La ruée linguistique vers le Web

Ludovic Tanguy
CLLE-ERSS : CNRS et Université de Toulouse le Mirail

Octobre 2013

Résumé : Cet article propose un panorama des usages du Web en linguistique de corpus. À travers une présentation de différents travaux, il aborde les considérations méthodologiques et techniques, en mettant en avant les difficultés que rencontrent les linguistes face à cette source particulière de données langagières. En prenant exemple sur des travaux menés sur l'acquisition de données en morphologie extensive, je discute le statut des données, ainsi que de la position peu confortable dans laquelle les moteurs de recherche placent les chercheurs, et la façon dont ils doivent en permanence s'adapter à un matériau irremplaçable mais difficile d'accès.

Abstract : This paper presents an overview of the linguists' use of the Web as a corpus. Across several experiments, it exposes both methodological and technical aspects, while explaining the difficulties encountered. Focusing on past work on extensive morphology, I discuss the particular status of this source of textual data. One important point is related to the difficulties posed by web search engines, and how we must constantly evolve our approach in order to continue using the Web as an elusive source of useful data.

Mots-clés : Corpus, Web, TAL, morphologie extensive

Keywords : Corpus, Web, Natural Language Processing, extensive morphology

Préambule

NOTE : Cet article est tiré de mon mémoire d'Habilitation à Diriger des Recherches en linguistique, soutenu en septembre 2012 à l'université de Toulouse le Mirail sous le titre « *Complexification des données et des techniques en linguistique : contributions du traitement automatique des langues aux solutions et aux problèmes* » (Tanguy, 2012). Ce contexte initial d'écriture explique (mais n'excuse pas) l'emploi massif de la première personne et la mise en avant de mes propres travaux.

L'arrivée du Web dans la linguistique (en attendant l'arrivée de la linguistique dans les modes d'accès au Web) a changé tout un ensemble d'habitudes de recours aux données langagières. Comme le disaient avec un enthousiasme qui fait plaisir à voir Adam Kilgarriff et Gregory Grefenstette (Kilgarriff et Grefenstette, 2003) : « *The corpus of the new millenium is the Web.* ». Il est devenu un nouvel objet d'étude en tant que source de nouveaux types de textes, un réservoir d'attestations et une manne pour les approches quantitatives du TAL. Chacun y trouve des avantages sur les corpus classiques, que ce soit la masse, la variété des genres et des langues, l'évolution permanente, qui semblent globalement compenser les inconvénients qui lui valent de nombreux détracteurs (l'opacité du contenu et son hétérogénéité, les biais de ses modes d'accès, l'absence de représentativité, etc.).

Ma propre pratique du Web comme corpus remonte à plus de 10 ans, et a concerné principalement l'acquisition de ressources lexicales pour la morphologie extensive, qui a été l'occasion d'une collaboration avec les collègues morphologues de l'ERSS (Marc Plénat, Nabil Hathout, Michel Roché, Gilles Boyé) et d'ailleurs (Fiammetta Namer et Stéphanie Lignon de l'ATILF, Georgette Dal de STL). Bien qu'il ne s'agisse pas d'une utilisation *main stream* au sein de la communauté du « Web comme corpus », cela m'a permis de me confronter à l'ensemble de ses problématiques, et d'en suivre activement l'évolution.

Je commencerai par un panorama des usages du Web en linguistique de corpus et en TAL, puis je détaillerai les principaux aspects méthodologiques de l'accès à cette ressource. Je présenterai ensuite de façon synthétique les travaux d'acquisition lexicale avant de proposer quelques pistes pour continuer à explorer ce matériau.

1 Une courte histoire du Web : évolution des modes d'accès et des pratiques

Si le Web est un objet encore très jeune (il vient juste de fêter ses vingt ans), son histoire est déjà très remplie. Son apparition et son développement ont bouleversé un grand nombre d'activités humaines, et les tartes à la crème ne manquent pas pour décrire cet état de fait, des autoroutes de l'information des années 1990 aux réseaux sociaux omniprésents de la fin des années 2000. Je vais tenter ici de retracer l'impact qu'il a eu sur une grande partie des travaux en linguistique de corpus et en TAL, et de montrer comment ceux-ci ont dû également évoluer rapidement pour s'adapter à un objet sujet à de très nombreux changements.

1.1 Vue chronologique des usages du Web en linguistique

Je vais commencer ce tour d'horizon par une frise chronologique qui servira de guide à cet article. J'ai donc tenté de retracer dans la figure 1 les principaux moments de son histoire, à la fois en tant qu'objet indépendant, mais aussi dans les activités de recherche en linguistique et en TAL

Cette frise met surtout le point sur les aspects techniques et institutionnels de cette évolution, en marquant certains événements qui concernent l'évolution du Web lui-même et l'impact que ceux-ci ont eu sur la communauté scientifique.

