l'on collecte comme données. Il peut s'agir de formes hors contexte, lemmatisées ou non[1]. Il peut s'agir de formes en contexte, c'est-à-dire de formes dans des textes. A ce niveau, une nouvelle distinction s'impose. Soit ces textes satisfont, ou visent à satisfaire, la définition du corpus comme « collection de données langagières qui sont sélectionnées et organisées explicitement selon des critères linguistiques *et extra-linguistiques* explicites pour servir d'échantillon d'*emplois déterminés d'une langue* » (cf. Habert (2000 : 13) qui complète la définition de Sinclair (1996)). Soit ce n'est pas le cas et on a affaire à des rassemblements arbitraires de données, dont la Toile est le prototype. Cette distinction repose sur trois ordres d'opposition. Par définition, les corpus visent une représentativité quand ils sont de vrais corpus et non ce qu'Anne Condamines nomme des « bases textuelles » (cf. Condamines 2005a : 18) : ils doivent être un échantillon de la langue[2]. Ils sont construits pour être une image fidèle de la langue. En revanche, la représentativité de la Toile n'est pas assurée, parce que cette dernière ne résulte pas d'une élaboration. La deuxième opposition met en jeu l'idée de stabilité. Alors qu'un corpus a des critères de clôture, qui déterminent ce qui y sera et n'y sera pas, la Toile n'a aucun critère conceptuel de clôture : s'y trouve tout ce qui peut être représenté sous forme numérisée. La dernière opposition fait jouer la notion de cumul. Il y a cumul des données dans un corpus : les données qui y sont y demeurent au fur et à mesure qu'il croît, parce qu'il a une croissance monotone. Ceci n'est pas forcément vrai pour la Toile, dont les données sont volatiles et peuvent disparaître d'une période d'interrogation à l'autre. On prendra ici le parti de classer sous les corpus les regroupements « opportunistes » (Marcus, Santorini & Marcinkiewicz 1993) de documents, comme les archives de certains périodiques tels *Le Monde*, *Le Monde diplomatique*, *Le Soir, Le Courrier International*, etc. Même si ces documents ne sont pas des corpus au sens strict, ils visent quand même à une représentativité de la langue et on peut mener à partir d'eux des études sur des rubriques présentant une unité thématique (Gaeta & Ricca 2003).

Dans les faits, les études de morphologie constructionnelle recourent à tous les types distingués ci-dessus, sauf peut-être à des corpus au sens le plus contraint du terme.

2. L'apport des ressources numérisées en morphologie

Nous avons identifié cinq raisons justifiant le recours à des données numérisées en morphologie constructionnelle.

[1] Le terme de *lemme* employé ici appartient au métalangage informatique. Il désigne la forme qui subsume toutes les variations de formes des mots du point de vue du traitement informatique. Les formes lemmatisées correspondent aux formes citationnelles qui se trouvent dans les dictionnaires. Par exemple toutes les formes d'un verbe ont pour lemme son infinitif, ou encore CE sert de lemme à *ce*, *ces*, *c'*, *cet*, *cette*.

[2] Ou d'un sous-langage au sens harrissien.

2.1. La première raison que nous invoquerons est qu'aussi bien les corpus au sens strict que les groupements opportunistes de documents ou la Toile permettent d'effectuer des études destinées à valider des hypothèses théoriques ou, le cas échéant, à les infirmer. Ces études peuvent être centrées sur la forme des lexèmes construits ou sur leur sens. Elles peuvent également mettre en jeu les contextes d'utilisation des formes de lexèmes. Sans le recours à ces diverses ressources numérisées, les hypothèses demeurent en effet souvent infalsifiables. Or, on connaît l'importance de la falsifiabilité dans le domaine des sciences. Nous donnerons quelques exemples d'hypothèses théoriques qu'ont permis de valider les ressources numérisées au §4.1.

2.2. La deuxième raison n'est pas propre à la morphologie, mais vaut également pour elle. Comme le notent Habert & Zweigenbaum (2002 : 88), « la validité de l'intuition d'acceptabilité ou inversement l'impression que tel énoncé est "impossible" » à la base des travaux en grammaire générative, et plus largement en linguistique, est mise à mal par l'avènement des très gros volumes de ressources textuelles sous forme numérisée. La morphologie n'y échappe pas. En effet, dans le domaine de la morphologie, jusqu'à l'avènement de ces ressources, les jugements d'acceptabilité qui étaient formulés pour tel ou tel lexème construit absent des principaux dictionnaires de langue générale ou les étapes intermédiaires reconstituées dans certaines analyses se fondaient la plupart du temps sur l'introspection du descripteur, cette dernière pouvant être biaisée par le résultat souhaité. C'est ainsi qu'une requête sur la Toile avec DÉSIMMORTALISABILISER[3], qui ne ramène aucune page, rend suspect ce verbe, que Corbin (1987 : 790) pose pourtant comme possible[4] (nous reparlerons des verbes en *-ABILISER* dans le §4.1.2.). Sont pour la même raison suspectes les deux étapes BUANDE, catégorisée comme nom, et BUANDIER, catégorisée comme adjectif, dont font l'hypothèse Corbin & Corbin (1991) au cours d'une analyse très serrée sur la suffixation par *-IER* du français, pour expliquer le nom BUANDIER ('personne qui lave le linge') à partir du verbe BUER ('faire la lessive')[5]. Inversement, ANXIEUSITÉ, marqué d'un astérisque dans Corbin (1987), compte 22 occurrences sur la Toile (requête effectuée le 12 novembre 2006), ce qui indique que l'allomorphie attendue (ANXIOSITÉ) n'est pas toujours réalisée par les scripteurs internautes. De la même façon, Corbin (1987 : 356) prédit que les verbes en *-ISER* correspondant à COSMIQUE et HUMORISTIQUE sont, respectivement, COSMICISER (vs *COSMISER) et HUMORISTICISER (vs *HUMORISTISER). Or, les résultats obtenus sur la Toile donnent dans les deux cas la préférence au verbe marqué comme impossible[6]. Ceci dit, il est difficile de tirer une conclusion ferme de ces résultats dans la mesure où l'on n'a aucun indice sur la cohérence de la compétence grammaticale (au sens technique) des scripteurs qui les ont produits.

[3] Suivant la convention initiée par Matthews (1974), on notera les lexèmes en petites capitales, leur formes fléchies en minuscules italiques. Les affixes seront notés en petites capitales italiques.

[4] Requête effectuée le 1/11/2006 avec les formes *désimmortalisabiliser*, *désimmortalisabilise*, *désimmortalisabilisent*, *désimmortalisabilisera*, *désimmortalisabiliseront*, *a désimmortalisabilisé*. Nous mentionnons des exemples de D. Corbin car, étant morphologue, elle a fait des hypothèses explicites sur le lexique potentiel. Ceci dit, le décalage entre les formes postulées comme (non) existantes par les linguistes et ce qu'offrent les données numérisées se posent aussi en syntaxe. Par exemple Dubois (1969 : 70) postule que (i) *Pierre a un chien* provient de (ii) *Pierre fait qu'un chien est de lui*. Or, comme on s'y attend, les moteurs de recherche ne fournissent aucune attestation pour (iii) *un chien est de lui*.

[5] La Toile fournit des occurrences de *buande*, mais il s'agit de coupes malencontreuses de *buanderie*. Quant à *buandier*, il est présent sur la Toile, mais en tant que nom.

[6] COSMISER et HUMORISTISER ramènent respectivement 23 et 2 pages, contre 4 et 0 pour COSMICISER HUMORISTICISER (requête effectuée le 12/11/06 avec les formes *Xiser*, *Xise*, *Xisent*).

2.3. L'utilisation de données massives, en particulier de la Toile, permet également de faire jaillir des régularités, impossibles à observer sur des échantillons de langue plus petits (dont les dictionnaires, sous quelque forme que ce soit, ou même les cueillettes effectuées au hasard des lectures). Plusieurs des travaux menés par Marc Plénat, seul ou en collaboration, font apparaître que la consultabilité de gros volumes textuels sur support électronique permet de mettre en évidence des phénomènes qui resteraient sinon inaperçus (cf. par exemple Plénat (1997) sur la suffixation en -*ESQUE*, Plénat (2002) sur la suffixation en -*ISSIME*, Hathout, Plénat & Tanguy (2003) sur la suffixation en -*ABLE*[7]). Nous détaillerons pour notre part l'exemple des adjectifs en -*IEN* dans le §4.2.

2.4. Une quatrième utilisation possible des données numérisées en morphologie réside dans les recherches sur la productivité morphologique. En effet, comme on le verra au §4.3., les mesures de productivité que l'on peut effectuer pour telle ou telle règle de construction de lexèmes sont par définition consubstantielles aux corpus dans lesquelles elles ont été effectuées.

2.5. Enfin, de façon plus inattendue peut-être, les données numérisées permettent également d'effectuer des études sur les manques et de donner du sens aux lacunes observables, à condition toutefois que l'on s'entoure d'un certain nombre de précautions que nous détaillerons au §4.5.

Avant de donner quelques exemples d'études illustrant ce qui précède, nous ferons quelques précisions sur l'apprêt que doivent subir les corpus avant qu'on puisse les utiliser.

3. Apprêt des données numérisées

Les données langagières livrées sur les supports commerciaux (CD-ROM) ou telles qu'on peut les récupérer sur la Toile ne peuvent pas être utilisées telles quelles parce qu'elles contiennent des informations qui en parasiteraient l'étude. Elles nécessitent d'être apprêtées et ce sont les procédures d'apprêt que nous allons décrire maintenant. Ces procédures sont au nombre de quatre : nettoyage, sélection, étiquetage et lemmatisation. Les deux premières doivent être menées quelle que soit l'étude de morphologie constructionnelle envisagée. Les troisième et quatrième dépendent, dans les faits, de l'objectif de l'étude : ainsi, les études en morpho-phonologie peuvent se dispenser de l'étiquetage, celles portant sur la productivité se situent en amont de la lemmatisation. Ces procédures peuvent en partie être exécutées automatiquement. Leur validation passe toutefois par une étape manuelle.

