Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie

Georgette Dal, Fiammetta Namer

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie

Dal, Georgette* & Namer, Fiammetta**

*UMR 8163, STL, CNRS & Universités Lille 3, Lille 1
** UMR 7118, ATILF, CNRS & Université de Lorraine
georgette.dal@univ-lille3.fr, fiammetta.namer@univ-nancy2.fr

1 Introduction

Pour mener leurs études, les morphologues, et, singulièrement parmi eux, ceux qui traitent du lexique construit, sont pris entre deux tensions : utiliser les données issues des dictionnaires, réputées fiables parce que passées au travers du crible de l'institutionnalisation, ou se fonder sur des données authentiques, produites dans des situations écologiques. Bien qu'inspirant encore de la défiance, le Web devient alors la ressource par excellence. Les ressources numérisées telles que les archives de journaux constituent un moyen terme entre ces deux pratiques : elles permettent d'effectuer des recherches basées sur l'usage, tout en offrant la garantie d'une certaine tenue linguistique de la part des scripteurs, la plupart du temps identifiables.

Dans le présent papier, nous nous demanderons si les dictionnaires méritent la confiance que continuent de leur octroyer les morphologues ou, plus exactement, si une morphologie qui puise ses données dans les seuls dictionnaires peut encore avoir cours au 21e siècle. Pour engager cette discussion, nous commencerons par poser la question de savoir ce que signifie exister, pour un lexème construit. Nous exposerons ensuite ce que constitue une morphologie basée sur l'usage, puis donnerons un aperçu des types de recherches que permettent de réaliser les ressources non dictionnairiques.

2 Exister pour un lexème construit

La défiance que continue d'inspirer le Web à au moins une partie des morphologues et, inversement, la confiance qu'ils ont dans les dictionnaires en matière de lexique construit ont partie liée avec la question de la légitimité des données sur lesquelles se fondent les analyses, donc avec celle de leur existence.

À la question de savoir ce que signifie exister, pour un lexème construit, deux réponses sont possibles : un lexème construit existe dès lors qu'il est dans le dictionnaire ; un lexème construit existe dès lors qu'il est dans l'usage.

Pour le non linguiste (français ?), voire pour le linguiste non morphologue, la seule existence qui vaille en matière de mots est dictionnairique. Nous citerons le postulat tacite que mentionne Jean Pruvost[1] : « tout ce qui est mentionné dans les dictionnaires est indiscutable, et le mot qui n'y figure pas n'existe pas. D'où cette remarque habituelle au point de devenir un cliché : "Ce mot n'est pas français, il n'est pas dans le dictionnaire" ». À leur manière, témoignent de cet état de fait les extraits suivants, relevés sur la Toile dans des forums de discussion en novembre 2011 :

> « Pourquoi le mot *islamophobie* n'existe pas dans le dictionnaire Larousse ? »
>
> « Pourquoi *appréhendable* ne figure-t-il pas dans le dictionnaire ? »
>
> « Désolé mais la corbeautière n'existe pas dans le dictionnaire »
>
> « Pourquoi le verbe *futiliser* n'est pas dans le dictionnaire ? – Parce qu'il n'existe pas »

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100217

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

> « Le mot qui vient spontanément à l'esprit est *émin*ç*age*, malheureusement il n'existe pas puisqu'on ne le trouve pas dans le dictionnaire. Serait-il vraiment fautif de l'utiliser ? »

De même, en 1997, Danielle Corbin s'interrogeait sur la notion de mot existant, et finissait par conclure que, pour pouvoir être dit comme tel, un mot devait être répertorié dans un grand dictionnaire de référence synchronique :

> « La notion de mot existant implique une instance d'actualisation qui permette d'authentifier le mot. Cette instance peut être, au choix, les dictionnaires, les productions langagières primaires ou encore le stock lexical que croit connaître le linguiste en situation d'effectuer une classification de ce point de vue. Par rapport à cette instance, l'existence d'un mot est une notion discrète. Pour des raisons de commodité pratique (seuls les dictionnaires offrent un accès objectivable aux unités actualisées, même s'ils ne peuvent répertorier que celles qui figurent dans les documents à partir desquels ils sont constitués), j'entendrai ici par *mot existant* un mot répertorié dans l'un ou l'autre des grands dictionnaires actuels de référence, à savoir, pour l'essentiel, le *Grand Robert de la langue française* (*GRLF*) et le *Trésor de la langue française* (*TLF*) » (Corbin, 1997 : 79).

Pour exister, il suffit dans ce premier cas d'être dans le dictionnaire. Encore faut-il savoir de quel dictionnaire il est question : dictionnaire de langue générale, dictionnaire encyclopédique ou dictionnaire de spécialité, mono- ou multivolume, dictionnaire synchronique ou diachronique, dictionnaire monolingue ou multilingue, etc.

L'autre façon d'exister, pour un lexème construit, est d'être présent dans l'usage, indépendamment de son attestation dictionnairique, car, comme l'écrit très justement un internaute à propos des mots en général : « [s]i les mots avaient dû attendre un dictionnaire pour exister, le dictionnaire n'existerait pas ». Dans ce cas, un mot existe dès lors qu'il est produit, à l'oral ou à l'écrit, par un locuteur ou scripteur, et qu'il est interprétable. Nous reviendrons sur la façon de capter l'usage dans le § 3.

Les deux modes d'existence qui précèdent constituent deux vues différentes, qui peuvent être le cas échéant complémentaires. L'option pour l'un ou l'autre mode dépendra alors de l'objet que s'assigne le morphologue, dans la mesure où les lexèmes construits auxquels il aura accès ne seront pas les mêmes.

Par définition, une morphologie basée sur les dictionnaires prendra dans ses filets des archaïsmes ou des mots marqués du diacritique « vieux » (cf. 1), des mots a priori analysables comme construits apparus en français du XI^e^ au XXI^e^ siècle pour les éditions les plus récentes, éventuellement empruntés construits dans leur langue source (cf. 2), de fréquences très variables qui vont de l'hapax d'auteur (cf. 3a) au lexème à haute voire très haute fréquence (cf. 3b)[2]. On note d'ailleurs (cf. 4) que des lexèmes présentés comme hapaxiques dans le *TLF* peuvent avoir un usage fréquent sur le Web :

(1) accélératif ; accortise ; accueillance ; accueillement ; accusataire ; accroissance ; acérique ; acotylédon ; actinifère ; admonitif ; assaillement ; bi-hebdomadaire (« Qui se fait, qui paraît toutes les deux semaines ») ; billonnage (*s.v.* **billonner**) ; brûlotier (*s.v.* **brûlot**) ; buissonnier (« Qui vit dans les buissons ») ; caléfacteur ; cannamelliste (*s.v.* **cannamelle**) ; célestiel (*s.v.* **céleste**) ; célérimètre ; chapitrer ; chauvinique (*s.v.* **chauvin**) ; chenâtre ; chouriner ; concubinaire ; doreloterie ; drageoir ; écorcherie ; embabouiner ; encourtiner ; endormement ; ignorable (*s.v.* **ignorer**) ; flammette (*s.v.* **flamme**) ; huissier (« Gardien d'une porte ») ; outrecuider ; vaillantise

(2) XI^e^ siècle : commandement (1050) ; douloureux (1050 ; du b. lat. *dolorosus* « douloureux ») ; nobilité (1050 ; empr. au lat. *nobilitas*)

XII^e^ siècle : âpreté (1190 ; du lat. *asperitas*) ; bombancier (ap. 1170) ; chantable (déb. XII^e^ s.) ; tiédeur (fin XII^e^ s.)

XIII^e^ siècle : arrangement (XIII^e^ s.) ; moutardier (1292) ; plumasseau (1240)

XIV^e^ siècle : aiguiseur (1300) ; insinuation (1319 ; empr. au lat. *insinuatio*) ; laitage (fin XIV^e^ s.)