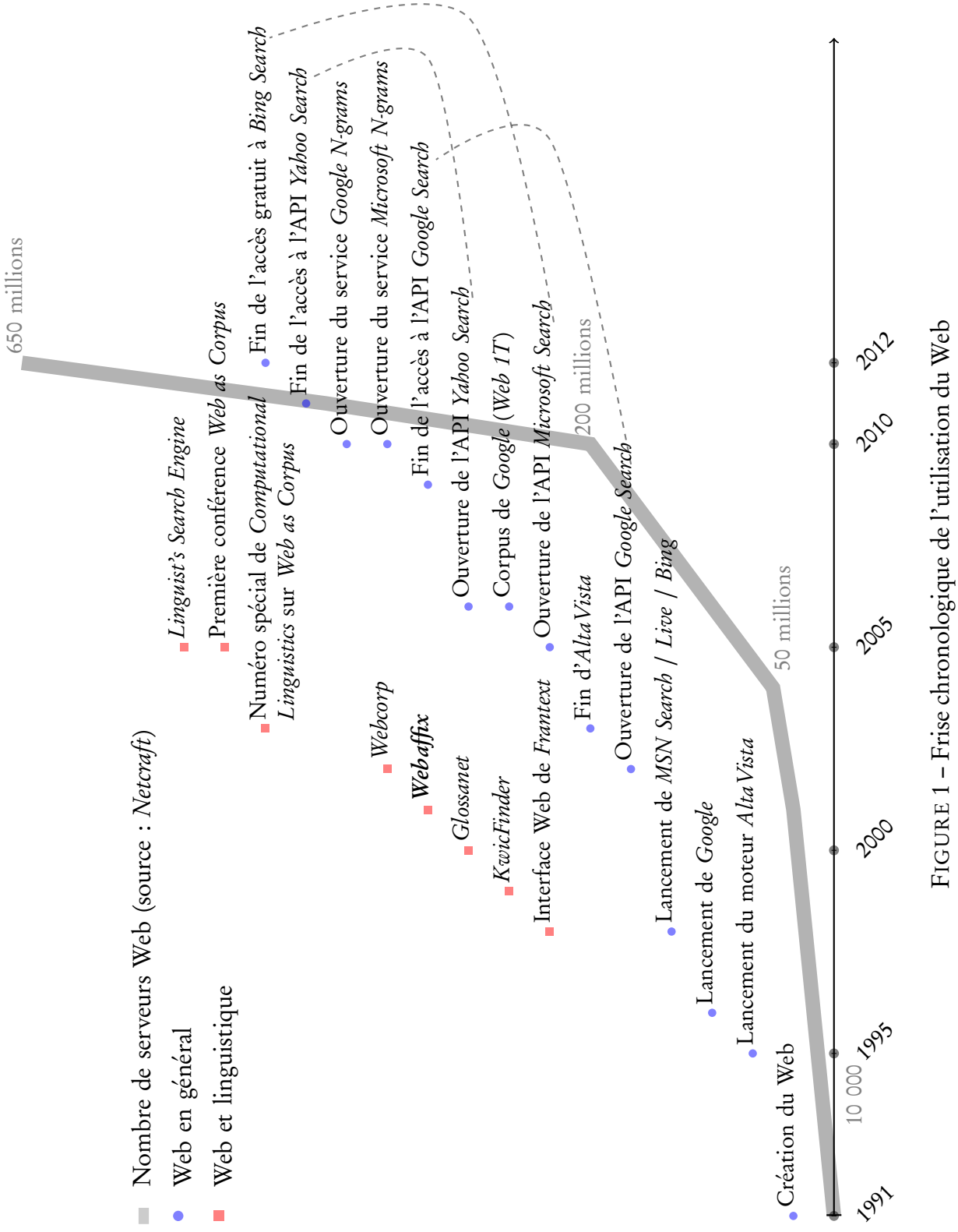

FIGURE 1 – Frise chronologique de l'utilisation du Web

- Sur le plan de l'**évolution générale du Web**, la frise commence bien entendu par la création du Web en 1991, et à partir de là son évolution rapide : le nombre approximatif de sites Web recensés a rapidement explosé comme l'indique la courbe. Les indications viennent des enquêtes menées par la société *Netcraft* [1] ; le nombre de pages lui-même est impossible à estimer sérieusement. Les premières années du Web ont vu se développer un grand nombre de moteurs de recherche, dont seuls quelques-uns sont désormais actifs et utilisés.
- En termes de **pratiques pour les travaux en linguistique**, je n'ai pas développé la diffusion d'outils et de ressources accessibles via le Web (à part la base Frantext, qui a rapidement su profiter de ce mode d'accès pour être utilisé par de très nombreux chercheurs). Par contre, j'ai positionné sur la frise plusieurs outils permettant une exploitation linguistique du Web : ces outils détaillés dans la suite de l'article ont commencé à apparaître dès 1998 et se sont rapidement multiplié. Les questionnements sur l'usage du Web en linguistique et en TAL ont très vite été organisés par la communauté scientifique, à travers des numéros spéciaux de revues et la création de groupes d'intérêt et de conférences dédiées.
- J'insiste également dans la frise sur les **aspects techniques** de l'utilisation du Web, dont la rapide évolution a eu des conséquences importantes sur les activités scientifiques. La masse de données langagières disponible sur le Web est essentiellement accessible *via* les moteurs de recherche, si bien que ce sont leurs décisions techniques qui conditionnent les exploitations déployées en TAL et en linguistique outillée. On peut distinguer trois grandes phases : l'utilisation « sauvage » de ces moteurs par des programmes spécifiques, suivie d'une période de compromis entre l'enthousiasme des chercheurs et le coût (pour les moteurs) des interrogations massives, et enfin une période actuelle de retranchement des moteurs de recherche, qui privilégient dans le meilleur des cas la mise à disposition de la communauté scientifique de ressources statiques ou de produits dérivés, comme des bases de données de séquences de mots (n-grammes).

Si les premiers utilisateurs du Web comme source de données langagières sont les chercheurs en TAL, la communauté plus large des linguistes s'est rapidement intéressée à la question, et a pu bénéficier des avancées techniques produites par les premiers. Je commencerai néanmoins par le volet linguistique.

1.2 Usages du Web en linguistique : des attestations faciles d'accès aux doutes sur leur valeur

Il existe une gamme d'outils très simples, très pratiques et très faciles à installer, permettant d'avoir accès rapidement et sans apprendre de langage de requête abscons à une quantité très importante de textes au format numérique : les moteurs de recherche sur le Web. Il n'est désormais plus rare de trouver dans des exempliers de conférence en syntaxe ou en sémantique lexicale des exemples d'énoncés dont il est dit prosaïquement qu'ils « viennent de Google », comme ces extraits de plusieurs articles de linguistique récents qui adressent des probléma-

1. `http://www.netcraft.com/survey/`

tiques variées :

– La fragmentation syntaxique :

(34) Dans ce livre, il parle de deux papas et de leur enfant. Enfant, qui est victime de jugements injurieux de la part de ses camarades d'école [Google]

– Les euphémismes :

(1) I don't know, I must be a dolt because I can't seem to change the color. (Google)

– Les emplois de prépositions particulières :

(18b) Le village est coiffé par des champs dominant de près de 300 mètres. [Google]

On remarquera que l'origine exacte des exemples n'est pas évoquée, ni la nature des documents dont ils sont extraits, que ce soit comme ici une critique de livre, une discussion de forum ou un journal de voyage.