3.1. Nettoyage

Pour trier les données langagières brutes que nous livre l'interrogation de corpus, une première série de tris est nécessaire qui utilise des critères formels exclusivement. Les formes qui ne satisfont pas les critères formels sont de plusieurs types :

1) Les coquilles. Les corpus textuels, notamment journalistiques, peuvent contenir beaucoup de coquilles. Sur un ensemble de 177 formes se terminant par -*et* et de 62 se terminant par -*ette*, Fradin, Hathout & Meunier (2003) aboutissent à 86 séquences

[7] Ces derniers prônent le bien-fondé de ce qu'ils appellent une « approche extensive » de la suffixation en -*ABLE*.

inanalysables comme lexèmes en *-et* et 38 comme lexèmes en *-ette*. Ces coquilles sont majoritairement des fautes de frappe (1a) ou de découpe (1b) (la forme attendue est entre parenthèses) :

(1) a bugdget (budget), nettoiemenet (nettoiement), prpojet (projet), épouvette (éprouvette), avainet (avaient), gagdet (gadget), vollet (volley),
b accueilet (accueil et), sadette (sa dette), le travail d'Aimé J acquet (Jacquet), quecette (que cette), nitrompette (ni trompette)

Selon les cas, les coquilles seront simplement corrigées, corrigées mais en conservant la trace de leur correction, ou rejetées (cf. §3.2).

2) Des acronymes ou noms de marques. Suivant les procédés morphologiques étudiés, ils sont plus ou moins nombreux. Pour *-ET*, on relevait 36 acronymes relevant du domaine de l'informatique *(telnet, calvanet)*.

3) Les noms propres quand ils apparaissent en minuscules. Ici encore, leur nombre varie avec le procédé étudié : *malhuret, tourmalet, pinochet, villette, brossolette*… Pour les noms de marque devenus noms communs *(mobylette, sanisette)*, il faut fixer une ligne à suivre en fonction de l'étude.

4) Les formes appartenant à d'autres langues, qui ne sont pas des emprunts : *sunset*, *wilayet*, *znaet*, *svet*.

5) Les formes relevant d'une autre catégorie lexicale que la catégorie étudiée. Par exemple, dans une étude sur *-ET*, on ne veut pas inclure les formes verbales en *-ette (projette, refeuillette)* ; ou encore, on ne veut pas inclure les noms comme *inconfort* dans une étude sur les adjectifs en *IN-*. Ces problèmes de catégorisation peuvent être redoutables au point d'empêcher l'étude d'un phénomène sur corpus[8]. Pour y remédier, on a recours alors souvent à un niveau de préparation automatique des corpus qui consiste à attribuer la catégorie grammaticale la plus probable à chaque mot-forme, en fonction de son contexte phrastique. Cette tâche dévolue aux étiqueteurs est brièvement décrite en §3.3, ainsi que les inconvénients qui peuvent s'y rattacher.

Au terme de ce premier tri, on aboutit aux données apprêtées. L'écart entre les données brutes et les données apprêtées est plus ou moins important. Il varie en fonction de la nature du marquage phonique / graphémique associé au procédé. Dans le cas de la suffixation en *-ET*, il est très important. Ainsi, le corpus choisi — cinq années du journal *Libération*, de 1995 à 1999 inclus totalisant 87 millions d'occurrences — contenait 617 000 unités dont la finale est *-et*, *-ets*, *-ette* ou *-ettes*, chiffre qui inclut la conjonction *et*. Or, même en admettant que *et* augmente artificiellement le nombre de formes, le nombre total de noms en *-ET* conservé (= données apprêtées) ne s'élève qu'à 270 441, ce qui reste très élevé pour une analyse manuelle.

3.2. Sélection

Les unités qui satisfont les critères formels ne relèvent pas forcément toutes du procédé étudié. Pour le déterminer, il faut faire entrer en ligne de compte les critères morphologiques, qui permettent d'éliminer :

[8] C'est le cas en italien où l'étude sur corpus de la suffixation diminutive en *-ETT-* (*-etto/a/i/e)* butte sur l'existence de très nombreux participes passés (et adjectifs) avec la même terminaison. De ce fait, le repérage des formes en *-etto/a/i/e* ne dispense pas le linguiste d'un immense travail de nettoyage qui rend ce type d'étude très coûteuse en temps.

1. Les lexèmes comportant une suite graphique accidentellement identique à l'affixe étudié. Il s'agit la plupart du temps de lexèmes non construits[9], comme SQUELETTE, BLETTE et AMPHET *(<* AMPHÉTAMINES*)* pour la suffixation en *-ET*, FAIBLE*,* DOUBLE et AFFABLE pour la suffixation en *-ABLE*.

2. Les lexèmes qui présentent bien l'affixe étudié, mais pour lesquels cet affixe ne résulte pas de la dernière opération constructionnelle. Par exemple pour la suffixation en *-ABLE*, INFAISABLE et INFRÉQUENTABLE, s'ils présentent bien le suffixe *-ABLE*, sont construits en dernier lieu par la préfixation en *IN-* (bases FAISABLE et FRÉQUENTABLE) et non par la suffixation en *-ABLE* (cf. les bases impossibles *INFAIRE et *INFRÉQUENTER). Ou encore pour la suffixation en *-ET* : *(se retrouver sur la) cybersellette*, *minicamionnette*.

Dans certains cas, plusieurs histoires constructionnelles sont possibles, comme pour IMMOBILISABLE qui peut en dernier lieu être construit par la suffixation en *-ABLE* sur le verbe IMMOBILISER ou bien par la préfixation en *IN-* sur l'adjectif MOBILISABLE. Ici encore, le choix à faire dépend de l'étude envisagée.

3. Les lexèmes inanalysables comme construits en français contemporain. Il s'agit d'emprunts au latin et au grec comme IMPECCABLE, INÉLUCTABLE, ALACRITÉ ou à d'autres langues (BAGUETTE*,* BUDGET*,* FARIGOULETTE*,* QUINTETTE). Mais il peut s'agir aussi de formes construites à un état antérieur de la langue et devenues opaques comme HOULETTE (< afr HOULER 'lancer, jeter') ou *galet* (< afr GAL 'caillou'). Il peut s'agir enfin de formes qui n'ont jamais été construites (ASSIETTE*,* DISETTE*,* ÉCHAUGUETTE).

Le cas 3 pose parfois des problèmes difficiles à résoudre. Deux types de facteurs entrent en ligne de compte pour décider qu'un lexème est construit : les connaissances linguistiques du locuteur d'une part ; les critères qu'on sélectionne pour l'analyse, d'autre part.

Pour ce qui est du locuteur, selon qu'on se place du point de vue d'un locuteur ordinaire n'ayant jamais fait de latin ni de grec ou bien du point de vue d'une personne ayant cette connaissance, les lexèmes INDÉLÉBILE ou INEXTINGUIBLE ne seront probablement pas considérés de la même façon.

Quant à l'analysabilité, on peut distinguer trois degrés : (i) reconnaissance immédiate de l'affixe et de la base, ou d'un de leurs allomorphes, et capacité à les voir comme combinés par une règle de construction de lexèmes : ce cas correspond à l'analysabilité classique telle qu'elle est définie, par exemple, dans Chomsky & Miller (1963) ; (ii) reconnaissance conjointe de l'affixe étudié et d'une base du français (ou d'un de leurs allomorphes) mais la combinaison des deux ne permet pas d'obtenir le sens de l'unité en question e.g. BAGU-ETTE ; (iii) reconnaissance de l'affixe ou de la base mais pas des deux e.g. IMPAVIDE ou JASSERIE. Le choix de l'analyse dépend dans une large mesure des hypothèses qu'on admet à propos des connaissances du locuteur. Voir §4.3.

Le traitement des lexèmes de ce type nécessite de fixer les critères qui déterminent ce qui est analysable, au regard de l'étude menée, préalablement à toute validation. Celle-ci peut dépendre des objectifs de l'étude. Fradin, Hathout et Meunier (2003) par exemple considèrent comme analysables en français les lexèmes qui, d'une part, sont segmentables en une partie suffixale *-et* ou *-ette* et une partie radicale appartenant au lexique (e.g. CALCUL-ETTE), d'autre part, relèvent d'une règle de construction de lexèmes disponible (la règle qui forme des N d'instruments en *-ET*). Selon un autre point de vue,

[9] Un lexème construit est un lexème complexe qui peut être engendré / analysé par les règles de la morphologie qui construit les unités ayant vocation à devenir des unités lexicales e.g. SUFFIXAT-ION. Un lexème est complexe sans être construit s'il peut seulement être analysé partiellement e.g. ROY-AUME (Corbin 1987).

Grabar et al. (2006b) considèrent comme construits selon un procédé donné les lexèmes contenant l'affixe exposant du procédé et présentant un élément de sens attaché au procédé en question.

Quand, dans un corpus donné, il s'agit de sélectionner automatiquement les données pertinentes pour l'analyse, on peut identifier trois classes de variantes orthographiques non reçues, en fonction de leur impact sur les résultats du traitement en morphologie :

(1) les écueils qui empêchent l'identification du procédé constructionnel analysé. Par exemple, la graphie *nettoiemenet* masque la présence d'un suffixé en -*MENT,* et *portabe* ne peut être décompté comme un dérivé suffixé en -*ABLE* ;
(2) les séquences qui, à l'inverse, vont introduitre du bruit dans le calcul des lexèmes construits par un procédé donné. Par exemple, à l'issue d'une sélection automatique des données, éventuellement lemmatisées, *vollet* (au lieu de *volley*) est fautivement comptabilisé parmi les dérivés en -*ET* ;
(3) les variantes, enfin, qui n'ont aucune incidence sur l'analyse morphologique d'un procédé particulier ; ainsi *conjuguable* (au lieu de *conjugable*) ou *communiquable* (vs. *communicable*) sont régulièrement analysés comme construits par la suffixation en -*ABLE*.