XV^e^ siècle : ventricule (1478 ; empr. au lat. *ventriculus*) ; ventru (1490) ; verbosité (1496-97 ; empr. au b. lat. *verbositas, -atis*)

XVI^e^ siècle : alpestre (1555 ; empr. à l'ital. *alpestre*) ; ameublir (XVI^e^ s.) ; antagoniser (ap. 1500 ? ; empr. au gr. « lutter »)

XVII^e^ siècle : dominoterie (1640) ; invariabilité (1616) ; vertébral (1675)

XVIII^e^ siècle : acidifiable (1786) ; aiguiserie (1784) ; amygdaloïde (1752 ; empr. au gr. « en forme d'amande »)

XIX^e^ siècle : bimétallique (1876) ; centrifuger (1871) ; débondage (1805-08)

XX^e^ siècle : activeur (1953) ; apolitique (1927) ; référentiel (1965) ; verdunisation (1928)

(3) a- amicoter ; félibrique (*s.v.* **félibre**) ; feuilletonisme (*s.v.* **feuilloniste**) ; fleurance (*s.v.* **fleurer**) ; finassement (*s.v.* **finasserie**) ; fuyeuse (*s.v.* **fuir**) ; gavrochiner (*s.v.* **gavroche**) ; gigantisation (*s.v.* **gigantesque**) ; goémoneux (*s.v.* **goémon**) ; gouvernementomane (*s.v.* **gouvernement**) ; hanchu (*s.v.* **hanche**) ; hargnosité (*s.v.* **hargne**) ; heural (*s.v.* **heure) ;** homme-chèvre (*s.v.* **homme**, élém. de compos.) **;** hybridage (*s.v.* **hybridation**) **;** ignorable (*s.v.* **ignorer**) ; imagette (*s.v.* **image**) ; immédiatiser (*s.v.* **immédiat**) ; incursionner (*s.v.* **incursion**) ; infinitéisme

b- anticipation (fréq. abs. littér. 271) ; apprentissage (fréq. abs. littér. 479) ; brasserie (fréq. abs. littér. 308) ; délicieux (fréq. abs. littér. 3596) ; fournisseur (fréq. abs. littér. 409) ; indigence (fréq. abs. littér. 337) ; libération (fréq. abs. littér. 920) ; monarchie (fréq. abs. littér. 1855) ; travailleur (fréq. abs. littér. 1453)

(4) gigantisation (602) ; ignorable (52100) ; imagette (4330000)

En revanche, échapperont à la description par définition les lexèmes construits non encore passés par le sas de l'entrée dans le dictionnaire, fussent-ils très fréquents (cf. 5[3]), les créations à la volée comme celles sous (6)[4] ainsi que l'ensemble des mots ne relevant pas du projet éditorial du ou des dictionnaires consultés :

(5) américanité (145096) ; bien-pensance (134000) ; contre(-)productif (425000) ; doublonner (9930) ; européanité (23507) ; fiabiliser (225000) ; gravage (637200) ; livrage (13900) ; mesurette (422000)

(6) mais je t'ai **poissonifié d'avril** (si ça existe :)D) en te disant que c'était Charlemagne

Bientôt **plurieliste**....si ça se dit [à propos de la Citröen Pluriel]

Ouh lala Bloodshed a juste l'esprit **partageur** (je sais pas trop si ça se dit mais tant pis)

Tout d'abord, il faut impérativement que tout artiste/auteur soit inscrit sur les forums **résaldiens** (je sais pas si ça se dit mais ça sonne con donc je garde)

anniversaire et **mensuversaire**!!!si ça se dit lol!

3eme étape : **l'éfritage** (je sais pas si ca se dit mais on s'en fou...)

Dans le pire des cas il faudrait que vous fassiez venir un expert en "**dératisation**" (je ne sais pas si ça se dit...) pour éliminer ces rongeurs

Après avoir "**décalcairisé**" (je ne sais pas si ça se dit ?), voilà le résultat

après c'est question de choix et d'**assumage** (lol, ché pas si ça se dit ça !)

Sensationnel, **adrénalinique** (j'sais pas si ça se dit, mais quand même adrénalinique !)

Pour info, le but est de voir, ou plutot d'entendre une amélioration dans la "**suavité**" (je ne sais pas trop si ça se dit...) des timbres comme annoncée par Frank

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

> Je ne pense pas réussir à finir cette mission, mais merci pour le **repoussement** (je sais pas si ça se dit) de 24 heures
>
> Je comprends parfaitement les gens qui trouvent ça bizarre d'être **émétophobiste** (si ça se dit) parce que je trouve ça bizarre aussi

Relativement à la question qui fonde la présente réflexion (quelles données prendre en compte pour mener des analyses en morphologie ?), la question à se poser est alors : « quel est le mode d'existence le plus intéressant relativement à l'objectif que s'assigne le morphologue ? ». S'il s'agit pour lui de mener des études diachroniques (par exemple, combien de lexèmes en *-able* sont apparus en français entre les XIII^{e} et XVIII^{e} siècles ?), alors les dictionnaires sont de précieux outils. De même, ils pourront également servir à dégager des ébauches d'hypothèses pour l'étude d'un procédé donné (cf. par exemple Dal & Namer, 2010b, à propos de la suffixation en *-ance*).

En revanche, si le morphologue se donne comme objectif de dégager les régularités à l'œuvre dans le lexique construit actuel, plutôt que de se trouver face au défi de proposer une analyse si possible unifiée ou de construire un modèle pour des données relevant d'époques différentes, dont certaines se sont opacifiées et chargées d'idiosyncrasies au fil des siècles si bien qu'il ne sait plus quelle langue il étudie (s'agit-il de synchronie ? d'achronie ? de panchronie ?), il convient de privilégier l'existence dans l'usage, indépendamment de l'attestation dictionnairique, ce d'autant plus qu'il pourra tirer partie de la mise en contexte des lexèmes construits. À charge pour lui d'affiner au préalable la notion d'usage : usage de qui ? Quel degré de confiance apporter aux locuteurs/scripteurs ? Etc. C'est ce à quoi nous nous employons dans le paragraphe suivant, après avoir défini ce qu'est l'usage.

3 Morphologie basée sur l'usage

Les tenants de la linguistique basée sur l'usage se donnent pour but d'expliquer le fonctionnement de la langue à partir de ses fonctionnalités. Préconisé par les fonctionnalistes (Greenberg, 1966 ; Givón, 1979), formalisé dans un premier temps par Bybee (1985) dans son *modèle dynamique*, ou par Langacker (1991) (*modèle basé sur l'usage*), ce mouvement, rejoint par les linguistes cognitivistes, est désormais connu sous le nom de théorie basée sur l'usage (*Usage-Based Theory*, cf. Barlow & Kemmer, 2000 ; Langacker, 2000 ; Bybee, 2001). Il présente la structure de la langue comme le résultat de contraintes imposées par son usage, et que l'on peut résumer ainsi : la langue est comme elle est, car elle s'adapte aux besoins des locuteurs.

3.1 Représenter l'usage

Par conséquent, une forme existante en usage se caractérise par sa fréquence d'emploi, son registre, le type de discours où elle apparaît, les facultés langagières du locuteur, l'époque où elle a été prononcée ou écrite, etc. L'usage comporte donc théoriquement de l'écrit, de l'oral, les productions actuelles, passées, ou en cours d'élaboration. L'accès direct à ce que recouvre l'usage n'est pas envisageable, et, pour identifier si un mot existe, le morphologue doit nécessairement se faire une représentation approximative et simplifiée de l'usage, qui en constitue une projection raisonnable quantitativement et qualitativement.

C'est dans ce but qu'il va avoir recours à des données textuelles numérisées. Celles-ci sont en mesure de modéliser l'univers du discours, car elles sont à la fois massivement accessibles et facilement exploitables. Depuis une dizaine d'années, ces deux atouts mettent, en quelque sorte, l'usage à la portée du linguiste, grâce d'une part à la démocratisation spectaculaire des capacités de stockage des ordinateurs de plus en plus performantes, et d'autre part à l'évolution des techniques informatiques de recherche d'information, qui simplifient et accélèrent la fouille de ces grandes quantités de données informatisées.

Parce qu'elles ont des origines, des contenus et des finalités différents, nous distinguons ici deux types de ressources numérisées : les corpus et les données de la Toile.