Toutefois, la méfiance vis-à-vis de la source de données persiste, et bien souvent à juste titre, pour preuve cet extrait :

> « *Quand on consulte un moteur de recherche comme Google, on observe que les deux clitiques sont en effet attestés dans des cas contraires à la norme, ce qui suggère que la confusion semble concerner non seulement des stades plus anciens du français, mais la langue actuelle. Il est évident que l'utilisation de Google comme corpus demande des précautions, dans la mesure où on ne connaît pas l'identité des utilisateurs, qui peuvent en outre laisser des fautes de frappe. Une recherche sur Frantext sur une période récente et sur un corpus oral soigneusement transcrit serait évidemment nécessaire afin de vérifier l'hypothèse [...].* »
> (Lamiroy et Charolles, 2010)

De fait, bien souvent ce type d'exemples concerne des emplois marginaux des structures étudiées, et ils sont utilisés pour démontrer la relativité des contraintes qui sont supposées les régir, grâce à des locuteurs qui s'expriment dans un média bien moins rigide et contrôlé que les articles de journaux ou les textes littéraires qui fournissent l'essentiel des attestations. En quelque sorte, le Web comme source d'énoncés joue alors le rôle des corpus d'oral spontané difficiles à trouver et souvent très limités en volume.

Des études spécifiques sur le Web (comme celle de Resnik *et al.* (2005)) montrent même son intérêt spécifique pour trouver des exemples massifs de structures syntaxiques jusqu'ici déclarées impossibles (on verra le cas également en morphologie dans cet article).

Un autre usage très classique des moteurs de recherche tel qu'il transparaît dans un grand nombre de publications de linguistique est l'utilisation des fréquences telles qu'indiquées par le nombre de pages renvoyées pour une requête. Ces valeurs sont utilisées pour indiquer que le phénomène visé est fréquent (« *plus de 10.000 occurrences selon Google* »), ou au contraire très rare (« *aucune occurrence, même sur Google* »), ou encore pour comparer deux emplois concurrents. De nombreux débats ont eu lieu sur la fiabilité des chiffres renvoyés par les moteurs de recherche. Le plus célèbre repose sur l'expérience menée par Jean Véronis [2] qui a mis au jour le manque total de cohérence de ces résultats pour les principaux moteurs de recherche. Quoiqu'il

2. Voir à ce sujet son blog sur `http://blog.veronis.fr/`

en soit ces valeurs sont, lorsqu'elles indiquent des variations très importantes, une estimation à peu près aussi fiable que celles fournies par un corpus « classique ». Pour plus d'exemples d'études de ce type, voir notamment (Lüdeling *et al.*, 2007) et (Hundt *et al.*, 2007).

Dans le même ordre d'idée, le Web peut également être utilisé pour des études comparatives, en ciblant des productions de différentes communautés linguistiques, comme le fait Wooldridge (2004) lorsqu'il compare le français de France et celui du Québec, opération rendue très simple par la configuration d'un moteur de recherche ou la spécification d'un suffixe de domaine particulier (`fr` versus `ca` par exemple).

Un autre type d'étude concerne cette fois le Web comme objet, et non plus comme outil ou simple source d'exemples diversifiés. De nombreux travaux visent à observer et caractériser les nouveaux modes d'expression induits par le développement rapide des moyens de communication. Que ce soit les discussions en ligne, les forums ou les blogs, ces données facilement accessibles permettent d'étudier le comportement langagier (plus ou moins) spontané de communautés, et renouvellent certaines questions en analyse linguistique à différents niveaux (voir par exemple Mourlhon-Dallies *et al.* (2004)). On y voit notamment des études sur les écarts à la norme dans ces modes d'expression, des analyses conversationnelles, ou encore des approches sociolinguistiques s'intéressant à la formation de communautés autour de ces lieux particuliers d'échange. Tous ces travaux semblent exprimer une certaine circonspection par rapport aux objets étudiés, et une question récurrente concerne l'existence ou non d'une réelle nouveauté des pratiques langagières.

Le même type de question est soulevée par les travaux qui cherchent à inventorier et à caractériser les *genres du Web*. Je citerai notamment ceux de Marina Santini (Santini, 2007) qui propose de voir dans ces nouveaux objets textuels une hybridation de genres existants, mais laisse la question ouverte, tout en approchant également cette problématique par le biais d'une classification automatique en genres des pages Web. Cette problématique des genres se décline également à des niveaux plus locaux, comme les travaux de Valette et Rastier (2006) dans le cadre du projet Princip qui visait la caractérisation linguistique et la détection automatique de sites Web racistes.

Il est donc rassurant de constater que la communauté scientifique semble aborder ces données avec dynamisme et circonspection, mais en évitant un enthousiasme aveugle face au manque de fiabilité des données. Les autres utilisateurs du Web comme source de données, les chercheurs en TAL et autres développeurs d'applications d'ingénierie linguistique font peut-être étalage de moins de doutes, tant sont nombreuses et apparemment sans limites les utilisations qui peuvent être faites de cette masse de données.

1.3 Usages du Web en TAL : apologie de la quantité et de la diversité

Le Web et le TAL vivent une histoire commune depuis l'apparition du premier, ne serait-ce que par la problématique de la recherche d'information et les besoins en ingénierie linguistique que les applications ont fait émerger. Je me concentrerai ici sur une petite partie des travaux, en ciblant ceux qui (comme nous l'avons fait à l'ERSS) cherchent à extraire du Web des données linguistiques exploitables.