Toute variante non reçue, quelle que soit la classe à laquelle elle appartient, va cependant altérer les résultats des traitements lexicométriques (par exemple en augmentant artificiellement le nombre d'hapax). Face à chaque variante posant problème, et en fonction des objectifs qu'il se fixe, l'utilisateur peut choisir entre deux stratégies : corriger les variantes considérées comme non reçues, ou les mettre à l'écart. La deuxième est la plus efficace d'un point de vue informatique ; en revanche, les résultats obtenus sont biaisés : en éliminant des formes, on risque a priori de se priver de cas intéressants, et on crée ainsi du silence. La solution symétrique est celle de la correction systématique des variantes. Cette stratégie a le mérite de garantir des résultats entièrement fiables. Par contre, elle est aussi coûteuse que la précédente est économique, puisque la correction doit le plus souvent être manuelle.

Quelle que soit la stratégie adoptée, elle se fonde sur une typologie minimale des variantes graphémiques qu'on peut rencontrer (voir notamment Catach 1995). Le tableau 1 en propose une, à titre indicatif, qui articule les deux dimensions du code graphémique et de la prononciation visée par le scripteur. Le code graphémique présente lui-même deux niveaux : le premier, qu'on qualifiera de graphotactique[10], énonce les combinaisons graphémiques possibles pour la langue en question ; le second, qui relève de la norme, au sens de Coseriu (1967), stipule quelles sont les formes graphémiques recevables dans la langue en question (des variantes étant possibles). Le code graphémique visant explicitement à transcrire la prononciation des locuteurs (à un moment donné de l'histoire et de la société), la question de savoir si une forme graphémique correspond à la prononciation de l'unité linguistique que le locuteur / scripteur avait en tête se pose et constitue le troisième paramètre à prendre en compte.

[10] Ce terme est fabriqué sur le modèle de *phonotactique* (partie de la phonologie qui stipule quelles sont les combinaisons de phonèmes possibles pour les unités lexicales d'une langue). *Graphémotactique* aurait été plus exact, mais probablement moins heureux.

Graphotactique			
+			–
Correspond à la prononciation			
+		–	
Orthographe reçue			
+	–		
(a)	(b)	(c)	(d)

Tableau 1. Typologie des variantes graphémiques

(a) mots bien orthographiés : *erreurs, propose, avaient*.
(b) la forme est graphémiquement possible et correspond à une prononciation possible, mais son orthographe n'est pas celle reçue pour le mot en question, laquelle est donnée entre parenthèses : *camionette (camionnette), moufette (mouffette), conjuguable (conjugable), vollet (volley).*
(c) la forme est possible dans le code graphémique du français mais sa prononciation ne correspond pas (selon la norme) au mot visé en question (en fait, c'est souvent un non mot) : *nettoiemenet (nettoiement), avainet (avaient), perpet (perpète).*
(d) la forme enfreint le code graphotactique du français : *prpojet (projet), bouqet (bouquet), brossser (brosser), flèmme (flemme)*.

Les cas (b) induisent du bruit (*vollet* / *volley*) ou sont neutres (*conjuguable* / *conjugable*). En effet, alors qu'on peut considérer ce dernier comme un dérivé déverbal en -*ABLE*, on est fondé à rejeter *vollet* et *nettoiemenet* comme étant des dérivés en -*ET*, parce qu'ils n'ont aucune base, possible ou attestée. Dans le cas de *nettoiemenet*, toutefois, cela augmente le silence. Le cas (d) regroupe ce qu'on classe comme des fautes de frappe.

En résumé, on retiendra qu'il faut expliciter les critères formels qu'on utilise et garder la trace de ce qui est éliminé. Ceci est plus vrai encore pour les critères de sélection. Cette démarche offre l'avantage de permettre de comparer les ensembles de formes obtenues et tester les analyses sur chacun d'eux. Ceci peut être particulièrement utile quand on a des questions liées à l'opacité. De cette manière, au lieu de rester dans l'ombre, les critères de tri deviennent des paramètres contrôlables intervenant dans le traitement.

Le tri de certains de ces cas peut être facilité par l'application d'un analyseur morphologique, soulageant ainsi l'intervention humaine. Souvent, un tel outil n'est applicable que sur des données étiquetées et lemmatisées. La chaîne de traitement que constitue cette succession de tâches réalisées automatiquement est présentée ci-dessous, à l'exclusion de l'analyse morphologique. Les lexèmes validés à la suite de ces étapes constituent le matériel de travail du morphologue.

En dehors des cas qui viennent d'être mentionnés, on fait l'hypothèse que toutes les formes relevées en corpus ou sur la Toile sont considérées comme valides. Toutefois le fait qu'une forme soit attestée ne suffit pas à en faire une donnée recevable : il faut qu'elle soit reproductible par d'autres locuteurs.

3.3. Etiquetage

L'étiquetage catégoriel est la procédure qui consiste en l'annotation automatique des formes au moyen de leur catégorie lexicale la plus probable. L'attribution de cette catégorie est le fruit du croisement de diverses techniques, chacune exploitée à des degrés divers en fonction de l'étiqueteur considéré : règles de grammaires locales

ordonnées[11], apprentissage statistique, consultation d'un dictionnaire. La catégorie calculée peut être, suivant les étiqueteurs, complétée par des informations morpho-syntaxiques caractérisant la forme. L'étiquetage est une étape clé dans l'apprêt des corpus, dont seules peuvent se passer les études de morphophonologie. La valeur des résultats des recherches sur corpus est tributaire de la justesse des catégories produites par l'étiqueteur ; par conséquent, le choix de celui-ci est crucial.

TreeTagger (Schmid 1994) est un étiqueteur publiquement disponible à des fins de recherche[12], qui peut traiter plusieurs langues (dont français, anglais, allemand, italien). Les résultats qu'il produit en font l'un des outils « libres » les plus fiables. Son fonctionnement est probabiliste : il se fonde sur l'utilisation d'arbres de décision et se sert d'un dictionnaire de petite taille. Le système comprend également un module de segmentation. Lors de l'étiquetage, le lemme et certaines informations flexionnelles (temps pour les verbes, type du déterminant, etc.) sont calculés. Les résultats de l'étiquetage sont affichés sous forme de triplets réunissant la forme fléchie, la catégorie, le lemme. Les conventions d'étiquetage catégoriel utilisées par TreeTagger illustrées par les exemples du tableau 2 sont : déterminant de type article (DET:ART), pronom de type possessif (PRO:POS), adverbe (ADV) et nom propre (NAM).

Forme Fléchie	Etiquette	Lemme	Forme Fléchie	Etiquette	Lemme
Les	DET:ART	Le	-c'	ADV	<unknown>
hyperliens	PRO:POS	<unknown>	Héliopolis	NAM	<unknown>

Tableau 2. Sortie de TreeTagger

Comme en témoigne le tableau 2, la valeur arbitraire <unknown> est attribuée en guise de lemme aux unités inconnues du dictionnaire de l'étiqueteur (*hyperliens, -c', Héliopolis*) ; la présence de ces mots inconnus (*hyperliens*) ainsi que les mauvaises segmentations (*-c'*) peuvent entraîner la production d'erreurs d'étiquetage. Les erreurs les plus fréquentes de TreeTagger concernent les noms propres (*Héliopolis*), rarement reconnus comme tels.

3.4. Lemmatisation

L'identification du lemme correspondant à une forme catégorisée n'est pas toujours une étape d'apprêt indispensable : en particulier, on peut en faire l'économie dans le cadre de calculs de productivité (cf. §4.3). En tout état de cause, le lemmatiseur doit permettre de conserver la valeur de la forme d'entrée, de manière à ce que le corpus apprêté soit exploitable dans toutes sortes d'études.

Le lemmatiseur du français Flemm (Namer 2000) répond à cette exigence et permet également de corriger certaines des erreurs d'étiquetage produites par TreeTagger. Flemm est un programme conçu pour lemmatiser les formes catégorisées, après avoir contrôlé, voire rectifié, les étiquetages et segmentations erronés. Ce programme est en outre capable d'effectuer l'analyse flexionnelle des mots inconnus. Le lemmatiseur Flemm se sert de règles pour produire le lemme de ces mots inconnus ainsi que l'ensemble des traits flexionnels calculables hors contexte. Le dictionnaire de Flemm est réduit à quelques milliers d'exceptions. Le fait que ce programme soit basé sur

[11] Par exemple, quand elle précède une forme préalablement reconnue comme nom, la forme *la* est de catégorie 'déterminant'; et quand elle précède ce qui a été identifié comme un verbe conjugué, il s'agit d'un 'pronom'.

[12] URL : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html. Les résultats de TreeTagger sont généralement très bons, et le temps d'exécution très rapide.

règles le rend capable de lemmatiser les mots dits inconnus, et de gérer les analyses multiples. Le résultat produit est de type (2) :

(2) Forme Fléchie/Catégorie:Traits Flexionnels/Lemme(:Famille)

La vérification des résultats de l'étiquetage, qui précède la lemmatisation proprement dite, se déroule en deux temps :

1. Décollement des ponctuations en début et fin de mot, et s'il y a reconnaissance de la séquence graphique obtenue, affectation de la catégorie grammaticale idoine ;

2. Evaluation de la validité de la catégorie affectée par l'étiqueteur en fonction de la terminaison du mot en présence : en cas d'incompatibilité, la catégorie choisie par Flemm est à la fois la plus vraisemblable du point de vue de la graphie du mot, et la plus proche fonctionnellement de la catégorie rejetée.

En reprenant l'échantillon du tableau 2, la vérification s'applique aux entrées *-c'* et *hyperliens* :

— une fois la séquence *c'* isolée du tiret, elle est reconnue comme pronom (PRO) ;

— la classe fermée des pronoms possessifs (PRO:POS) n'incluant pas *hyperliens*, ce dernier est recatégorisé nom commun (NOM).

Le résultat de l'application de Flemm sur l'échantillon en question est donné dans le tableau 3, qui montre la lemmatisation de mots inconnus (*Héliopolis*), éventuellement après réétiquetage (*hyperliens, c'*), et l'ajout de traits flexionnels (le nombre (**p**)luriel sur les entrées *Les* et *hyperliens*).