3.2 Usage et corpus numérique

Un corpus résulte de la constitution d'observables linguistiques à partir de données empiriques issues de l'usage et censées échantillonner celui-ci. Le recueil des données, leur organisation et leur calibrage doivent permettre de construire une ressource combinant homogénéité et exhaustivité : sélection et délimitation qualitative des textes en fonction de leur genre, type, nature, registre, et choix du seuil d'occurrences validant la pertinence quantitative du corpus à constituer.

L'origine des données recueillies doit satisfaire le type de recherche qui intéresse le morphologue : documents oraux retranscrits, interactions, journaux d'information, articles scientifiques, œuvres littéraires, ….

Pour qu'il puisse explorer le corpus efficacement, les données contenues doivent être annotées : les descriptions, le plus souvent sous forme de balises xml, définissent les parties du discours, les traits syntaxiques, morphologiques ou sémantiques, renseignent sur l'auteur d'un fragment de texte ou précisent une rubrique de journal (sur la réalisation des tâches d'apprêt des corpus, cf. Fradin et alii, 2008).

Le corpus est souvent organisé en bases de données, de manière à optimiser, pour l'utilisateur, les tâches d'exploration et d'analyse. Celles-ci comportent, entre autres, les calculs lexicométriques (décompte du nombre d'occurrences, du nombre de types pour chaque (forme de) lexème, mesure de fréquence), la constitution de listes (vocabulaire, n-grammes[5], cooccurrences), ou la navigation (identification des contextes syntaxiques, des concordances, des voisins distributionnels). À titre d'exemples, le lecteur pourra se reporter aux bases lexicales *Les voisins de LeMonde* et *Les voisins d'En Face* (Bourigault 2002), à la base de textes *Frantext* (Martin, 1994), à l'interface *Weblex* (Heiden, 2004) ou à celle donnant accès aux corpus bilingues alignés du parlement européen *EuroParl* (Koehn, 2005). Citons pour finir trois sources très souvent exploitées dans les travaux de morphologie des langues anglo-saxonnes :

- le *British National Corpus*, base textuelle de l'anglais comportant 90% d'écrit et 10% de retranscriptions de l'oral de la fin du 20e siècle, totalisant environ 100 millions d'occurrences,

- le réseau lexical *Wordnet* (Fellbaum, 1998, 2005) dont une émanation libre de droit, *WOLF*, est en train de voir le jour pour le français (Sagot & Darja, 2008),

- la base de données *Celex* (Baayen et alii, 1995), élaborée à partir de dictionnaires, documents littéraires et journalistiques de l'allemand, l'anglais et le néerlandais.

Le plus souvent, pour leurs études consacrées au français, les morphologues se servent de deux types de corpus de textes : les archives de journaux (notamment les articles de *Le Monde*) et les œuvres littéraires (en particulier celles que recense la base de données textuelles *Frantext*). Dans un cas comme dans l'autre, ce choix résulte du niveau soutenu du registre de langue. En ce qui concerne le consensus autour de *Le Monde*, il relève de la conviction partagée (à tort ou à raison) selon laquelle le contenu de ce journal reflète au mieux la langue générale.

3.3 Usage et données de la Toile

Avant d'examiner en détail ce qu'exister veut dire pour un mot trouvé dans un corpus numérisé et ce que cela implique en matière de recherche en morphologie, nous consacrons cette section à la présentation de la Toile, en tant que source exclusive de données. En effet, si le morphologue a souvent recours au contenu d'Internet, tout ce qu'on peut y trouver n'est pas nécessairement qualifiable de ressource propre. Comme chacun sait, la Toile dite visible[6] (c'est-à-dire la partie accessible d'Internet indexée par les moteurs de recherche commerciaux) archive et duplique, entre autres, des collections de textes, d'images, de vidéos ou de sons, produites à l'origine pour être consultées ou diffusées via un autre support (papier, CD, DVD, CD-ROM, etc.). Parfois, un « mot qui existe sur Internet » est par conséquent une forme que l'utilisateur aurait pu trouver sur un autre médium, mais qu'il est plus facile et rapide de retrouver sur la Toile, qui fonctionne alors comme une immense bibliothèque, ou cyberthèque, régulièrement mise à jour.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

À titre d'exemple, et sans la moindre prétention à l'exhaustivité, nous pouvons citer, pour le français, différents portails :

- le *CNRTL*, donnant accès à des dictionnaires et lexiques du français,

- le *WebLettres*, qui répertorie les sites contenant l'intégralité des textes d'œuvres littéraires libres de droits,

- *CisMef*, qui permet d'accéder aux comptes rendus hospitaliers, articles et cours de médecine, bases de données terminologiques et autres sites de documents liés à la spécialité biomédicale.

Il est donc important de distinguer la Toile productrice originale de données de celle qui sert de dépositaire à des copies d'informations disponibles ailleurs. Dans le second cas, le morphologue y récupère des formes qu'il pourrait trouver sur d'autres médias : dans des dictionnaires, des bases de données, des archives de journaux, des corpus de textes littéraires ou des articles scientifiques.

Dans la suite de cet article, nous faisons le choix d'appeler « données de la Toile » les seules formes produites exclusivement sur et pour la Toile. Pour l'essentiel, il s'agit des blogs, des wikis, des sites interactifs et coopératifs et des forums de discussion, les autres espaces de production textuelle des internautes (intranets, email, tweets, messagerie instantanée) étant rarement indexés par les moteurs de recherche.

Le morphologue qui effectue sa recherche sur la Toile interroge un moteur de recherche au moyen de mots-clés, en se servant d'un navigateur. Les mots-clés correspondant aux index du moteur de recherche ramènent des pages ou URL. L'adresse de ces URL est souvent un bon moyen d'identifier les contenus originaux des duplicatas de corpus.

En dehors de cet accès aléatoire et manuel, le morphologue dispose d'un certain nombre d'outils pour effectuer des recherches plus fructueuses et efficaces en ligne, ou pour se constituer des corpus de textes à partir des données de la Toile. Il peut également disposer de ressources numériques échantillonnant le contenu de la Toile, par la taille et la variété des contenus. Ces trois moyens d'accès aux données en lignes sont brièvement présentés ici.

3.3.1 Outils d'interrogation automatique de la Toile

Les outils d'interrogation automatique sont conçus autour d'interfaces d'application (API) spécifiques aux moteurs de recherche commerciaux. Ces *Web Search APIs* servent à générer des robots qui soumettent des requêtes au moteur à la place de l'utilisateur, et qui sont capables de classer les résultats des requêtes en fonction d'un certain nombre de paramètres (nombre de pages, date de création de celles-ci, fragment du contexte de chaque requête effectuée). Certaines de ces applications ont été conçues spécifiquement pour répondre aux besoins exprimés par les morphologues d'accéder de façon automatique aux données lexicales de la Toile, afin d'y déceler les créations permettant de valider des hypothèses (*Walim*, cf. Namer, 2003) voire d'en ébaucher de nouvelles (*Webaffix*, cf. Hathout & Tanguy, 2005), en flexion et en construction. D'autres dispositifs, comme l'interface *WebCorp* (et son émanation plus récente *WebCorpLive*, cf. Renouf et alii, 2007), offrent au linguiste la possibilité d'interroger indifféremment plusieurs moteurs de recherche au moyen d'une seule et même requête, et organisent l'affichage des séquences de textes comportant la forme requise par l'utilisateur de manière à en optimiser l'étude. Les résultats obtenus par ces outils, qui ne demandent pas d'habileté particulière en programmation, sont limités par le type de requête effectuée : par exemple, *WaliM* a besoin d'une liste de formes candidates comme sources de requêtes. Les outils sont eux-mêmes tributaires des changements de politique imposés par les moteurs de recherche commerciaux : par exemple, *Webaffix* a dû renoncer à l'utilisation de motifs (expressions régulières) lors de l'interrogation du moteur *Altavista*. Enfin, on notera que les robots d'interrogation ne contrôlent pas l'origine des pages ramenées : ce mode de fonctionnement garantit par conséquent une variété diastratique et diatopique des contenus récupérés. En revanche, il impose une vigilance dans l'exploration de ces données, dont la validité est souvent mise en

cause (cf. les nombreuses critiques à l'encontre de la Toile en tant que réservoir de corpus, par exemple Lüdeling et alii, 2007).