Les premiers chercheurs de TAL à avoir su profiter du Web semblent être ceux qui s'intéressaient à l'ingénierie multilingue, que ce soit pour l'acquisition de lexiques de transfert ou pour alimenter des systèmes de traduction automatique statistique (Grefenstette, 1998; Resnik, 1999). Un des avantages du Web est en effet, en plus de la variété des langues qui y sont représentées (malgré l'hégémonie évidente de l'anglais), le grand nombre de sites traduits en plusieurs langues (qu'ils soient institutionnels ou commerciaux). Des indices explicites et fiables peuvent être utilisés pour repérer qu'une page Web contient la traduction d'une autre, et permettent de construire automatiquement un corpus parallèle.

Du côté des ressources monolingues, on peut citer la récolte automatisée et à grande échelle d'entités nommées (Jacquemin et Bush, 2000) et nos propres travaux sur l'extension de lexiques morphologiques et la découverte de néologismes (Hathout et Tanguy, 2002; Hathout et Tanguy, 2005; Hathout *et al.*, 2009).

Les exemples ne manquent pas non plus du point de vue plus *hard-core* du TAL, à savoir l'exploitation de grandes quantités d'informations de bas niveau extraites de données textuelles non enrichies. La plupart des besoins pour l'entraînement de systèmes statistiques de TAL se résument à des séquences courtes de mots (n-grammes) avec une fréquence associée. Les utilisations de ces données très frustes (appelées *modèles de langue*) recouvrent le traitement de bas niveau comme l'étiquetage morpho-syntaxique, l'analyse syntaxique, la désambiguïsation, mais aussi les calculs plus sémantiques sur des principes distributionnels (classification d'unités lexicales, repérage de relations). Le besoin principal de ces types d'approches est une quantité très importante de données, qui compense en partie la pauvreté des informations. Le Web est donc un choix logique pour accumuler ces grandes quantités de n-grammes. De plus, les compagnies qui construisent et gèrent les principaux moteurs de recherche Google (Brants et Franz, 2006), Microsoft (Wang *et al.*, 2010) et Yahoo ! ont récemment mis à disposition de la communauté scientifique de telles données, dont le succès a été immédiat. Voir notamment (Lin *et al.*, 2010) pour des exemples d'utilisation de ces données (ici de Google) et Keller et Lapata (2003) pour l'exploitation des n-grammes avant la distribution des données des moteurs.

Des approches plus fines peuvent concerner également l'interrogation du Web pour désambiguïser le rattachement prépositionnel en analyse syntaxique (Gala, 2003), pour vérifier la validité d'une dérivation morphologique (Namer, 2003) ou encore pour trouver des reformulations (Duclaye *et al.*, 2002).

Dans tous ces travaux, le Web est vu comme un réservoir indifférencié de textes à analyser, indépendamment des grandes questions sur leur nature. Cette caractéristique des approches ultra-massives des données en TAL se retrouve également dans le mouvement actuel en linguistique de corpus visant la constitution de corpus génériques de plus en plus volumineux ; il se trouve que les corpus actuels les plus volumineux (et de loin) sont issus du Web.

La remise en cause de la nature et de la qualité des données y est nettement moins présente que pour les usages directs par des linguistes (qui, eux, regardent bien évidemment les données de plus près). Quelques exceptions sont à noter toutefois, notamment lorsque certaines expériences mettent au jour des résultats aberrants. Les données fournies par Google notamment, ont causé quelques émois dans la communauté, lorsque Hal Daumé [3] a découvert que les sé-

3. http://nlpers.blogspot.com/2010/02/google-5gram-corpus-has-unreasonable.html

quences de cinq mots les plus fréquentes correspondants au schéma *the X Ved the Y* étaient dans l'ordre (et avec des fréquences de plusieurs dizaines de milliers d'occurrences) :

the surveyor observed the use
the rivals shattered the farm
the link entitled the names
the trolls ambushed the dwarfs
the dwarfs ambushed the trolls

Il semblerait que ces fréquences disproportionnées pour des séquences aussi improbables soient le fait des nombreuses pages Web de *spam* générées par milliers pour leurrer les moteurs de recherche et amener leurs utilisateurs vers des sites souvent peu recommandables. Comme on le verra plus loin, il s'agit d'un des nombreux problèmes du Web, mais qui peut avoir des conséquences si les données sont utilisées sans discernement.

Pour reprendre la notion de complexification des données et des techniques, il est clair au vu des quelques exemples de pratiques décrits dans cette première section que le Web est un lieu très significatif des changements que je cherche à dégager dans Tanguy (2012). Du point de vue de la linguistique empirique, les données accessibles via le Web sont certes nombreuses, mais surtout très difficiles à décrire et à exploiter, et posent un ensemble de questions plus épistémologiques sur le statut des données elles-mêmes. Le TAL a permis de développer un ensemble assez vaste (étant donnée la brièveté de la période) d'outils pour permettre leur exploitation : si ces outils ne sont pas eux-mêmes d'une grande complexité de conception ni d'utilisation, ils ont amplifié les problèmes liés à la nature et au contrôle des données manipulées, les rendant elles aussi encore plus difficiles à évaluer, et produisant des résultats qui posent un ensemble de problèmes dans leur validation.

2 *Web for corpus* versus *Web as corpus*

Avant de détailler les aspects plus techniques qui sous-tendent les travaux qui exploitent le Web, il convient d'en faire une première typologie, en plus de la distinction des objectifs entre linguistique et TAL. De Schryver (2002) propose de distinguer les approches du Web pour constituer un corpus (Web *for* corpus) de celles qui considèrent l'ensemble du Web accessible comme un seul gros corpus (Web *as* corpus). Ces deux façons d'aborder la question ont des implications méthodologiques importantes, même si dans certains cas on peut remplacer l'une par l'autre. Je commencerai par quelques précisions terminologiques sur la notion de corpus dans ces deux locutions.