Forme Fléchie	Etiquette	Lemme	Forme Fléchie	Etiquette	Lemme
Les	DET(ART)**:p**	le	-c'	**PRO**	**ce**
hyperliens	**NOM:p**	**hyperlien**	Héliopolis	NAM	**Héliopolis**

Tableau 3. Sortie de Flemm

4. Quelques exemples d'utilisation des ressources numérisées en morphologie

Dans cette section, nous donnerons quelques exemples d'utilisation effective de ressources numérisées en morphologie. Nous verrons successivement des cas où ces ressources ont permis de valider des hypothèses théoriques (§4.1), de mettre en évidence des paramètres intervenant dans le mode de sélection d'une règle de construction de lexèmes (§4.2), d'effectuer des études sur la productivité morphologique (§4.3). Le §4.4. relatera des études sur les manques.

4.1. Validation d'hypothèses théoriques

Nous présentons ici succinctement trois études développées ailleurs. Les deux premières (§4.1.1) se servent de la Toile pour valider des hypothèses, sémantique et formelle. La troisième (§4.1.2) repose crucialement sur les contextes d'utilisation effective des lexèmes construits.

4.1.1. Verbes en *-ABILISER* et noms en *-AISITÉ*

Dal & Namer (2000) et Dal & Namer (2005) s'intéressent aux règles de construction de lexèmes (désormais RCL) formant respectivement des verbes désadjectivaux en *-ISER* et des noms de propriété en *-ITÉ* mettant en jeu un toponyme. Chacune de ces études part

d'une hypothèse théorique ou d'une observation préalables portant sur les bases refusées par chacune de ces RCL :
— la règle de suffixation en -*ISER* construit difficilement des verbes décrivant un changement d'état à partir de bases adjectivales suffixées par -*ABLE* (*LAVABILISER). Cette incompatibilité découle de la caractérisation sémantique proposée dans ces études des suffixations en -*ABLE* et en -*ISER* : les dérivés en -*ABLE* expriment le caractère potentiel de la propriété décrite par leur verbe-base, alors que les verbes en -*ISER* présentent la propriété décrite par leur base adjectivale comme constituant l'état effectivement atteint par le référent du complément direct du verbe construit à l'issue de l'effectuation du procès (ÉGALISER) ;
— les adjectifs ethniques en -*AIS* ou -*OIS* sont faiblement compatibles avec la RCL formant des noms de propriété en -*ITÉ*. Dans les dictionnaires, la donnée cruciale est le nom de propriété attesté FRANCITÉ correspondant à l'adjectif FRANÇAIS (*vs* FRANÇAISITÉ), sachant qu'une requête sur la Toile ne fournit que quelques occurrences de FRANÇAISITÉ.

Dans les deux cas, le protocole expérimental était identique : un programme engendrait automatiquement une liste de lexèmes construits, supposés impossibles (verbes en -*ABILISER* dans un cas, noms en -*AISITÉ*, -*OISITÉ* dans l'autre). Le robot WaliM (Namer 2003b) recherchait automatiquement via Yahoo™ les pages contenant ces formes de lexèmes. L'absence de réponse a été quasi-systématique, ce qui a permis de conforter les hypothèses formulées au préalable :
— s'agissant des verbes en -*ABILISER*, seule une poignée d'entre eux, attestés de longue date d'après le *TLF,* se retrouvent sur la Toile (COMPTABILISER, CULPABILISER, FIABILISER, NAVIGABILISER, (IM)PERMÉABILISER, RENTABILISER, STABILISER). Sur les 1 287 autres, créés pour les besoins de la démonstration (ABOLISSABILISER, ABORDABILISER…), seuls 6 ont obtenu des résultats positifs (COMMUTABILISER, NOTABILISER, PORTABILISER, POTABILISER, SOCIABILISER, VARIABILISER). On remarque que le résultat est constant, quels que soient le nombre des requêtes et la période à laquelle elles ont été effectuées ;
— s'agissant des noms en -*AISITÉ* / -*OISITÉ*, créés, eux aussi, pour les besoins de l'étude (ALBANAISITÉ, ANTILLAISITÉ, FINLANDAISITÉ, HONGROISITÉ, PORTUGAISITÉ, QUÉBECOISITÉ, CHINOISITÉ, FRANÇAISITÉ, IRLANDAISITÉ, ISLANDAISITÉ, etc.), ils renvoient toujours un nombre quasi-nul de pages indexées, quel que soit le moteur de recherche interrogé, ce qui confirme la tentative de généralisation que nous avons faite à partir de la seule observation de FRANCITÉ. Deux stratégies de repli sont observées : la base du nom de propriété est soit un radical supplétif de l'adjectif en -*AIS* (-*OIS*) (cf. MAGYARITÉ, ANTILLANITÉ), soit le toponyme servant de base à l'adjectif ethnique (ALBANITÉ, FINLANDITÉ, PORTUGALITÉ, QUÉBECITE…). Ce dernier constat mériterait à son tour d'être exploité, en ceci qu'il semble indiquer que, du point de vue de la construction de lexèmes, un toponyme et l'adjectif construit qui lui correspond semblent interchangeables, sans préjudice sémantique pour le résultat. Sans corpus, l'étude n'aurait pas pu être menée.

4.1.2. Adjectifs en *INXABLE*

Dal et al. ((à par.)) porte sur les adjectifs en *INXABLE* du français. Il fait l'hypothèse que, quand elle s'applique à des adjectifs en -*ABLE*, la préfixation en *IN*- construit des

adjectifs exprimant la non-satisfaction d'une propriété attendue[13]. Si elle est juste, cette hypothèse s'assortit de la double prédiction suivante :

(I) Pour un nom recteur donné, la situation normale est que soit l'adjectif en *-XABLE*, soit le construit en *INXABLE* fassent défaut.

(II) Si tous deux coexistent en corpus : (a) soit ils ne sont pas utilisés avec les mêmes noms recteurs, (b) soit, s'ils partagent le même nom recteur, le référent de ce dernier possède la propriété qu'exprime l'adjectif en *-XABLE* d'une façon non standard pour la catégorie d'objets nommés par lui.

Le recours à un corpus opportuniste comme *Le Monde* doublé de recherches sur la Toile a permis de tester la validité de cette double prédiction, et, par ricochet, de l'hypothèse initiale. En effet, les différents cas de figure imaginés a priori s'observent tous massivement dans *Le Monde* et sur la Toile :

— ainsi, conformément à la prédiction (I) (que reformule à sa façon la prédiction IIa), seul l'adjectif IMPUBLIABLE est attesté dans *Le Monde* 1997 avec un nom recteur désignant des types d'écrits comme ROMAN, NOTES ou INÉDITS. Cette observation s'explique par le fait que la publiabilité fait partie des propriétés attendues de ces types d'écrits, si bien qu'un principe de pertinence s'oppose à ce que l'on parle de roman publiable, sans plus de précision ;

— IMPRATICABLE / PRATICABLE, en revanche, partagent la plupart de leurs noms recteurs. Ces derniers réfèrent soit à des voies de communication (ROUTE, PISTE, RUE, etc.), soit à ce que nous avons appelé des dispositifs mentaux ou physiques (IDÉE, PROCÉDÉ, PROCÉDURE, PROPOSITION, etc.). L'emploi, récurrent, de PRATICABLE avec des noms qui se satisfont également de l'adjectif IMPRATICABLE demande donc explication. Or, un examen précis du contexte d'utilisation de PRATICABLE dans les années 1997 et 1999 du *Monde* a montré que, conformément à la prédiction (IIb), soit il n'exprime pas une propriété attendue à un degré standard pour la catégorie d'objets dont il prédique une propriété (par ex. « La piste est sinueuse, parfois difficilement praticable »), soit les entités qui possèdent cette propriété attendue sont les seules à avoir conservé cette propriété (par ex. « Les embouteillages monstrueux sur les seules routes praticables… »), si bien que, dans de tels contextes, il devient pertinent de mentionner la possession de cette propriété ;

— le couple INJOIGNABLE / JOIGNABLE confirme la prédication (IIb). Dans les années 1997 et 1999 du *Monde*, ces deux adjectifs expriment exclusivement des propriétés d'individus. Or, la quasi-totalité des contextes d'utilisation d'*INJOIGNABLE* exprime la perte de la propriété 'pouvoir être joint' (par téléphone) des individus dont cette propriété est prédiquée (par ex. « des amis devenus injoignables »), alors que JOIGNABLE figure systématiquement en cooccurrence avec « à tout moment » (« à tout instant »), exprimant, du même coup, une propriété au-delà de ce qui est attendu. Par un effet de corpus, nous n'avons trouvé aucun DIFFICILEMENT JOIGNABLE dans les deux années du *Monde* explorées, mais la Toile nous en a fourni plusieurs (par ex. « Je serai en vacances du 17/07 au 06/08, et serai difficilement joignable durant cette période »).

Avec cette étude, l'utilité des corpus opportunistes et de la Toile comme lieux de vérification d'hypothèses s'est de nouveau trouvée confirmée.

4.2. Mise en évidence de paramètres intervenant dans la sélection d'une RCL

[13] Les raisons pour lesquelles l'hypothèse est restreinte aux bases en *-ABLE* tiennent au fait qu'en synchronie, la préfixation en *IN-* sélectionne de façon quasi-exclusive ce type morphologique de bases, comme le confirme une requête menée sur les hapax legomena des années 1995 et 1999 du *Monde*.

Pour illustrer l'utilité des ressources numérisées dans la mise en évidence de paramètres entrant dans la formation de RCL, nous résumerons ici des résultats décrits dans (Lignon 2000) portant sur la suffixation en -*IEN* du français.