3.3.2 Outils de constitution de corpus à partir de la Toile

Les programmes de constitution de corpus issus de la Toile sont construits à partir de *crawlers* (ou robots d'indexation), qui explorent systématiquement la Toile et téléchargent, à partir d'un ensemble d'URL racines, les contenus des pages qui en constituent l'arborescence, suivant des consignes limitant la profondeur de l'exploration, spécifiant le type de contenu recherché, évitant les doublons, etc. Contrairement aux robots ci-dessus, la conception même de ces applications impose le contrôle des ressources linguistiques extraites de la Toile. Ainsi, *Glossanet* (Fairon et alii, 2008) permet à l'internaute de se constituer un corpus à partir des archives publiées en ligne par les quotidiens ; le système *BootCat* (Baroni & Bernardini, 2004) se sert d'une liste de mots-clés (ou plus précisément de *seed words*) qui vont servir d'amorce aux indexages successifs des pages à ramener, en fonction de critères décidés par l'utilisateur, et exprimés par ces mots-clés. Par exemple, la liste initiale « "greenhouse", "global warming", "acid rain", "rainforest", "carbon emission", "Kyoto environment" » sert d'amorce à la constitution de corpus de documents en anglais portant sur l'environnement. Naturellement, les résultats de ces plateformes ou interfaces ne sont pas directement utilisables en morphologie, contrairement aux sorties des robots d'interrogation de moteurs de recherche. En effet, une fois constitué son corpus de textes bruts à partir des documents de la Toile, le morphologue doit encore être capable de le rendre exploitable, en le soumettant à divers programmes d'apprêt, selon l'objectif qu'il cherche à atteindre : segmenteur, étiqueteur grammatical, analyseur morpho-flexionnel, concordancier, fréquenceur, etc.

Les robots d'interrogation d'Internet, dont les résultats dépendent de la finalité visée, et les constructeurs de corpus de textes depuis la Toile, qui imposent, eux, des recherches guidées par les données, se partagent les avantages et les inconvénients. Contrairement aux seconds, les premiers ne demandent aucune compétence en informatique, et les données recueillies permettent au morphologue de procéder à des études de variation ; les constructeurs de corpus, en revanche, leur garantissent une homogénéité textuelle, et leur confèrent ainsi les caractéristiques proches de celles des corpus contrôlés. Un troisième type de ressources issues de la Toile, qui convient aux usagers linguistes non-informaticiens, est présenté ci-dessous.

3.3.3 Ressources textuelles issues de la Toile

De plus en plus de ressources textuelles simulant le contenu de la Toile sont mises à disposition des internautes dans le cadre de différents projets, accompagnées d'interfaces qui en facilitent l'accès et l'exploitation. Parmi les corpus proposés, citons deux types de documents ayant des structures différentes, et donc d'usage différent en morphologie. Il est possible tout d'abord de se procurer l'ensemble des n-grams (c'est-à-dire des séquences comportant n mots) ayant servi d'index à Google jusqu'en 2006 : *Google Index 1T-5gram* (Brants & Franz, 2006). Cet ensemble de séquences comporte de un (n=1) à six (n=6) mots formes, et est disponible sous la forme de deux ressources distinctes : l'une contient un milliard d'occurrences pour l'anglais, l'autre cent millions d'occurrences pour dix autres langues européennes. Un tout autre type de données textuelles, également disponible pour une large gamme de langues présentes sur la Toile, fait son apparition depuis quelques temps. Ces ressources se veulent à la fois d'une fiabilité digne de celle des corpus numérisés contrôlés (§ 3.2) et d'une couverture comparable, qualitativement et quantitativement, à celle de la Toile. C'est pour répondre à ce double objectif que se sont développés les projets autour du consortium *WaCky* (Baroni et Bernardini eds, 2006). Des corpus sont construits grâce à l'outil *BootCat* (cf. § 3.3.2), à partir de documents en ligne dans un grand nombre de langues, et réunissant de gros volumes de textes (par exemple, *frWaC* comporte 1,6 milliard d'occurrences). Les données des corpus sont annotées au moyens d'informations catégorielles et morpho-flexionnelles produites par l'étiqueteur et lemmatiseur *Treetagger* (Schmid, 1994), ce qui facilite l'utilisation de la part de linguistes non spécialistes du TAL.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

4 Quels types de recherches basées sur les ressources ?

Les ressources auxquelles donnent accès les corpus numérisés, la Toile, ou la conjonction des deux ont permis aux morphologues de faire des découvertes que les données dictionnairiques seules n'auraient jamais pu rendre possibles. C'est par ce constat que Hathout et alii (2009) introduisent leur synthèse concernant l'apport des données numérisées en morphologie, sans toutefois distinguer corpus numérisés et données exclusivement trouvées sur la Toile. C'est ce à quoi se consacre cette dernière section, où nous comparons « corpus » et « Toile » suivant les types de résultats en morphologie que ces ressources peuvent faire émerger.

4.1 Ressources numérisées, contrôlées ou non

La réalisation de certaines recherches en morphologie requiert la collaboration des deux types de ressources informatisées, sans qu'il soit nécessaire de distinguer l'apport spécifique de chacune.

4.1.1 Rôle des contextes

La première d'entre elles concerne la présence de contextes, consubstantiels à la notion de données textuelles, et qui peuvent jouer différents rôles dans l'analyse de la forme étudiée : spécification du sens, désambiguïsation d'une forme inventée potentiellement polysémique, identification d'une volonté d'infraction, de démarquage.

Souvent (Hathout, Plénat, Tanguy, 2003 ; Lignon, à paraître), le contexte est le seul indice permettant de comprendre le sens d'un occasionnalisme (cf. 7, 8, 9). Mais il sert également à désambiguïser les formes sous-spécifiées. Par exemple, dans leur étude de 2003 portant sur la formation des adjectifs en *-able*, du français, N. Hathout, M. Plénat et L. Tanguy soulignent, grâce aux contextes, l'étendue des relations possibles que peut entretenir un nom avec le verbe-base de l'adjectif en *-able* qui le modifie : en particulier, quand la base est un verbe transitif, le nom s'interprète certes comme un objet direct (cf. 10a), mais peut également correspondre au destinataire (cf. 10b). Dans Namer (à paraître), les contextes servent à montrer la polysémie des verbes construits sur base adjectivale (cf. 11a) qui parfois sont à mettre en relation sémantique avec un groupe nominal ('salaire mensuel', en 11b). Enfin, le contexte sert à faire émerger les créations, qui, par jeu, sont sciemment transgressives (cf. 12) (Koehl, 2010), ou celles qui sont créées sous l'effet de rafales (cf. 13) (Dal et alii, 2004) :

(7) L'effet de ce mascara : des cils **millionisés**.

(8) « **L'innumérisme** est à la maîtrise des nombres, du raisonnement et du calcul ce qu'est l'illettrisme à la maîtrise de la langue », explique le ministère de l'éducation nationale.

(9) le chat, y s'est **auto-soixante-neuviser**... comment c'est souple les chats !

(10) a ... d'une bonne fin doivent être notées : la relation de Paul avec Dieu par Jésus-Christ est toujours chaude et personnelle, il demeure **enseignable** et soumis à ...

b **...** Vous progresserez étape par étape. Nous ne recherchons uniquement des personnes sérieuses, **enseignables** et décidées d'agir dès maintenant. ...

(11) a Pour **mensualiser** ses impôts, le plus pratique et rapide est d'utiliser le service de paiement.

b Si vous n'êtes pas dans ce cas-là, vous devez **mensualiser** votre assistante maternelle soit en année complète, soit en année incomplète.

(12) Un peu de '**lachitude'** et beaucoup de '**diplomaterie'** si c'est pour garder des copains.