2.1 Abus du terme « corpus »

De nombreux débats ont abordé la définition de ce qu'est un corpus en linguistique, et se sont tout naturellement adressés au cas particulier du Web. La tendance généralement exprimée est de refuser le statut de corpus au Web, pour plusieurs raisons. Sinclair (2004) est sans doute le plus radical (dans les citations qui suivent, c'est moi qui souligne) :

> « *The World Wide Web is* not a corpus*, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective.* »

Rundell (2000) est un peu moins sévère :

> « *The big question here is about the actual value of the web as a corpus. In fact, of course, it is* not a corpus at all *according to any of the standard definitions : what it is is* a huge ragbag of digital text*, whose content and balance are largely unknown. [...] So the first caveat is that the web should not be regarded as a representative sample of English (or any of its other languages), and cannot therefore be used as a basis for making reliable generalizations about linguistic behaviour.* »

Ces deux remarques sont à mettre en regard de la tradition anglaise de la constitution de corpus de référence (le BNC en étant l'étalon), construit comme représentatif d'un état de langue en veillant à l'équilibre quantitatif des différents genres de texte, en plus de la description précise de l'origine des textes.

Dans une autre tradition, mais tout aussi radical, Rastier (2005) met en avant la notion plus générale de critère linguistique pour pouvoir attribuer le statut de corpus à une collection de textes :

> « *De fait, tout regroupement de textes ne mérite pas le nom de corpus. Ainsi une banque textuelle peut regrouper des textes numériques de statuts divers : aucun critère linguistique ne permet cependant leur totalisation, sauf l'hypothèse que la langue leur confèrerait une unité a priori ; mais même organisée en base de données, une banque textuelle ne devient pas pour autant un corpus.*
>
> *Un hypertexte n'est pas non plus par nature un corpus : soit c'est l'équivalent numérique d'un codex dont les renvois internes sont des liens hypertexte ; soit c'est l'hypertexte indéfini et il se confond avec le web – qui n'est* pas un corpus *mais, un euphémisme s'impose,* une aire de stockage, voire une décharge publique. »

Il faut donc considérer le Web comme, au mieux, une simple collection de textes dont la nature, la taille et la distribution sont mal connues (voire inconnaissables). Par contre, dans les usages concrets qui en sont faits dans les travaux de linguistique descriptive et de TAL, il semblerait qu'il soit utilisé en tant que corpus, au sens d'un ensemble de productions langagières exploitées à des fins linguistiques.

L'abus de langage est donc de mise (même si regrettable) dans la plupart des travaux décrits ici, mais avec toutefois une variation importante d'une situation précise à une autre.

2.2 Constituer un corpus à partir du Web

L'utilisation du Web comme *source de corpus* implique que le chercheur va utiliser une collection de textes extraits du Web. Dans ce type de cas, l'usage du terme corpus est moins choquant, puisque le linguiste a la possibilité de définir lui-même, et sur des critères qu'il est en mesure d'expliciter, la constitution de la collection de documents.

Si cette activité peut très bien se faire à petite échelle, en glanant des pages Web (ou des sites entiers), elle peut également se faire de façon plus massive en faisant appel à des processus automatisés.

Différents outils ont en effet été proposés pour assister le repérage, le nettoyage et dans certains cas l'étiquetage de pages Web. Les outils proposés sont par exemple BootCat (Baroni et Bernardini, 2004; Sharoff, 2006) qui à partir d'une liste de mots-clés génère un ensemble de requêtes automatisées et les soumet à un moteur de recherche pour télécharger des pages (en identifiant de façon itérative de nouveaux termes de requête). A ce travail de repérage s'ajoutent des procédures de normalisation des formats (extraction du seul contenu textuel des pages, harmonisation des codages de caractères, etc.), et la possibilité d'effectuer un étiquetage automatique). Le fait de pouvoir préciser un ensemble de mots-clés de départ, et la procédure de *bootstrapping* pour étendre ceux-ci permet de cibler des corpus thématiques. De la même façon le *GrosMoteur* de Kim Gerdes [4] effectue quant à lui un parcours du Web (*crawling*) et intègre un outil d'interrogation des pages moissonnées avec une interface dédiée.

Ces efforts techniques importants sont pour la plupart dûs à la création d'un groupe d'intérêt de l'association ACL (*SIG-WAC, Special Interest Group on the Web as Corpus* [5], organisateur notamment de la conférence WAC). Le groupe Wacky [6] a également construit par ce type de méthodes des corpus génériques pour les principales langues occidentales et les a mis à la disposition de la communauté. On peut ainsi disposer gratuitement de corpus de plusieurs centaines de millions de mots (voire plus) prêts à l'emploi (Baroni *et al.*, 2009), pour peu que l'on dispose des moyens de stockage. Certains de ces corpus sont également interrogeables directement en ligne.

Dans ces cas-là, par contre, il est évident que la notion traditionnelle de corpus est encore plus malmenée : un moissonnage automatique ne permet guère de contrôler quoi que ce soit dans la sélection des textes, que celui-ci soit déclenché par le chercheur lui-même, ou bien qu'il se serve de masses de textes prêtes à l'emploi.

2.3 Corpus et outils prêts à l'emploi

Je me contenterai ici de signaler quelques outils qui proposent un accès direct à des corpus issus du Web.

Le plus complet de ces outils n'est malheureusement plus en activité, mais a représenté l'effort le plus abouti pour fournir un accès à des corpus issus du Web. Il s'agit du *Linguist's Search Engine* de Philip Resnik et Aaron Elkiss de l'université du Maryland [7] (Resnik *et al.*, 2005) qui proposait un accès direct avec une interface innovante à une importante collection de pages Web étiquetées et parsées.