Le fait de disposer de séries suffisamment grandes de formes construites par une RCL est crucial en morphologie. La perception d'un phénomène n'est en effet pas la même selon qu'on dispose d'une seule occurrence d'une forme construite sur un corpus de 1 000 lexèmes, ou de six occurrences sur un corpus de 3 000. Même si, en termes de pourcentage, les deux situations restent proches, on peut difficilement parler d'exception dans le second cas, alors qu'on pourrait être tenté de le faire pour le premier. C'est ce qui se passe pour les troncations de plus d'un phonème déclenchées par la suffixation en -*IEN*. En effet, cette dernière peut avoir pour effet de tronquer les lexèmes auxquels elle s'applique. Si l'on se fonde sur les données que fournissent les dictionnaires, la troncation porte alors sur le dernier phonème de la base (par ex. *Canada* > [kanadjɛ̃] (vs [kanadajɛ̃]), cette troncation pouvant résulter de l'application d'une haplologie (par ex. *Italie* > [italjɛ̃] (vs [italijɛ̃])). Les dictionnaires ne fournissent qu'un cas où plus d'un phonème semble tronqué : il s'agit de SOGDIANE > SOGDIEN[14]. Il est difficile de tirer une quelconque conclusion à partir de ce seul cas, sauf à remarquer que la base est un dissyllabe dont la dernière syllabe comporte un yod, et de corréler cette dernière remarque au fait que -*IEN* commence également par un yod. Pour évaluer dans quelle mesure la troncation de plus d'un phonème est liée à la présence d'un yod dans la base, une étude a été menée à partir de la version numérisée du journal *Le Soir*, enrichie d'occurrences d'adjectifs en -*IEN* glanées dans San Antonio. Ces deux méthodes conjuguées ont permis d'aboutir au tableau 4.

Concaténation (20)	Troncation d'un phonème (6)	Troncation de plus d'un phonème (12)
Bases de 1 ou 2 syllabes (12) *Alcyon > alcyonien* *Audiard > audiardien* *Berlioz > berliozien* *Berrias > berriasien* *concierge > conciergien* *Fabius > fabiusien* *gruyère > gruyérien* *Guillaume > guillaumien* *Hayek > hayekien* *Ignace > ignacien* *Mauriac > mauriacien* *Vian > vianien*	Bases de plus de 2 syllabes (6) *Antonioni > antonionien* *Bolletieri > bolletierien* *Culioli > culiolien* *Dostoïevski > dostoïevskien* *maffiosi > maffiosien* *Mariano > marianien*	Bases de 2 syllabes (2) *Phidias > phidien* *Sogdiane > sogdien*
Bases de plus de 2 syllabes (7) *Condillac > condillacien* *Fouratier > fouratiérien* *Gabriel > gabriélien* *Le Corbusier > corbusierien* *Moebius > moebiusien* *Prokofiev > prokofiévien* *Robespierre > robespierrien*		Bases de plus de 2 syllabes (10) *Bérurier > bérurien* *Chateaubriand > chateaubrien* *gougnafier > gougnafien* *La Salpêtrière > salpétrien* *Microsianie > microsien* *Milésius > milésien* *palétuvier > palétuvien* *Pépé Moustier > pépémoustien* *postérieur > postérien* *Stradivarius > stradivarien*

[14] Selon le *TLF* : **SOGDIEN**, -IENNE, adj. et subst. masc.**I.** *Adj., HIST.* Propre à l'ancienne Sogdiane (région de l'ancienne Perse entre Boukhara et Samarkand), à ses habitants, à sa culture, à sa langue. **II.** *Subst. masc.***A.** *HIST.* Habitant de l'ancienne Sogdiane. **B.** *LING.* Langue appartenant au groupe oriental de la famille iranienne et à la période du moyen-iranien.

Epenthèse (1) Base de 2 syllabes (1) *Badiou* > *badioulien*

Tableau 4. Suffixation en *-IEN* avec base contenant un yod

Ce tableau permet de dégager les résultats suivants :
— la présence d'un yod dans la syllabe pénultième ou antépénultième de la base ne s'accompagne pas toujours d'une troncation, et l'on retrouve ici les trois modes d'adjonction du suffixe *-IEN* que l'on trouve pour des bases sans yod, par ordre de fréquence : concaténation, troncation, épenthèse ;
— dans les cas de troncation d'un phonème ou plus, les bases mono ou dissyllabiques sont minoritaires (2/18), alors qu'elles sont majoritaires pour les cas de concaténation et d'épenthèse (13/20).

On en conclut que, tout en étant sensible aux contraintes dissimilatives selon lesquelles les locuteurs français évitent la consécution de deux phonèmes identiques ou quasi-identiques de part et d'autre d'une frontière constructionnelle (colonne 3), la suffixation en *-IEN* satisfait également des contraintes de taille, si bien que les locuteurs évitent de tronquer des bases de moins de trois syllabes, même si elles comportent un yod (colonne 1). Les cas qui échappent à cette description (bases de une ou deux syllabes tronquées (2/15) ou bases de plus de deux syllabes concaténées (7/23)) sont explicables à l'aide d'autres paramètres que nous ne développerons pas ici.

Sans une augmentation significative de la taille de la base de données en *-IEN*, il aurait été impossible d'identifier les paramètres de taille et d'euphonie intervenant dans le choix de l'adjonction par troncation de plus d'un phonème de la base, ni de dégager les résultats que résume le tableau 4.

4.3. Etudes sur la productivité morphologique

Un autre type de recherche où l'utilisation des corpus est capitale concerne les mesures de la productivité morphologique[15].

Depuis Schultink (1961), d'un point de vue qualitatif, on s'accorde à définir la productivité morphologique comme l'aptitude d'une RCL à produire de nouveaux lexèmes de façon non intentionnelle. Plusieurs méthodes de calcul ont été proposées pour tenter d'objectiver la productivité d'une RCL au-delà de la propre intuition du descripteur, du hasard des rencontres et/ou des attestations dans les dictionnaires. Les plus usitées actuellement ont été proposées par H. Baayen, avec d'éventuels aménagements (Gaeta & Ricca 2003). Il s'agit de la « productivité au sens strict » (Baayen & Lieber 1991 : 817) et la « productivité globale ». La première, notée *P*, exprime la propension d'un procédé à former de nouveaux lexèmes (Baayen 1993 : 181). Elle est destinée à comparer la productivité de procédés à l'intérieur d'un même corpus (Baayen & Renouf 1996), ou entre corpus différents (Baayen 1994). La seconde, notée *P**, exprime la probabilité, pour un hapax legomenon construit, qu'il relève d'une RCL donnée.

On ne détaillera pas ici ces deux mesures. On retiendra toutefois qu'elles sont consubstantielles de la notion de corpus textuels, dans la mesure où elles ne valent que et rien que pour les corpus sur lesquels elles ont été effectuées.

[15] Pour un point sur la notion de productivité morphologique, (cf. Dal 2003).

Des études de productivité sur corpus ont ainsi été menées dans différentes langues et ont permis de vérifier des hypothèses et d'observer certaines spécificités des procédés morphologiques étudiés. Par exemple, Brunet (1981) a effectué une étude sur un corpus diachronique du *Trésor de la langue française* et a pu observer, entre autres, que la suffixation en -*IQUE* semble être propre au domaine technique. En effet, ce suffixe montre des fréquences remarquables au début du XIX^{e} siècle, période qui correspond à un développement industriel et une production de documentation technique intenses.

Dans des travaux plus récents, qui font suite aux travaux de H. Baayen cités plus haut, l'étude de la productivité est effectuée à partir de corpus synchroniques. Keune, Van Hout & Baayen (2006) observent ainsi, sur un corpus oral du néerlandais, qu'il existe une relation entre le locuteur et l'emploi des procédés morphologiques, et donc de leur productivité. Les critères qui entrent en jeu sont relatifs à l'âge, au sexe, à l'éducation et à la provenance géographique des locuteurs. Les auteurs de (Grabar et al. 2006a) ont effectué une étude sur un corpus journalistique du français, le *Monde*. Ils ont analysé des procédés morphologiques dans l'ensemble des articles parus en 1995 dans ce périodique en distinguant ses rubriques thématiques. Bien que la suffixation par -*ABLE* paraisse productive pour l'ensemble des articles de l'année, les auteurs observent des différences selon les rubriques : la RCL en -*ABLE* est productive dans les rubriques *Société*, *Livres* et *Agenda*, elle ne l'est pas dans la rubrique *International*. Le travail présenté dans Baayen (1994) est encore plus révélateur de la spécificité des procédés affixaux et montre que leur productivité peut même servir de critère pour effectuer automatiquement une typologie des textes. En faisant une étude comparative sur l'anglais entre les nouvelles, les livres pour enfants, les textes religieux et les textes officiels, l'auteur montre que les affixes permettent de faire une distinction assez nette entre ces quatre catégories. Par exemple, les affixes d'origine latine sont utilisés surtout dans les nouvelles et les livres pour enfants, tandis que les textes officiels et religieux ont une préférence nette pour les affixes d'origine germanique.

Comme cela avait été noté auparavant, les indices de productivité obtenus suite à de telles études restent spécifiques aux corpus étudiés. Une généralisation est toutefois possible si :
— les observations sont répétées sur plusieurs échantillons ;
— elles sont faites sur des corpus diachroniques ;
— elles sont faites dans des corpus synchroniques suffisamment grands pour assurer une meilleure précision des indices de productivité.

Les groupes de comparaison entre corpus ou rubriques d'un même corpus journalistique comme *Le Monde*, par exemple, peuvent être élaborés selon différents principes :
— dans Namer (2003a), l'ensemble des documents disponibles est retenu pour mener l'étude morphologique selon le principe rappelé en introduction de « more data is better data » ;
— dans (Fradin, Hathout & Meunier 2003; Gaeta & Ricca 2003) en revanche, un échantillonnage est effectué en nivelant les données par le nombre d'occurrences des affixations étudiées : les données sont comparables si elles présentent un nombre de formes comparable pour les RCL étudiées. Parmi ces occurrences, ce sont les nombres de types et d'hapax qui permettent de définir ensuite la productivité effective des règles ;
— dans Grabar & Zweigenbaum (2003), les corpus sont considérés comme comparables lorsqu'ils comportent des nombres d'occurrences proches : la productivité

des règles étudiées est fonction des dimensions stylistiques comme la spécialisation, le sous-domaine et le genre des corpus comparés.

En tout état de cause, quelles que soient les études menées en matière de productivité morphologique et les principes retenus pour mener à bien ces études, il est important de souligner qu'avant l'avènement des ressources textuelles sous forme numérisée, il était inenvisageable de mener de telles études, dans la mesure où elles auraient nécessité un travail manuel fastidieux. Le fait que les mesures de H. Baayen soient contemporaines de l'apparition de ressources numérisées de grande taille est à cet égard révélateur.