(13) (ouf … enfin) cette fois encore nous démarrons dans les véhicules, **levage, douchage, mangeage** et enfin 30 minutes de **rangeage** et nous voilà.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

4.1.2 Néologismes, et disponibilité des patrons morphologiques

Pour, entre autres, Baayen & Lieber (1991 : 801-2) et Lieber (1992 : 1), seuls les patrons morphologiques permettant de former de nouveaux lexèmes, autrement dit les patrons disponibles en synchronie, intéressent la théorie. La difficulté, avec cette affirmation, est de déterminer cette disponibilité, car les lexèmes marqués comme néologiques par les dictionnaires, s'ils l'ont été au moment de leur création, ne le sont plus guère lorsqu'ils entrent dans le dictionnaire. C'est le cas des exemples sous (14) empruntés au *TLF,* dont la néologie remonte, pour certains, au 19e siècle. Pour chacun d'entre eux, à la suite de l'extrait du *TLF*, nous indiquons entre crochets le nombre d'occurrences sur la Toile au 15 mars 2012 :

(14) **adjurateur** : « *Néol. d'aut.* Qui abjure. ● Au dernier moment, serré dans mes bras cinq amis (... Trois ont fait ce qu'ils ont pu pour nous tuer et le quatrième, un peintre, m'a lâché avec la plus ignoble candeur, (...). C'est mon calviniste abjurateur du 18 décembre 1897). L. BLOY, *Journal,* 6 janvier 1899, p. 289 » [12200]
absorptivité : « **HIST. — Néol**. du XIXe s. (1834, *cf.* étymol), terme sc. de phys. ou de chim. Aucune docum. en dehors de *Ac. Compl.* 1842, *DG, Lar. 20e, Lar. encyclop.* et *Pt* ROB. qui le signale comme apparaissant en 1839; selon *DG* il s'agit d'un néol. » [89400]
abstentionnite : « *POL., **néol**.* Tendance, considérée comme une maladie du corps social, à pratiquer l'abstention ● Toutefois, on peut admettre (...) que « l'abstentionnite » a moins sévi qu'au cours d'élections précédentes, ... *Combat,* 27 avr. 1959, p. 1, col. 2. » [262]
accumulatif : « ***Néol**.* Qui porte en soi un principe d'accumulation : ● N'a-t-il pas constamment lutté contre deux éternels instincts d'erreur opposés, d'une part contre l'instinct d'inertie accumulative d'une scolastique attardée qui s'attachait dans la tradition chrétienne à des éléments accidentels et périssables, d'autre part contre un instinct de dissociation dépensive représenté à cette époque par le mouvement averroïste et qui a donné ses fruits plus tard dans l'humanisme anthropocentrique des temps modernes? J. MARITAIN, *Humanisme intégral,* 1936, p. 222. » [12800]
accaparation : « ***Néol**. d'aut.* Action d'accaparer : ● C'était un grand jeune homme de vingt-huit ans environ, pâle sous la teinte brune dont le soleil avait doré son visage; une courte barbe noire, dure, serrée et frisée encadrait ses mâchoires saillantes et sa lèvre épaisse; son front large, très-développé par *les bosses* d'accaparation**,** se plissait de deux ou trois rides prématurées... M. DU CAMP, *Mémoires d'un suicidé,* introd., 1853, p. 9. » [14200]
achronie : « *PHILOS., **néol**. d'aut.* Caractère de ce qui se situe hors du temps, de la durée, du discontinu, et s'inscrit dans l'intemporel et le continu. *Achronie intérieure.* Sentiment d'absence de durée, d'intemporalité : ● Ce non-être de la durée, que mon vieillissement me rend sensible, fit aussi le drame d'Amiel; sa soif d'une « achronie *intérieure* », qui le libérerait de la division et de la dispersion intérieure, est profondément motivée par la condition humaine; son rêve d'*intemporalité* est vain, pour autant que c'est dans la *durée* qu'il nous faut créer et être fidèle; son mal du moins ne se ramène ni à une anomalie de caractère, ni à une erreur philosophique sur la signification du temps; car, à travers cette anomalie et cette erreur, Amiel a porté à son extrême point de lucidité la tristesse d'être durée. P. RICŒUR, *Philosophie de la volonté,* 1949, p. 427. » [3300]
automatisable : « ***Néol**., rare.* Susceptible d'être traité selon les procédés de l'automatique » [128000]
autopsiste (*s.v.* **autopsie**) : « **néol**. Médecin qui pratique des autopsies. „Le fameux médecin légal Tardieu, l'*autopsiste* de tous les sadismes de la société`` (E. et J. DE GONCOURT *Journal,* 1866, p. 264) » [187]

Ce qui a pu apparaître comme une profession de foi il y a vingt ans à un moment où les ressources numériques commençaient seulement à émerger est toutefois désormais aisément vérifiable grâce à l'essor qu'elles ont pris depuis, qu'il s'agisse de ce que nous avons appelé plus haut les corpus numériques ou de la Toile. Ainsi, une simple exploration de la Toile faite le 23 novembre 2011 au moyen des requêtes « permettre/permettez ce/le/un néologisme », « pardonner/pardonnez ce/le/un néologisme », « autorisez/autoriser ce/le/un néologisme », « accordez/accorder ce/le/un néologisme », « oser/ose

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

ce/le/un néologisme », restreintes aux pages : France, apparues sur la Toile il y a moins d'un an ramène 90 résultats utiles[7]. Ces résultats mettent en évidence qu'après la composition ou ce qui s'y apparente (17/90), les suffixations par *-iser* (8/90), *-esque* (*id*.), *-ien* (6/90) et *-itude* (5/90) sont régulièrement impliquées dans la formation de néologismes ou de ce que les scripteurs estiment comme tels. Les néologismes en *-isation* (5/90) sont également bien représentés. Le tableau 1 donne la totalité des résultats avec leur type et fournit quelques exemples (ou l'ensemble des néologismes quand ils n'excèdent pas deux) :

Type	Nbre	Exemples	Type	Nbre	Exemples
Composés	17	miso-islamie, politophobe	autoX	2	autocontextualiser, autosaisir
Xiser	8	extrême-droitiser, psychédéliser	Xance	2	bien-pensance, inversance
Xesque	8	sitcomesque, Woody Allenesque	Xiste	2	absentétiste, morchelliste (< Morchella)
Xien	6	fresnoisien, ursawaïen	Xment (adv)	2	irrécupérablement, marketinguement
Xisation	5	lyonnisation, sentimentalisation	Xaire	1	expositionnaire
Xitude	5	balaisitude, greekitude	X(i)al	1	massial
Convertis N/V	5	émuler, virguler	Xif	1	suspensif
(in)Xable	4	hakisable, incautionnable	Xifier	1	indignifier
Xisme	4	diplômisme, TROLLisme	Xite	1	testiculite
Xique	3	baltajique (<Baltaija), panthéonique	Xoïde	1	moboïde
Xissime	3	hispagnolissime, vulgairissime,	aX(N)	1	a-pertinence
Xité	3	constativité, moellosité	inX(N)	1	incommunication
Analogie	3	incommunication (excommunication), littératrice (calculatrice)			

Tableau 1. Types de néologismes construits relevés sur la Toile
(nov 2010-nov 2011 ; requêtes contraintes)

4.1.3 Non-existence d'une forme

Une autre tâche qui rend indispensable le recours à l'ensemble formé par les corpus et le contenu de la Toile concerne l'identification des formes lexicales inexistantes, qui en disent autant de la langue en usage que les formes existantes voire davantage. En effet, la légitimité d'une forme trouvée en ligne peut toujours être remise en cause, en raison des doutes concernant son origine. En revanche, l'absence d'une forme, quand elle est recherchée dans le réservoir documentaire que constitue l'ensemble formé par la Toile et les corpus numérisés, est un indice fort qu'au moment où la requête est soumise, aucun scripteur, quels que soient son niveau de langue ou son érudition, n'a eu ni l'envie ni le besoin de créer la forme recherchée. En effet, la source explorée est suffisamment représentative pour refléter la production langagière réelle, tant par le volume des données contenues que par leur variété linguistique[8].