Serge Sharoff et ses collègues de l'université de Leeds [8] proposent une série de corpus extraits du Web, étiquetés et lemmatisés, que l'on peut interroger avec l'outil CQP (Corpus Query

4. http://grosmoteur.elizia.net/
5. http://www.sigwac.org.uk/
6. http://wacky.sslmit.unibo.it/
7. http://lse.umiacs.umd.edu/
8. http://corpus.leeds.ac.uk/internet.html

Processor), développé par l'IMS de l'université de Stuttgard (Christ, 1994). Ils proposent des corpus dans plusieurs langues, dont le français.

Adam Kilgarriff propose à travers son Sketch Engine [9] l'interrogation de nombreux corpus issus du Web, en mettant l'accent sur des utilisations en lexicographie, dans la tradition britannique (Atkins et Rundell, 2008). On y trouve ainsi de nombreuses fonctionnalités, en plus de la recherche par patrons, comme l'examen des collocations et une analyse distributionnelle. Malheureusement, l'accès à ce service est payant.

Glossanet est un des services proposés depuis longtemps, et sa dernière version (Fairon *et al.*, 2008) vise spécifiquement des sources de données dynamiques du Web sous la forme de flux RSS. Par un système d'abonnement (gratuit), la dernière version de Glossanet [10] permet à un utilisateur de sélectionner (ou de définir) un ensemble de flux (dépêches d'agence de presse, journaux, blogs, etc.) et d'appliquer dynamiquement à leurs contenus une requête sous la forme d'un patron morphosyntaxique. Au final, ce service permet d'effectuer une sorte de veille linguistique en exploitant spécifiquement les aspects les plus dynamiques (et également les mieux organisés) du Web.

Une dernière source de données issues du Web peut également, dans certaines circonstances, être obtenue *via* un partenariat avec les opérateurs des moteurs de recherche eux-mêmes, qui sont bien entendu les mieux placés pour disposer de données, puisque leur activité consiste justement à les trouver et à les indexer. Si les acteurs majeurs comme Google et Yahoo ne proposent pas directement ces données (uniquement les n-grammes comme indiqué plus haut), nous avons notamment eu la chance de pouvoir collaborer avec la compagnie Exalead et disposer d'un échantillon important de pages Web issues de leur index.

L'autre façon d'aborder le Web (*as* corpus) est d'utiliser directement un moteur de recherche pour accéder aux données, et c'est ce dont je vais parler plus en détails dans les prochaines sections, puisque c'est le mode d'accès que j'ai privilégié pour mes travaux.

3 Utilisation des moteurs de recherche : un passage obligé

Les moteurs de recherche sont des outils très pratiques pour les interrogations à la volée et la recherche d'attestations, mais ils introduisent un biais dans l'accès aux données que constitue le Web dans son ensemble. Dans tous les cas, il s'agit d'un point de passage obligé de la quasi-totalité des exploitations du Web comme corpus ou source de corpus. Le TAL entretient donc des rapports compliqués avec ces outils, objets de nombreuses critiques et sources de frustrations, d'autant plus que les intérêts scientifiques se heurtent souvent à des principes industriels et commerciaux, et de plus en plus à une concurrence entre la recherche académique et les équipes de TAL dont disposent désormais les compagnies qui créent et exploitent ces moteurs (ce que le nombre important de transferts de chercheurs de la recherche publique vers ces structures traduit bien). Dans cette section un peu plus technique nous allons voir concrètement les modes d'utilisation de ces outils pour accéder au Web comme corpus

9. `http://www.sketchengine.co.uk/`
10. `http://glossa.fltr.ucl.ac.be/`

3.1 Utilisation directe : la *googleologie*

On l'a vu, l'utilisation directe d'un moteur de recherche pour rechercher des attestations d'une unité lexicale ou d'une structure particulière est désormais une opération très courante. Mais il est clair qu'un moteur de recherche sur le Web n'est pas un concordancier ni un système d'interrogation de corpus, et que ses fonctionnalités sont limitées pour plusieurs raisons :

- Ce sont des systèmes de recherche d'*information*, et leur objectif est d'accéder au contenu (à la *sémantique* des textes comme disent souvent les informaticiens travaillant dans ce domaine). Si bien que les fonctionnalités de ces moteurs pour faciliter cet accès sont souvent des obstacles lorsque l'on s'intéresse spécifiquement à la forme : impossibilité de prendre en compte les variations de casse ou la ponctuation, gestion automatique de la flexion, correction automatique d'erreurs, utilisation d'équivalences lexicales diverses, etc. Parfois même les termes de la requête sont purement et simplement absents des documents renvoyés, puisque d'autres mécanismes sont déployés pour associer un document à une requête (dont le fameux principe d'utilisation des termes dans les liens hypertextes qui pointent sur un document pour indexer celui-ci, au lieu de son contenu).
- Ce sont des systèmes qui doivent gérer une quantité impressionnante de données, et qui ont une exigence d'efficacité et de rapidité. Les modes d'accès complexes au matériau textuel sont des sources notoires de ralentissement du processus de recherche : une simple expression régulière plutôt qu'une forme exacte entraîne un temps de calcul largement supérieur. Il existe donc au final très peu de fonctionnalités en dehors de la recherche d'une séquence de formes précises.
- L'unité de données pour un moteur est une page Web, pas une phrase ni un segment de texte quelconque. Cela signifie (en plus du problème que les fréquences affichées ne peuvent donc correspondre à des occurrences) qu'il est difficile de contrôler la portée d'une requête, ou encore d'accéder directement au contexte recherché.

Cela n'empêche pas ces outils d'être tout à fait utilisables pour des recherches portant notamment sur des unités lexicales simples, ou pour des locutions. Lorsqu'il s'agit de séquences plus complexes, s'apparentant à des patrons lexico-syntaxiques, les choses deviennent plus compliquées. Mais il reste encore quelques possibilités, notamment par le biais de requêtes utilisant des jokers, comme :

```
"plus on * plus on"
```

qui permet de rechercher des séquences dans lesquelles l'étoile peut être une série de mots quelconques (d'une taille non contrôlable, mais généralement de 1 à 5 mots).