4.4. Etudes sur les manques

Pour terminer ce point consacré aux types d'études qu'ont permis de mener les ressources numérisées en morphologie, nous dirons quelques mots de l'exploitation qui peut être faite des manques, que la vastitude de ces ressources rend tout à fait pertinents.

Le manque devient dans ce cas révélateur de plusieurs phénomènes :

— il peut témoigner du caractère désuet ou technique de certains lexèmes figurant dans les dictionnaires ;

— il peut servir à vérifier empiriquement le bien-fondé d'hypothèses qu'on peut formuler à l'endroit d'une RCL (nous en avons donné un aperçu au §4.1.).

En tout état de cause, ce n'est qu'avec l'arrivée de ressources numérisées facilement interrogeables que l'absence de données lexicales est apparue comme pouvant constituer un fait tangible et exploitable.

Les études relatives aux manques nécessitent cependant que l'on prenne un certain nombre de précautions dans l'exploitation des corpus utilisés à cette fin. Tout d'abord, leur contenu doit être maximisé en fonction de l'usage souhaité, sinon la raison pour laquelle un lexème manque pourra être imputée au caractère non exhaustif du corpus de travail. Pour cette raison, et à moins que les lacunes à étudier ne concernent un domaine de spécialité, c'est la Toile qui est généralement choisie comme source de données.

Cependant, le contenu de la Toile n'est accessible que via le filtre que constituent les moteurs de recherche, et l'on est en droit de se poser la question de la fiabilité de ceux-ci pour retrouver l'ensemble des pages contenues dans la Toile. A ce sujet, Foenix-Rioux (2002) estimait qu'au mieux 20% du *Web visible* [16] était accessible par un moteur de recherche, et que ce pourcentage avait tendance à baisser. Un rapport récent, (Gulli & Signorini 2005), revient sur cette prévision pessimiste, en faisant état d'une nette amélioration de la couverture des moteurs, notamment grâce à l'évolution de leur technologie : sur les 11,5 milliards de pages estimées sur le Web visible, Google™ à lui seul en atteindrait plus de 76%, soit 69,6% de l'ensemble de la Toile[17]. Ainsi le repérage des manques est plus fiable, si bien que les conclusions qu'on peut en tirer sont plus vraisemblablement d'ordre linguistique que liées au système d'indexation des moteurs de recherche.

La connaissance de ces données, tout en nous permettant d'envisager les résultats obtenus avec une certaine confiance, nous incite néanmoins à une prudence relative. Ainsi, toute réponse apportée par l'intermédiaire d'un moteur devra être soumise à plusieurs vérifications :

---

[16] On nomme *Web visible* l'ensemble des pages qu'un moteur de recherche peut indexer, c'est-à-dire répertorier dans sa base de données. Le Web dit *invisible* comporte quant à lui les sites à accès réservé, les pages utilisant des technologies impropres à l'indexation (animations, codes informatiques,…) et les pages dynamiques, c'est-à-dire reliées à des bases de données.

[17] *Google*™ est le seul moteur en mesure de renvoyer des pages du *Web invisible*, qui constituaient déjà en 2002 1% de ses résultats, d'après l'URL : www.searchengineshowdown.com.

— tout d'abord par l'interrogation via un voire plusieurs autres moteurs de recherche. Ainsi, ABOLISSABLE et ABDOMINOSCOPIQUE sont tous deux référencés dans le *TLF*. ABOLISSABLE est absent de la Toile interrogée via Yahoo™, mais indexé via Google dans 171 documents, et doit donc être écarté de la liste des manques potentiels. Par contre, ABDOMINOSCOPIQUE, adjectif dénominal dont la base appartient au domaine médical, ne renvoie aucune page, quel que soit le moteur de recherche utilisé (Google™, Yahoo™, Altavista™, Seek™, Exalead™) ; la recherche échoue également lorsqu'on interroge CISMeF[18], un portail spécialisé dans les ressources médicales francophones[19]. Contrairement à ABOLISSABLE, il semble donc légitime d'admettre qu'ABDOMINOSCOPIQUE n'est plus utilisé.
— ensuite, par la répétition de la même requête à intervalles réguliers. La Toile peut en effet être vue comme une ressource de données fondamentalement synchroniques, dans la mesure où les pages indexées par les différents moteurs varient continuellement, en nombre et en contenu. Il devient dès lors nécessaire de vérifier que l'absence perdure au fil du temps, car seule l'absence répétée d'un lexème donné peut avoir un poids et servir d'indice pertinent ;
— enfin, les résultats obtenus sont à voir non pas comme des certitudes d'absences lexicales, mais plutôt comme de bons indicateurs de ces manques.

Un cas d'obsolescence lexicale a été mesuré à partir du contenu du *TLF*. Les 89 200 vocables qui constituent la nomenclature de ce dictionnaire ont constitué autant de requêtes qui ont servi à interroger le moteur *Yahoo* au moyen du robot WaliM (les précautions rappelées ci-dessus ayant été prises : requêtes répétées *n* fois à des intervalles de temps *t*). Parmi les requêtes qui ne renvoient aucune page à l'exclusion de celles qui contiennent des listes de mots, une grande partie semble morphologiquement construite (ASTRAGALISÉ, AVIVEUSE). L'absence d'attestation s'explique soit parce que le mot appartient à un domaine très spécialisé (ADIANTIFOLIÉ, ADMAXILAIRE) absent de la Toile, soit parce qu'il dénote une réalité disparue (BESQUINE, ancien bateau de pêche, ACCOUTREUR, ACCUBITEUR, noms de métier ou de fonction au moyen-âge), soit encore parce qu'il s'agit d'un hapax d'auteur (BADINGUEUSARD, dû à Jules Vallès, ou BARYTONNEUR créé par Colette).

En tout état de cause, cette expérience conduit ainsi à remettre en question l'intérêt en synchronie de 1,5% des entrées du *TLFi* (soit environ 1 200). Il ne s'agit pas ici de supprimer des dictionnaires les formes absentes de la Toile mais de prendre des distances par rapport aux attestations purement dictionnairiques pour les études de morphologie (cf. Gaeta & Ricca 2003).

5. Limites de l'apport des ressources numérisées en morphologie

On a vu jusqu'ici que, dans le domaine de la morphologie constructionnelle, les ressources numérisées sont devenues, en quelques années, proprement irremplaçables, à tel point qu'il est devenu pratiquement impensable de mener des études sans elles. Cependant, elles ne constituent pas toujours la panacée, et ne dispensent pas toujours le morphologue d'un travail fastidieux. Nous distinguons ici trois limites : la première est liée au format des requêtes (§5.1), le deuxième met en cause les RCL ne mettant en jeu aucun marquage segmental (§5.2), la troisième a trait à la représentativité des corpus (§5.3).

[18] URL: www.cismef.org/
[19] Requêtes effectuées le 2 septembre 2007, au moyen des formes *abdominoscopique* et *abdominoscopiques.*

5.1. Format des requêtes

La première de ces limites concerne le format d'interrogation des données. En effet, que ce soit pour récupérer toutes les occurrences de lexèmes construits par un procédé donné ou pour trouver l'attestation d'une forme construite, il faut trouver le format de requête adéquat. A ce stade, les formes fléchies d'un corpus constituent évidemment un obstacle, en particulier si l'on s'intéresse à la formation de verbes. Lorsqu'il s'agit de ressources contenues ou stockables sur un support de type CD-ROM, il suffit d'effectuer une lemmatisation préalable, et l'interrogation peut ensuite se faire par lemmes. Mais lorsqu'il s'agit d'interroger la Toile, la lemmatisation pose beaucoup de difficultés car la requête devra englober toutes les formes possibles, au moyen d'une liste ou d'un joker. La solution la plus répandue consiste alors à interroger la Toile sous la forme de requêtes multiples, ou de listes, contenant toutes ou les principales formes fléchies du lexème recherché. Ainsi, lorsque l'on a cherché une attestation du néologisme ALCOOTESTER, construit sur le nom ALCOOTEST, seule la forme fléchie à la deuxième personne du pluriel, *alcootestez*, a fourni un résultat via Google. Le format d'interrogation des données constitue donc un obstacle qui, bien que contournable, doit néanmoins être pris en compte lors de la constitution et de l'apprêt du matériel de travail.

5.2. Les RCL dépourvues d'exposant graphémique

Les corpus numérisés montrent également leur limite dans l'étude de certaines RCL, en particulier celles qui ne s'accompagnent d'aucun marquage segmental (exposant). C'est le cas de tous les procédés de conversion, que ce soit la conversion nom > verbe ($SCIE_N$ > $SCIER_V$), verbe > nom ($GARDER_V$ > $GARDE_N$), adjectif > verbe ($ROUGE_A$ > $ROUGIR_V$), adjectif > nom ($BLEU_A$ > $BLEU_N$) ou nom > adjectif ($ROSE_N$ > $ROSE_A$). Dans ces cas-là, aucun critère formel ne permet de faire une requête générale de type « trouver toutes les occurrences d'adjectifs issus de la conversion de noms » puisque, formellement, rien ne distingue le nom-base de l'adjectif dérivé, et ce quel que soit le type de données utilisées : Toile, dictionnaire en version numérique, collection de données sur CD-ROM, etc. Dépourvus de toute spécificité formelle, ces construits sont donc beaucoup plus difficiles à récupérer automatiquement que, par exemple, les noms suffixés en -*ETTE*, les adjectifs préfixés en *IN*- ou encore les verbes suffixés en -*ISER*. Les ressources textuelles numérisées se révèlent donc inexploitables pour ce qui est de récupérer plus facilement et en grand nombre les occurrences de lexèmes construits par conversion. Leur seul intérêt consiste à fournir, à condition que le corpus soit au préalable étiqueté morphosyntaxiquement, tous les adjectifs du corpus, à charge pour le morphologue de déterminer manuellement pour chacun s'il est converti depuis un nom ou pas.