Un manque est constaté quand l'interrogation sur le Web au moyen d'une forme servant de requête ne renvoie aucun résultat. Cette absence de résultat conduit à trois conclusions possibles : (i) c'est un mot devenu obsolète, (ii) la forme est impossible, (iii) son existence est inutile car elle ne répond à aucun besoin :

(i) Il est important en morphologie d'identifier les mots obsolètes, c'est-à-dire enregistrés dans les dictionnaires mais non usités. En effet, de telles formes sont susceptibles de fausser des résultats en alimentant de façon erronée des calculs de fréquence, ou en parasitant des systèmes de contraintes par des propriétés qui ne concernent plus les formes utilisées en synchronie. L'expérience, relatée dans Fradin et alii (2008), a ainsi montré qu'au moins 1,5% des entrées du *TLF* n'indexaient (presque) aucune page Web, les quelques résultats ramenés correspondant au contenu du *TLF* via le *CNRTL* ou des sites miroirs, ou renvoyant à des dictionnaires historiques (par exemple, *accoutreur, accouvi, acouiner*).

(ii) L'élaboration et la validation d'hypothèses morphologiques sont indissociables de la notion d'impossibilité, quand celle-ci s'observe sur une série de formes. Leur absence constitue la preuve du bien fondé d'une hypothèse morphologique donnée. Namer (2012) teste ainsi en ligne les contraintes portant sur la nature de la relation entre les constituants N et V des verbes composés obtenus par rétroformation, et qui expliquent pourquoi *cardiobattement* est attesté, contrairement à *cardiobattre*, qui correspond à un lexème ne satisfaisant pas ces contraintes.

Enfin, dans l'espace instable que constitue la Toile, une même recherche, répétée à plusieurs années d'intervalle afin de confirmer l'absence de formes présupposées mal formées, conforte l'hypothèse de leur impossibilité. Ainsi, Dal & Namer (2010a), reprenant une enquête menée dans (Dal & Namer, 2005), confirment l'absence systématique de noms de propriété ethnique en *-ianité* ou *-éanité* (les requêtes *estonianité, bohémianité* par exemple, ne ramènent aucun résultat), à partir d'adjectifs en *-ien* et en *-éen* dans lesquels ces finales sont précédées d'un /n/ (*estonien*) ou d'un /m/ (*bohémien*).

(iii) Signalons pour finir qu'il ne faut pas confondre absence et impossibilité ou obsolescence : une forme jamais employée en usage peut simplement signifier que le besoin de sa création ne s'est pas encore fait sentir. Ainsi, sur la Toile, les formes du verbe *débénaliser* (<*Benali*) ont une fréquence d'emploi quasi-nulle avant janvier 2011 (1 seule occurrence datant d'octobre 2009 de ce verbe sous sa forme infinitive). Toujours pour cette même forme, en novembre 2011, la fréquence est de 57, elle monte à 741 en mars 2012.

Dans de nombreuses études exploitant les données de la Toile, les auteurs expliquent l'absence d'une partie des données dont l'existence est pourtant prévue, car conforme à un modèle morphologique : on pourra se reporter, entre autres, à Lignon & Namer (2010), à propos des conditions de l'émergence des verbes néologiques de la forme *Xionner* convertis de noms *Xion*, où *Xion* est lui-même dérivé du verbe *X*, auquel *Xionner* fait concurrence.

4.1.4 Inventaire de variantes morphophonologiques

Un dernier questionnement où l'apport de la Toile peut être décisif est le repérage de variantes morphophonologiques, pour autant que la graphie en rende compte et qu'on soit assuré qu'il ne s'agisse pas d'écarts de graphie. L'inventaire de ces variantes est essentiel en morphophonologie ; en effet, alors que les dictionnaires, guidés par des considérations normatives, ont tendance à les éliminer, l'exploration de la Toile permet au contraire de les mettre au jour. Parmi les travaux témoignant de cet état de fait, citons (Plénat et alii, 2002), à propos de la variante *-este* de *-esque* après vélaire (par exemple *titaniqueste***)**, ou encore Lignon & Roché (2011), sur l'identification et l'analyse de formes présentant la variante /ɛ̃/ de /jɛ̃/ (par exemple, *chicagoen, sardouen*).

4.2 Ressources numérisées calibrées

Alors que l'ensemble des ressources numériques, indépendamment de leur origine, permet d'observer la disponibilité des patrons morphologiques en synchronie, la mesure de leur productivité demande au morphologue qu'il effectue ses calculs sur des corpus clos. Les mesures les plus utilisées pour cette tâche sont celles de R. Baayen (cf. notamment Baayen, 1992), éventuellement amendées (cf. Gaeta & Ricca, 2003). Pour R. Baayen, l'indice de productivité d'un patron se mesure toujours en corpus et correspond au quotient du nombre d'hapax résultant de l'application du patron analysé par le nombre d'occurrences

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

de mots-formes de lexèmes illustrant ce même patron, l'approximation étant que, plus on avance dans le corpus, plus la probabilité est forte qu'un hapax corresponde à un néologisme[9]. La plupart du temps, les ressources utilisées pour cette tâche de comptage sont des archives de journaux (cf. Baayen & Lieber, 1991 ; Baayen & Renouf, 1995 ; Gaeta & Ricca, 2006 ; Grabar et alii, 2006), mais il pourrait aussi s'agir de corpus réalisés à partir de la Toile (cf. § 3.3.2.).

4.3 Données extensives

Dans le champ de la morphologie, comme dans d'autres disciplines de la linguistique, on cherche, à partir de données langagières, à dégager des régularités, éventuellement exprimables sous forme de règles ou de contraintes, et qui conduisent à modéliser l'observable. En d'autres termes, on se demande : « Qu'est-ce qui existe, et comment l'existant a-t-il été créé ? ». Mais une autre question fondamentale, liée à l'intérêt grandissant pour des disciplines comme la psycholinguistique ou la sociolinguistique, porte sur les *raisons* ayant conduit à la production d'une formation langagière donnée. En morphologie, cet intérêt, qui se manifeste sous la forme de l'interrogation : « Pourquoi cette forme existe-t-elle ? », ne trouve de réponse que par la fouille des données extensives auxquelles seule la Toile peut donner accès. En effet, non seulement le Web accumule des documents indexés dont la quantité croît à une vitesse impressionnante[10], mais, de plus, certaines caractéristiques de ces documents sont parfois directement identifiables. Par exemple, l'URL qui permet d'y accéder peut comporter une extension (*.fr*, *.be* ...) qui indique son origine géographique (cf. 15) ce qui aide à repérer la nature diatopique de certaines constructions (Namer & Villoing, 2008) :

(15) *flashoir* (radars routiers, Belgique), *bavardoir* (forum de discussion, Québec), *shootoir* (local aménagé pour les drogués par les municipalités suisses, Suisse), *aiguisoir* (taille-crayon, Québec), *garoir* (parking, Zaïre)

L'adresse du site permet aussi parfois de classer les documents suivant leur type textuel. Les URLs comportant la séquence 'forum' dans leur nom de domaine, par exemple, hébergent un contenu qui met en jeu des interactions langagières appartenant à des registres variés.

Mais en dehors de ces indices, ce sont évidemment les contextes d'emploi des formes examinées qui servent à identifier la nature diastratique des variations observées. Le recours aux données de la Toile est en effet particulièrement adapté pour le repérage de créations apparemment hors norme, car ne répondant pas au besoin habituellement identifié en construction lexicale de nommer de nouveaux concepts. Suivant la classification proposée par Roché (2009), qui reprend et complète l'opposition entre dénotation et désignation proposée dans Kleiber (1984), Lignon & Namer (2010) distinguent, grâce aux données relevées en ligne, deux principaux facteurs favorisant les créations lexicales qui concurrencent des formes attestées :

- le besoin conceptuel supposé : une forme comble ce que son auteur pense être un vide lexical. Ainsi, certains scripteurs ont forgé *redditionner*, afin de reconstruire le verbe apparenté à *reddition*, la véritable base (*rendre*) de ce nom entretenant avec lui un lien formel difficilement identifiable ;
- le besoin énonciatif : le néologisme partage avec une forme existante la même valeur référentielle. Sa création permet à l'auteur de se faire reconnaître au sein d'un groupe, avec qui il partage un même ensemble de codes, ou de se démarquer par une infraction volontaire des normes ou par l'emploi de jeux de mots, comme dans 16 :

(16) Je suis aussi allé me *mur-des-lamentationner* vétu d'une kippa en carton qu'on aurait cru tout droit sorti de chez McDo. Plus de détails très prochainement. ...