L'absence de fonctionnalités au-delà de cette possibilité de définir des séquences *à trous* peut entraîner l'emploi de diverses ruses, et généralement la multiplication des requêtes, c'est ce qui a notamment énervé Adam Kilgarriff dans (Kilgarriff, 2007) :

> « *Working with commercial search engines makes us develop workarounds. We become experts in the syntax and constraints of Google, Yahoo, Altavista etc. We become googleologists. The argument that the commercial search engines provide low-cost access to the web fades, as we realise how much of our time is devoted to working with and against the constraints that the search engine imposes.* »

Ces remarques pertinentes entraînent Kilgarriff et d'autres avec lui à proposer de se détourner de ces moteurs pour construire des corpus à partir du Web et à développer des outils d'interrogation comme ceux que la communauté a maintenant l'habitude d'utiliser (voir quelques exemples en section 2.3, page 10).

3.2 Services intermédiaires : les concordanciers du Web

Une autre alternative aux corpus statiques construits à partir du Web pour pallier la pauvreté linguistique des moteurs de recherche est représentée par plusieurs outils en ligne qui se proposent comme intermédiaires entre le chercheur et le moteur.

C'est le cas de Webcorp [11] (Kehoe et Renouf, 2002), le premier outil de ce type qui présente sous forme de concordances les résultats d'une requête à un moteur de recherche classique (au choix de l'utilisateur). De façon assez légère et rapide, il transmet la requête de l'utilisateur, et parcourt les documents renvoyés par le moteur pour extraire les contextes d'occurrence du ou des termes choisis, ainsi que la liste des cooccurrences. Bien qu'assez rudimentaire et ne proposant pas de nettes améliorations de la syntaxe de recherche, il constitue tout de même un progrès certain par rapport à l'interrogation directe des moteurs.

Le Web Concordancer (ou KwicFinder) d'Alan Fletcher [12] fonctionne exactement sur le même principe (Fletcher, 2006).

Ces deux outils exploitent la possibilité d'automatiser les requêtes transmises à un moteur de recherche, ce qui peut permettre d'autres types d'applications.

3.3 Accès automatisé aux moteurs de recherche : les *API*

Aux premiers temps de l'utilisation du Web comme corpus, les exploitations des moteurs de recherche se faisaient au travers d'outils adhoc qui jouaient en quelque sorte le rôle du navigateur Web qu'utilise un utilisateur normal. Pour une requête donnée (respectant la syntaxe du moteur visé), le programme devait construire l'URL d'interrogation correspondante, la transmettre au serveur du moteur de recherche, récupérer la réponse (sous la forme d'une page HTML) et l'analyser pour en extraire les résultats pertinents (nombre de pages, titres, adresses et si possible extraits des documents). Comme bien d'autres à cette époque (je parle des années 1999-2002), j'avais conçu de tels programmes pour différentes applications, notamment pour l'outil Webaffix dont je parlerai plus en détail dans la prochaine section. Le développement et le maintien de tels outils était très fastidieux, notamment parce que le moindre changement impromptu de la part du moteur (dans sa syntaxe d'interrogation, mais surtout dans sa façon de présenter les résultats) nécessitait une mise à jour de tout ou partie du système, il fallait donc très régulièrement en vérifier la stabilité.

L'intérêt de ce type d'outil était bien entendu de traiter de grandes quantités de requêtes, par exemple pour calculer la fréquence (ou une estimation de celle-ci, voir plus haut) d'un terme ou d'une expression, parfois à l'échelle d'un lexique entier. Dans ce cas, le nombre de requêtes pouvait bien entendu dépasser la centaine de milliers. Certains moteurs étaient plus circonspects

11. `http://www.webcorp.org.uk/`
12. `http://webascorpus.org/`

que d'autres par rapport à cette sollicitation massive. C'est Google qui le premier a purement et simplement interdit ce genre de pratiques en les détectant et en réagissant par une interdiction d'accès temporaire (je peux maintenant l'avouer, des années plus tard, si l'université de Toulouse 2 s'est vue privée de Google pendant quelques heures c'était de ma faute).

En contrepartie, les services de Google ont proposé un mode d'accès spécifique à leur moteur pour de telles utilisations sous la forme d'une API (un protocole informatique pour permettre la communication entre deux programmes). Ce service gratuit permettait, sur simple inscription, à la fois d'éviter les sanctions précédentes et d'accéder directement aux résultats sans avoir à fouiller dans la page de présentation destinée aux utilisateurs « normaux ». En contrepartie, le nombre d'interrogations était limité à 1000 requêtes par jour. Les autres moteurs de recherche concurrents, Yahoo ! puis Live (le moteur de Microsoft, désormais appelé Bing) lui ont emboîté le pas, avec des conditions similaires quoique généralement plus intéressantes en termes de nombres de requêtes autorisées.

Malheureusement, la tendance actuelle est à l'arrêt de ces modes d'accès, comme indiqué dans la frise de la figure 1. Google a rapidement suspendu ce service (d'abord en ne distribuant plus de codes d'accès, puis en le supprimant totalement), et c'est maintenant Yahoo ! qui vient de fermer ses portes, laissant la place libre au seul Bing qui propose désormais un accès payant (à partir de 5 000 requêtes par mois) .

Comme je l'ai évoqué plus haut, ce genre d'événements totalement indépendants de notre volonté met parfois fin à des efforts de recherche et de développement de plusieurs années. Les enjeux économiques qui entourent les moteurs de recherche sont désormais tels que la collaboration devient simplement impossible. Les fermetures de ces services mettent à mal l'ensemble des approches déclinées dans les sections précédentes, de la création d'un corpus à la volée jusqu'à l'interrogation indirecte et enrichie des moteurs.