La question se pose en termes assez similaires pour une recherche de (formes de) lexèmes construits résultant d'une règle de composition, qu'il s'agisse de lexèmes relevant de la composition ordinaire, ou de la composition dite savante ou néoclassique. Ainsi, les composés morphologiques ordinaires sont formés à partir de deux lexèmes, comme, par exemple les noms BALAI-BROSSE, PORTE-DRAPEAU, et l'adjectif GRIS-BLEU. Leur identification automatique est rendue impossible du fait de l'absence d'un marquage affixal commun. Certes, les constituants impliqués dans la formation des composés sont soumis à des contraintes sémantiques qui se manifestent par des restrictions catégorielles : par exemple, N-N, V-N et A-A sont des structures possibles de composés ordinaires, contrairement à V-A. Malheureusement, aucun étiqueteur n'est

en mesure de fournir la catégorie interne des composés (*cf.* §3.1.1), ce qui fait que cette connaissance n'est pas utilisable.

Le seul élément graphique qui semble pouvoir être exploité pour la collecte des composés ordinaires est le tiret. Cependant, l'usage de cette marque comme filtre de recherche à partir d'un corpus électronique présente trois inconvénients : (1) la présence du tiret n'est pas systématique dans les composés (on retrouve PORTECLEFS ou PORTE CLEFS aux côtés de PORTE-CLEFS) (voir Mathieu-Colas 1994) ; (2) elle ne dénote pas forcément un composé morphologique : même restreinte aux catégories N et A, une requête basée sur le tiret ramène également PORTE-À-PORTE, VERT-DE-GRIS, etc., qui sont des phrasèmes mais pas des composés ; (3) enfin, le tiret est invisible aux moteurs de recherche, ce qui le rend inexploitable pour collecter des composés sur la Toile. En d'autres termes, l'unique moyen de répertorier les composés (ordinaires ou savants) présents dans les ressources numérisées, consiste à se constituer artificiellement l'ensemble des lexèmes composés potentiels à partir de listes de constituants (par exemple, tous les noms issus des nomenclatures du *TLFi* sont combinés systématiquement deux à deux pour former la liste des composés N-N candidats) et de s'en servir comme autant de requêtes dans ces ressources.

5.3. Représentativité des corpus

La question de la représentativité des corpus se pose de façon particulièrement criante dans les études portant sur la productivité morphologique. On a dit qu'une RCL est productive quand elle est susceptible de construire de nouveaux lexèmes. On s'attend donc à ce qu'en étendant le corpus d'observation, on rencontre des attestations de nouveaux lexèmes construits grâce à elle. Ainsi, lors de l'étude de l'année 1995 du *Monde*, Grabar et al. (2006a) observent la croissance du nombre d'adjectifs suffixés en -*ABLE* différents au fil de l'année : même si cette croissance ralentit à mesure que le corpus s'étend, elle reste positive y compris dans les derniers mois de l'année. Cette définition de la productivité repose sur la disponibilité d'un corpus extensible, au moins en puissance. Le corpus examiné (par exemple, les articles de la rubrique *Société* d'une année du *Monde*) n'est alors qu'une partie, un échantillon d'un corpus plus grand (par exemple, l'ensemble des articles de la rubrique *Société* parus dans *Le Monde*). Ce corpus plus grand, potentiellement infini (ensemble des articles de la rubrique *Société* à paraître dans *Le Monde*), est caractérisé par des critères de sélection (ici, journal = *Le Monde*, rubrique = *Société*). Cette définition par critères (en intension) plutôt que par énumération de tous les textes à inclure (en extension) est une condition nécessaire pour que ce corpus soit extensible.

On peut considérer qu'une étude sur un corpus d'articles du *Monde* vaut pour mesurer la productivité de procédés morphologiques dans le corpus sans cesse croissant des articles publiés dans ce journal. Le corpus présent est utilisé comme un échantillon du corpus global, et sa représentativité par rapport à ce corpus global est à questionner. En effet, auteurs, dates de parution, thèmes traités, rubriques du journal, ces critères et bien d'autres sont susceptibles de provoquer des variations sensibles de la langue employée et donc des types de mots construits que l'on y rencontre. Par exemple, Grabar et al. (2006a) relèvent que le nombre de types d'adjectifs suffixés en -*ABLE* est plus important dans le corpus formé des articles de la rubrique *Société* du *Monde* 1995 que dans celui formé des articles de la rubrique *International*, pourtant presque deux fois plus grand en taille totale. Une différence de répartition des rubriques au cours du temps pourrait donc induire des différences notables dans la productivité morphologique observée.

On pourrait également être tenté de considérer qu'une étude sur *Le Monde* est représentative d'un corpus couvrant un genre textuel journalistique non limité à ce journal. On doit néanmoins s'attendre à des différences notables entre le vocabulaire du *Monde* et celui, par exemple, de *Libération*, pour prendre un autre grand quotidien national, ce qui limite cette représentativité.

En somme, la portée des mesures de productivité effectuées sur ces corpus-échantillons (comme plus largement celle de mesures quantitatives réalisées à d'autres paliers linguistiques) est à relativiser. Un enjeu est de préciser cette portée, en caractérisant les types de textes auxquels les mesures effectuées seraient généralisables. Mais il faut prendre garde d'éviter une caractérisation circulaire qui se fonderait précisément sur les propriétés de productivité de ces textes.

Notons enfin que si l'on s'intéresse à la productivité d'un procédé morphologique « en général », le corpus de travail est alors considéré comme un échantillon de « langue générale ». La question de sa représentativité se pose alors de façon encore plus pressante. Cette situation se retrouve dans les nombreux travaux sur la constitution de corpus « équilibrés » (voir par exemple la discussion qu'en fait Péry-Woodley (1995)).

5. **Conclusion**

Condamines (2005a) rappelle avec insistance que l'accès à des ressources numérisées à grande échelle a introduit une rupture pour la linguistique moderne (celle qui se fonde sur l'existence d'un différentiel grammatical (Milner 1989: 55-56)), dans la mesure où, désormais, « l'étude des phénomènes langagiers ne se fait plus seulement de manière introspective mais à partir de productions réelles » (Condamines 2005b : 35). Bien qu'elle contienne une part de vérité, cette vision présente les choses de manière naïve et partielle. D'une part, l'introspection n'acquiert tout son poids que si elle est partagée : tous les linguistes ont insisté sur le caractère reproductible que doit avoir le jugement de grammaticalité (ou d'acceptabilité) (Gross 1997: 74) et les études sérieuses se sont attachées à travailler sur des données qui satisfont cette exigence (voir, par exemple, les articles dans Godard (2003)). D'autre part, beaucoup se sont efforcés de pallier l'absence des ressources naguère disponibles par la collecte individuelle de données, lesquelles pouvaient figurer sur des supports très divers (livres, journaux, scripts, étiquettes, annonces, modes d'emploi, tracts, etc.). Or ces données sont tout aussi réelles que celles qui figurent sur la Toile ou dans des corpus électroniques. Le cheminement suivi dans cet article conduit à une vision plus nuancée des choses.

Tout d'abord, il est indéniable que les ressources numérisées de grande ampleur ont changé du tout au tout la perspective de travail pour le morphologue. En introduisant tendanciellement ce qu'on pourrait appeler une « saturation des possibles de langue », ces ressources rendent possible la validation d'hypothèses relatives au domaine d'application de certains procédés dérivationnels. Elles permettent aussi de constituer des séries de formes qui rendent possible la formulation des contraintes prosodiques et phonologiques à l'œuvre en morphophonologie. Enfin en faisant, pour ainsi dire, « parler les absents », elles confèrent à l'absence systématique d'une forme le statut d'hypothèse sur le système de la langue : l'absence traduit une lacune dans le système de la grammaire qu'il faudra traiter comme telle dans les descriptions linguistiques qu'on en donnera. Pour toutes ces raisons, il est clair que le changement quantitatif apporté par l'accès à des ressources numérisées massives s'est traduit par un changement qualitatif au niveau de la description des phénomènes linguistiques.

Il reste que ces ressources laissent intacte la nécessité du recours au jugement du locuteur face à des formes récupérées dans les textes. Cet aspect a été illustré à plusieurs

reprises dans cet article, à propos de l'apprêt des données notamment. Ce qui a changé toutefois, c'est l'introduction de la variation dans les données à des doses souvent importantes (au lieu d'une forme attendue, on en a plusieurs). Il s'agit d'un des effets les plus frappants du travail à partir « de données réelles ». Cela ne remet pas en cause fondamentalement le travail des linguistes, qui n'ont pas attendu de travailler sur des ressources numérisées pour traiter de la variation. Néanmoins, comme nous l'avons souligné ici, un des aspects nouveaux de la situation tient au fait qu'on peut reformuler comme une hypothèse à tester les décisions qu'on prend concernant le statut des données, pourvu qu'on en garde la trace : décider, par exemple, de considérer comme valides des données précédemment écartées parce qu'elles comportaient une faute de frappe et mesurer leur impact sur la productivité du procédé étudié. En bref, le travail sur des corpus de grande taille incite à un travail de collecte des données plus systématique et mieux contrôlé, au sens où le descripteur-linguiste doit rendre des comptes sur les sélections qu'il fait.