Enfin, Koehl (2010) présente des exemples de noms désadjectivaux en *-erie* récoltés en ligne et répondant à un besoin que Roché (2009) qualifie de morphologique : *courtoiserie, bourgeoiserie, barbarerie*. Contrairement aux noms attestés *courtoisie, bourgeoisie* et *barbarie*, ces formes résultent d'un mode de formation facilement repérable (/əʁi/ est plus facilement identifié comme suffixe que /i/), et s'insèrent immédiatement dans une série morphologique connue.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

5 Conclusion

Notre investigation a montré que, si le dictionnaire a constitué la ressource par excellence pour les recherches en morphologie pendant plusieurs décennies, il n'est désormais plus possible d'y puiser ses seules données, dès lors qu'on se donne pour objectif d'étudier le système morphologique en synchronie. À cet égard, l'apparition de ressources sous forme numérique a permis d'opérer un changement radical dans les recherches menées dans le champ, et d'inscrire ces dernières dans le courant basé sur l'usage.

Au-delà de ce résultat général, six points sont à retenir de ce qui précède :

1) Modulo la réserve qu'on a faite (vérification à intervalles réguliers), et à condition qu'on veille à pondérer les fréquences renvoyées par les moteurs de recherche (à ce sujet, cf. Roché 2011 : 36-37), seule la Toile témoigne de l'*inexistant*, et donc peut se considérer le dépositaire (des variations) de l'*existant*. Les autres ressources ne permettent pas au morphologue de prendre en compte les mots *impossibles* dans son analyse.

2) Les *variations diastratiques et diatopiques* qu'atteste le contenu du Web permettent au morphologue de classer les lexèmes construits suivant des paramètres identifiables grâce au contexte ou à l'URL, comme : niveau de langue, registre, domaine, région… . En cela, cette ressource est supérieure aux corpus standardisés, où la *variation langagière* et les *écarts à la norme* sont, par définition, minimaux.

3) En dehors des données textuelles que l'on peut trouver sous un autre medium, le contenu du Web est assimilable à de *l'écrit (quasi-)spontané*. Les nombreuses créations lexicales répondent le plus souvent à des besoins d'ordre pragmatique et énonciatif ; elles peuvent donc être analysées en fonction des *raisons* pour lesquelles elles ont été forgées, ce qui n'est pas aussi systématique le cas dans les corpus journalistiques ou littéraires.

4) Les données du Web peuvent être réunies en corpus, ce qui autorise les calculs de *productivité* morphologique. À ce jour, les corpus normalisés constituent toutefois de meilleures ressources pour le morphologue voulant réaliser des mesures liées à la productivité.

5) Les mesures et les recherches qui requièrent des *corpus apprêtés* font des ressources textuelles standardisées les meilleurs candidats, même si de nombreuses initiatives de transformation du Web en corpus en font un concurrent sérieux.

6) Enfin, une fois son modèle théorique élaboré, à partir des données recueillies ou inexistantes en ligne, en fonction des contextes d'emploi et des mesures lexicométriques auxquelles lui donnent accès les corpus, le morphologue peut se tourner vers le contenu du dictionnaire, qui lui a fourni une première piste d'analyse. Il peut y tester la résistance du modèle dans le temps, en le confrontant aux données enregistrées de longue date. Le dictionnaire sert alors de *témoin* de la longévité d'une règle morphologique dont le Web, puis les corpus, ont démontré la vitalité et permis de calculer la productivité en synchronie.

Références bibliographiques

Baayen, R.H. (1992). Quantitative Aspects of Morphological Productivity. *Yearbook of Morphology 1991,* 109-149.

Baayen, R.H. & Rochelle, L. (1991). Productivity and English Derivation: a Corpus-based Study. *Linguistics 29-5*, 801-843.

Baayen, R.H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Baayen, R.H. & Renouf, A. (1996). Chronicling the Times. Productive Lexical Innovations in an English Newspaper. *Language 72*, 69-96.

Barlow, M. & Kemmer, S. eds (2000). *Usage-based models of language*. Stanford: CSLI.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. *LREC 2004*, Lisbonne, 1313-1316.

Baroni, M. & Bernardini, S. eds. (2006). *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT.

Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *TALN-2002*, Nancy, France, 75-84.

Brants, T. & Franz, A. (2006). *Web 1T 5-gram version 1*. Philadelphia: Linguistic Data Consortium.

Bybee, J. (1985). *Morphology: a study of the relation between meaning and form*. Amsterdam, Philadelphia: John Benjamins.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Corbin, D. (1997). Entre les mots possibles et les mots existants : les unités à faible probabilité d'actualisation. In Corbin, D., Fradin, B., Habert, B., Kerleroux, F. & Plénat, M. (éds), *Mots possibles et mots existants*, Villeneuve d'Ascq : Université Lille 3, 78-89.

Dal, G. (2003). Productivité morphologique : définitions et notions connexes. *Langue française 140*, 3-23.

Dal, G., Fradin, B., Grabar, N., Lignon, S., Namer, F., Plancq, C., Yvon, F. & Zweigenbaum, P. (2008). Quelques préalables linguistiques au calcul de la productivité des règles constructionnelles et premiers résultats. In Durand, J., Habert, B. & Laks, B. (éds), *Actes en ligne du 1e Congrès Mondial de Linguistique Française (CMLF-08),* Paris : ILF, 1587-1599.

Dal, G., Lignon, S., Namer, F. & Tanguy, L. (2004). Toile contre dictionnaires : analyse morphologique en corpus de noms déverbaux concurrents *Colloque International sur "Les noms déverbaux"*, Villeneuve d'Ascq : Université Lille 3.

Dal, G. & Namer, F. (2010a). French property nouns based on toponyms or ethnic adjectives: A case of base variation. In Rainer, F., Dressler, W. U., & Luschützky, H.C. (eds), *Variation and Change in Morphology*, Amsterdam: John Benjamins, 53-74.

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

Fairon, C., Macé, K. & Naets, H. (2008). GlossaNet 2: a linguistic search engine for RSS-based corpora. *Proceedings of LREC 2008. Workshop WAC4*, Marrakesh, 34-39.

Fellbaum, C. éd. (1998). *WordNet: An Electronic Database*. MIT Press.

Fellbaum, C. (2005). WordNet and wordnets. In Brown, K. et al. (eds), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670.

Fradin, B., Dal, G., Grabar, N., Namer, F., Lignon, S., Tribout, D. & Zweigenbaum, P. (2008). Remarques sur l'usage des corpus en morphologie, *Langages 171*, 34-59.

Gaeta, L. & Ricca D. (2003). Italian Prefixes and Productivity: a Quantitative Approach. *Acta Linguistica Hungarica 50*, 89-108.

Gaeta, L. & Ricca D. (2006). Productivity in Italian word formation: a variable-corpus approcah, *Linguistics 44-1*, 57-89.

Givón, T. (1979). *On understanding grammar*. New York: Academic Press.

Grabar, N., Dal, G., Fradin, B., Hathout, N., Lignon, S., Namer, F., Plancq, C., Tribout, D., Yvon, F. & Zweigenbaum, P. (2006). Productivité quantitative des suffixations par *-ité* et par *-Able* dans un corpus journalistique moderne. In Mertens, P., Fairon, C., Dister, A. & Watrin, P. (éds), *Verbum ex machina, Actes de la 13e conférence sur le traitement automatique des langues naturelles,* Louvain-la Neuve : Presses universitaires de Louvain, 167-177.

Greenberg, J. (1966). *Language universals: with special reference to feature hierarchies*. The Hague: Mouton.

Hathout, N., Plénat, M. & Tanguy, L. (2003). Enquête sur les dérivés en *-able*. In Hathout, N., Roché, M. & Serna, N. (éds), *Cahiers de Grammaire*, Toulouse : ERSS, 49-91.

Hathout, N. & Tanguy, L. (2005). Webaffix : une boîte à outils d'acquisition lexicale à partir du Web. *Revue Québecoise de Linguistique 32*, 61-84.

Hathout, N., Namer, F., Plénat, M. & Tanguy, L. (2009). La collecte et l'utilisation des données en morphologie. In Fradin, B., Kerleroux, F. & Plénat, M. (éds), *Aperçus de Morphologie du français*, Paris : Presses Universitaires de Vincennes, 267-287.

Heiden, S. (2004). Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex, *JATD2004*, 577-588.

Kleiber, G. (1984). Dénomination et relations dénominatives. *Langages 76*, 77-94.

Koehl, A. (2010). Nominalisation en *-erie* à partir d'adjectifs en français et construction du sens : de l'occurrence à la propriété. In *Décembrettes 7 – International Conference on Morphology*. Toulouse.

Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*, MT Summit.

Langacker, R. (1991). *Concept, image, and symbol: The cognitive basis of grammar*. New York: Mouton de Gruyter.

Langacker, R. (2000). A dynamic usage-based model. In Barlow, M. & Kemmer, S. (éds), *Usage-based models of language*, Stanford: CSLI, 1-63.

Lieber, R. (1992). *Deconstructing Morphology. Word Formation in Syntactic Theory*. Chicago and London: The University of Chicago Press.

Lignon, S. & Namer, F. (2010). Comment conversionner les V-ion ? ou la construction de V-ionner par conversion. In Neveu, F., Muni Toke, V., Klinger, T., Durand, J., Mondada, L. & Prévost, S, (éds), *Actes en ligne du 2e Congrès Mondial de Linguistique Française,* La Nouvelle Orléans : ILF*,* 1009-1028.

Lignon, S. & Roché, M. (2011). Entre histoire et morphophonologie, quelle distribution pour *-éen* vs *-ien* ? In Roché, M., Boyé, G., Hathout, N., Lignon, S. & Plénat, M. (éds), *Des unités morphologiques au lexique*, Paris : Hermès, 191-250. .

Lignon, S. (à paraître). *-iser* and *-ifier* suffixation in French: verify data to verize hypotheses. In Hathout, N., Montermini, F. & Tseng, J. (éds), *Selected Proceedings of the 7th Décembrettes: Morphology in Toulouse,* München : Lincom Europa.

Lüdeling, A., Evert, S. & Baroni, M. (2007). Using Web Data for Linguistic Purposes. In Hundt, M., Nesselhauf, N. & Biewer, C. (éds), *Corpus Linguistics and the Web*, Amsterdam/New York: Rodopi, 7-24.

Martin, É. (1994). *FRANTEXT : autour d'une base de données textuelles. Témoignages d'utilisateurs et voies nouvelles*. Paris : Didier Érudition.

Namer, F. (2003). WaliM : valider les unités morphologiquement complexes par le Web. In Fradin, B., Dal, G., Hathout, N., Kerleroux, F., Plénat, M. & Roché, M., *Silexicales 3 : les unités morphologiques*, Villeneuve d'Ascq : CEGES, 142-151.

Namer, F. & Villoing, F. (2008). Interpréter les noms déverbaux : quelle relation avec la structure argumentale du verbe de base ? Le cas des noms en –OIR du français. In Durand, J., Habert, B. & Laks, B. (éds), *Actes en ligne du 1e Congrès Mondial de Linguistique Française (CMLF-08),* Paris : ILF, 1539-1557.

Namer, F. (2012). Nominalisation et composition en français : d'où viennent les verbes composés ?. *Lexique 20*, 173-205.

Namer, F. (à paraître). Adjectival bases of French *-aliser* and *-ariser* verbs: syncretism or under-specification?. In Hathout, N., Montermini, F. & Tseng, J., (eds), *Selected Proceedings of the 7th Décembrettes: Morphology in Toulouse,* München: Lincom Europa.

Plénat, M, Lignon, S, Serna, N, & Tanguy, L. (2002). La conjecture de Pichon. *Corpus et recherches linguistiques 1*, 105-150.

Plénat, M. (2011). Enquête sur divers effets des contraintes dissimilatives en français. In Roché, M., Boyé, G., Hathout, N., Lignon, S. & Plénat, M. (éds), *Des unités morphologiques au lexique*, Paris : Hermès, 145-190.

Renouf, A., Kehoe, A. & Banerjee, J. (2007). WebCorp: an integrated system for Web text search. In Hundt, M., Nesselhauf, N. & Biewer, C. (éds), *Corpus Linguistics and the Web*, Amsterdam/New York: Rodopi, 47-67.

Roché, M. (2009). Pour une morphologie *lexicale*. *Mémoires de la Société de Linguistique de Paris 17*, 65-87.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100217

Roché, M. (2011), Quelle morphologie ?, In Roché, M., Boyé, G., Hathout, N., Lignon, S. & Plénat, M. (éds), *Des unités morphologiques au lexique*, Paris : Hermès, 15-39.

Sagot, B. & Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. *TALN 2008*, Avignon, France.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging using Decision Trees. *International Conference on New Methods in Language Processing*, Manchester, 44-49.

---

[1] http://www.u-cergy.fr/dictionnaires/histoire_dico/mvd._dictionnairique_dico.html

[2] L'ensemble des données sous (1-3b) est emprunté au *Trésor de la Langue Française* (*TLF*). Les exemples sous (1) résultent de la requête « arch » et « vx » réalisée dans la version informatisée du *TLF*. Lorsque le diacritique porte sur une acception seulement, cette dernière est précisée. La fréquence donnée pour les exemples sous (3b) est celle que leur confère le *TLF* : on considère ici arbitrairement qu'à partir de 250, la fréquence est haute. Par défaut, les exemples mentionnés constituent des entrées autonomes. Lorsque tel n'est pas le cas, l'entrée sous laquelle ils apparaissent est précisée.

[3] Les exemples sous (4) résultent d'une requête effectuée sur Google entre le 1e et le 11 novembre 2011, pages : France. Le nombre entre parenthèses correspond au nombre de résultats ramenés, non nettoyés. Pour les verbes, la requête a été faite sous leur forme infinitive ; les noms ont été entrés au singulier et au pluriel, et les adjectifs sous leur quatre formes. Le dictionnaire de référence par rapport auquel est ici évaluée la néologie est le *TLF*.

[4] Les créations (ou considérées comme telles par leur auteur) sous (5) ont été relevées sur le Web au moyen des requêtes « si ça se dit » / « si ça existe ». L'orthographe d'origine est respectée. Il est intéressant de remarquer que certaines, comme *partageur* ou *dératisation,* correspondent toutefois à des lexèmes enregistrés dans le *TLF*, pris de nouveau comme dictionnaire de référence.

[5] Rappelons qu'un n-gramme, où n est un nombre entier positif, est une séquence de n unités quelconques (mots, caractères, phonèmes, etc.).

[6] La partie visible de la Toile est estimée en 2011 à environ 50 milliards de pages indexées, dont la moitié au moins est accessible via le moteur de recherche Google (cf. http://www.worldwideWebsize.com/).

[7] i.e. résultats nettoyés mettant en jeu des lexèmes analysables comme construits au moyen d'un patron ou par analogie (dans ce cas, l'analogue, précisé entre parenthèses, est présent dans le contexte).

[8] Il ne s'agit que d'un indice fort et non d'une certitude, car, comme nous l'indique l'un de nos rapporteurs que nous remercions pour cette remarque, l'expérience montre qu'une forme morphologiquement impeccable mais de faible fréquence peut être attestée sur la Toile à un moment donné et ne plus l'être quelque temps après, pour des questions de maintenance de site, par exemple. Idéalement, l'absence doit être constatée à intervalles réguliers pour qu'une conclusion puisse être tirée.

[9] Pour une présentation des différentes mesures de productivité, cf. Dal (2003) ; pour une application au français des mesures de Baayen, cf. notamment Dal et alii (2008).

[10]Pour une estimation du rythme d'accroissement des pages indexées sur la Toile, cf. http://news.netcraft.com/archives/category/web-server-survey.