Le dernier point à mentionner concerne le statut des données fournies par les ressources numérisées. Ainsi qu'on l'a dit, le seul fait qu'une forme soit attestée sur la Toile ou dans un corpus ne suffit pas à en faire une donnée recevable. Comme pour la collecte de données auprès de locuteurs, il faut encore qu'elle soit reproductible. Malheureusement, le nombre de formes recueillies pour chaque procédé étant souvent petit (en morphologie notamment), cette condition reste difficile à satisfaire. L'autre problème concerne le contrôle de la source lorsque les données sont issues de rassemblements arbitraires (Toile). Il arrive que les seules données qu'on glane proviennent d'études écrites par celui qui fait la requête. Ce problème peut certes être circonscrit facilement. Il n'en irait pas de même si des données fabriquées pour la circonstance étaient mises en grand nombre sur la Toile. On aurait alors une mystification du type de celle qu'on a eue avec l'homme de Piltdown, et elle ne serait pas plus facile à découvrir. Un dernier problème concerne la compétence des scripteurs. Pour savoir si une forme récoltée est une forme aberrante ou non, il faudrait avoir une idée de la compétence de celui qui l'a produite (et aussi de quelques autres paramètres : effet visé, etc.). Ceci n'est pas toujours facile à déterminer d'après le contexte, ce qui devrait renforcer la vigilance des utilisateurs. L'importance de ce point ressort mieux quand on sait que les psycholinguistes écartent de leurs données les réponses déviantes fournies par leurs sujets, car elles faussent les statistiques. On voit que le critère de sélection ultime qui prévaut pour la sélection des données linguistiques reste opératoire quel que soit le mode de collecte : en compétence auprès de locuteurs, à travers des attestations en corpus ou dans des rassemblements arbitraires de données. Dans tous les cas, il s'agit du jugement de grammaticalité, même si le recours aux corpus fait qu'il est mieux contrôlé et relativisé par la variation.

6. Bibliographie

Apothéloz Denis & Gilles Boyé. (sous presse). "Remarques sur la compositionnalité en morphologie". *Verbum* **XXVI** 4:375-385.

Baayen Harald R. 1993. "On frequency, transparency and productivity". *Yearbook of Morphology* 1992:181-208.

Baayen Harald R. 1994. "Derivational Productivity and Text Typology". *Journal of Quantitative Linguistics* **1** 1:16-34.

Baayen Harald R. & Rochelle Lieber. 1991. "Productivity and English derivation: a corpus-based study". *Linguistics* **29** 5:801-843.

Baayen Harald R. & Antoinette Renouf. 1996. "Chronicling the **Times**: Productive lexical innovations in an English newspaper". *Language* 72:69-96.

Brunet Etienne. 1981. "Les suffixes". In *Le vocabulaire français de 1789 à nos jours. D'après les données du Trésor de la langue française*. 415-493. Genève: Librairie Slatkine.

Catach Nina. 1995. *L'Orthographe française*. 3ème édition. Paris: Fernand Nathan.

Chomsky Noam A. & George A. Miller. 1963. *L'analyse formelle des langues naturelles*. Translated by Richard P. & N. Ruwet. 1968. Paris / La Haye: Gauthier Villars / Mouton.

Condamines Anne. 2005a. "Sémantique et corpus, quelles rencontres possibles?" In *Sémantique et corpus*, Condamines A. (ed). 15-38. Paris: Hermès / Lavoisier.

Condamines Anne (ed) 2005b. *Sémantique et corpus*. Paris: Hermès.

Corbin Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique*. 2 vols. Lille: Presses Universitaires du Septentrion. Original edition, Tübingen: Niemeyer.

Corbin Danielle & Pierre Corbin. 1991. "Un traitement unifié du suffixe **-er(e)**". *Lexique* 10:61-145.

Coseriu Emil. 1967. "Sistema, norma y habla". In *Teoría del lenguaje y lingüistica general*. 11-113. Madrid: Editorial Gredos.

Dal Georgette. 2003. "Productivité morphologique: définitions et notions connexes". *Langue française* 140:3-37.

Dal Georgette, Natalia Grabar, Stéphanie Lignon, Clément Plancq & Delphine Tribout. (à par.). "Les adjectifs en **inXable** du français". *Linguisticae Investigationes*.

Dal Georgette & Fiammetta Namer. 2000. "Génération et analyse automatiques de ressoureces lexicales construites utilisables en recherche d'information". In *Traitement automatique des langues pour la recherche d'information*, Jacquemin C. (ed). 423-446. Paris: Hermès.

Dal Georgette & Fiammetta Namer. 2005. "L'exception infirme-t-elle la notion de règle?" *Faits de langue* 25:123-130.

Dubois Jean. 1969. *Grammaire structurale du français: la phrase et les transformations*. Paris: Larousse.

Foenix-Rioux Béatrice. 2002. *Guide de recherche sur Internet*. Paris: Nathan.

Fradin Bernard, Nabil Hathout & Fanny Meunier. 2003. "La suffixation en ***-ET*** et la question de la productivité". *Langue française* 140:56-78.

Gaeta Livio & Davide Ricca. 2003. "Italian prefixes and productivity: a quantitative approach". *Acta Linguistica Hungarica* **50**:89-108.

Godard Danièle (ed) 2003. *Les langues romanes. Problèmes de la phrase simple*. Paris: CNRS Editions.

Grabar Natalia, Georgette Dal, Bernard Fradin, Nabil Hathout, Stéphanie Lignon, Fiammetta Namer, Clément Plancq, Delphine Tribout, François Yvon & Pierre Zweigenbaum. 2006a. "Productivité quantitative de la suffixation par **-Able** dans un corpus journalistique du français". In *Lexicometra. 8e JADT*, Viprey J.-M. (ed). 473-485. Besançon: Presses Universitaires de Franche-Comté.

Grabar Natalia, Georgette Dal, Bernard Fradin, Nabil Hathout, Stéphanie Lignon, Fiammetta Namer, Clément Plancq, Delphine Tribout, François Yvon & Pierre Zweigenbaum. 2006b. "Productivité quantitative des suffixation par **-ité** et **-Able** dans un corpus journalistique du français". In *Verbum ex machina. Actes de la 13e conférence sur le TALN*, Mertens P., C. Fairon, A. Dister & P. Watrin (eds). 167-175. Louvain: Presses Universitaires de Louvain.

Grabar Natalia & Pierre Zweigenbaum. 2003. "Productivité à travers domaines et genres: dérivés adjectivaux et langue médicale". *Langue française* 140:102-125.

Gross Maurice. 1997. "Synonymie, morphologie dérivationnelle et transformations". *Langages* 128:72-90.

Gulli Antonio & Alessio Signorini. 2005. "The indexable Web is more than 11.5 billion pages". In *Poster proceedings of the 14th international conference on World Wide Web*. 902-903. Chiba, Japan: ACM Press.

Habert Benoît. 2000. "Des corpus représentatifs: de quoi, pour quoi, comment?" In *Linguistique sur corpus. Etudes et réflexions*, Bilger M. (ed). 11-58. Perpignan: Presses Universitaires de Perpignan.

Habert Benoît, Adeline Nazarenko & André Salem. 1997. *Les linguistiques de corpus*. Paris: Armand Colin / Masson.

Habert Benoît & Pierre Zweigenbaum. 2002. "Régler les règles". *Traitement automatique des langues* **43** 3:83-102.

Hathout Nabil, Marc Plénat & Ludovic Tanguy. 2003. "Enquête sur les dérivés en **-able**". *Cahiers de grammaire* 28:49-90.

Jacquemin Christian. 1999. "Syntagmatic and paradigmatic representations of term variation". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 341-348. Baltimore: University of Maryland.

Keune Karen, Roeland Van Hout & Harald Baayen. 2006. "Socio-geographic variation in morphological productivity in spoken Dutch: a comparison of statistical techniques". In *JADT 2006*, Viprey J.-M. (ed). 571-580. Besançon: Presses Universitaires de Franche-Comté.

Lignon Stéphanie. 2000. La suffixation en **-ien**. Aspects sémantiques et phonologiques. Thèse de doctorat, Université de Toulouse le Mirail, Toulouse.

Marcus Mitchell, Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. "Building a large annotated corpus of English: the Penn Treebank". *Computational Linguistics* **19** 2:313-330.

Mathieu-Colas Michel. 1994. *Les Mots à traits d'union*. Paris: Didier Erudition.

Matthews Peter Hugoe. 1974. *Morphology*. 2nd Edition 1991. Cambridge: Cambridge University Press.

Milner Jean-Claude. 1989. *Introduction à une science du langage*. Paris: Le Seuil.

Namer Fiammetta. 2000. "FLEMM: un analyseur flexionnel du français à base de règles". *Traitement automatique des langues* **41** 2:523-547.

Namer Fiammetta. 2003a. "Productivité morphologique, représentativité et complexité de la base: le système MoQuête". *Langue française* 140:79-101.

Namer Fiammetta. 2003b. "WaliM: valider les unités morphologiquement complexes par le Web". In *Les unités morphologiques*, Vol. 3, *Silexicales*, Fradin B., G. Dal, N. Hathout, F. Kerleroux, M. Plénat & M. Roché (eds). 142-150. Villeneuve d'Ascq: SILEX: CNRS & Université Lille 3.

Péry-Woodley Marie-Paule. 1995. "Quels corpus pour quels traitements automatiques?" *Traitement automatique des langues* **36** 1-2:213-222.

Plénat Marc. 1997. "Analyse morpho-phonologique d'un corpus d'adjectifs en -**esque**". *French Language Studies* **7**:163-179.

Plénat Marc. 2002. "Jean-Louis Fossat: fossatissime. Note sur la morphophonologie des dérivés en **-issime**". In *Hommage à Jean-Louis Fossat*, Rabassa L. (ed). 229-248. Toulouse: CLID - Université de Toulouse 2.

Schmid Helmut. 1994. "Probabilistic part-of-speech tagging using decision trees". In *Proceedings of the International Conference on New Methods in Language*

Processing, Sima'an K., R. Bod, S. Krauwer & R. Scha (eds). 44-49. Manchester: UMIST.

Schultink Henk. 1961. "Produktiviteit als Morpfologisch Fenomeen". *Forum te Letteren* **2**:110-125.

Sinclair John. 1996. *Preliminary recommendations on Corpus Typology*. Technical Report N° 00. CEE. Bruxelles.

English abstract

The aim of this article is threefold. Firstly, it aims to recall how accessing large scale digitized data resources has qualitatively changed the way of doing morphology. Secondly, it shows that raw data extracted from such resources cannot be used as such and requires in-depth preparation in order to be properly exploited; it also shows that the procedures involved in such preparation have to be carefully made explicit since they have a strong impact on the results of queries submitted to sorted data. Finally, it will be argued that using large scale digitized data does not allow us to disregard speakers' judgements concerning grammaticality, which remain indispensable whenever one needs to determine whether a linguistic form is acceptable or not, but that it puts such judgements in perspective insofar as they have to be compared with what is attested in the data.