Les alternatives sont difficiles à mettre en place : la création d'un moteur de recherche, ou plutôt d'un *crawler* capable de parcourir le Web pour en extraire le contenu textuel est un travail de très longue haleine, et le coût matériel de son fonctionnement est colossal, bien hors de portée des budgets académiques. Il est fort possible que l'âge d'or du Web comme corpus soit derrière nous. Si c'est le cas, peut-être constaterons-nous au final que l'intense activité tant en TAL qu'en linguistique aura surtout concerné une augmentation du volume et de la variété dans les données utilisées plus qu'un changement de paradigme.

4 Webaffix : exploiter le Web pour une morphologie extensive

Mon expérience personnelle de l'exploitation du Web comme corpus est légèrement en dehors des principaux efforts de la communauté tels que je les ai présentés dans les sections précédentes. Hormis quelques travaux consistant à construire des corpus ciblés (notamment pour un projet consistant à étudier l'emploi de termes recommandés dans différents types de sites Web (Rebeyrolle *et al.*, 2007), ou pour différents mémoires réalisés par des étudiants de master), l'essentiel de mes efforts ont porté sur l'exploitation du Web pour y rechercher des créations lexicales. Ces travaux ont beaucoup évolué car ils ont subi de plein fouet les évolu-

tions technico-culturo-économiques du Web et des moteurs. Dans le contexte le plus favorable (de 2001 à 2003), il m'a donc été possible (avec Nabil Hathout) de construire et de distribuer Webaffix, un outil complet et opérationnel qui a rendu de nombreux services à la (petite mais active) communauté des morphologues qui s'intéressent à la morphologie dérivationnelle et qui perçoivent l'intérêt d'utiliser des données volumineuses.

4.1 Objectifs et principes

L'idée initiale a germé lors d'une discussion avec Marc Plénat, et nous avons voulu voir s'il était possible de repérer automatiquement de nouveaux adjectifs dérivés en *-esque* pour compléter l'impressionnante collection qu'il avait déjà réunie au fil des années dans différents corpus (Plénat, 1997). M. Plénat avait déjà vu l'opportunité d'exploiter le Web, et passait de longues heures à tester l'existence de nouveaux dérivés dont il devait auparavant supposer l'existence. L'idée de Webaffix était de remplacer ses hypothèses par une méthode purement inductive.

Il est important de noter que dès le début les phénomènes visés dans ces travaux sont très rares : si la création lexicale (et notamment par suffixation) est un phénomène relativement courant à l'échelle de l'évolution de la langue, lorsque l'on aborde des corpus les fréquences observées sont très basses. Nos dernières estimations (Hathout *et al.*, 2009) pour l'ensemble des suffixes déverbaux de noms d'actions (des suffixes très productifs comme *-tion*, *-ment* et *-age*) est que moins d'une page Web sur 200 contient une nouvelle forme lexicale de ce type.

La première version de Webaffix exploitait une fonctionnalité unique des moteurs de recherche de l'époque : la possibilité d'utiliser des troncations dans les termes de la requête. Seuls deux moteurs proposaient ce type d'accès au contenu : Altavista et Northern Light. Cette fonctionnalité a malheureusement disparu des deux moteurs au fil de leurs changements de propriétaires, de logiciel et/ou de positionnement sur le marché. Il était en effet possible de taper comme requête une chaîne comme `xxx*esque` et de voir le moteur l'interpréter comme *tout mot commençant par* `xxx` *et se terminant par* `esque`. Il était toutefois nécessaire de préciser les trois premières lettres (ou les quatre premières pour Northern Light) afin de limiter la complexité du calcul. La solution était donc simple : il suffisait de générer l'ensemble des combinaisons de lettres envisageables à l'initiale, de décliner les requêtes correspondantes et de s'arranger pour ne pas prendre en considération les formes déjà connues (qu'elles soient des entrées d'un lexique générique ou les formes nouvelles déjà répertoriées) et le tour était joué. Ce n'était rien de bien complexe une fois qu'on avait réussi à automatiser l'interrogation du moteur, en utilisant un programme spécifique comme indiqué plus haut.

Le principe de cet outil a ensuite été étendu pour y ajouter un module d'analyse morphologique des dérivés rapportés, afin d'en identifier la base, et ainsi proposer en sortie des informations morphologiques plus complètes. C'est sur ce point qu'est intervenue la collaboration avec Nabil Hathout, puisqu'il travaillait déjà sur l'automatisation des processus d'analyse morphologique et disposait de méthodes opérationnelles et robustes pour cette tâche (Hathout, 2000).

La tâche principale nécessaire pour rendre cet outil utilisable concernait la lutte contre la quantité impressionnante de bruit généré par cette méthode.

4.2 Problématique du filtrage

On continue parfois à parler du Web comme *poubelle planétaire*, et certains linguistes ont utilisé ce terme pour exprimer leur doute quant à son statut de corpus. En tout cas il est clair qu'on y est confronté à un ensemble de textes dont la nature et la qualité ne se retrouvent dans aucun corpus constitué à partir d'autres sources. J'ai donc très rapidement dû mettre en place des procédures automatisées de filtrage pour traiter les différentes sources de bruit rencontrées dans les résultats bruts, c'est-à-dire en automatisant le téléchargement et l'analyse des documents pour en extraire l'unité lexicale et son contexte.

4.2.1 Sources d'erreurs communes

Les types d'erreurs suivants ont été identifiés et traités comme indiqué ci-dessous. Les choix parfois drastiques sont justifiés par une volonté de limiter le travail de dépouillement, et par le fait qu'étant donnée la masse, un mot légitime finira bien par y résister.

- **Absence de mot** correspondant au schéma. Entre le moment où une page est indexée par le moteur et celui de l'interrogation, plusieurs modifications ont pu entraîner l'impossibilité de trouver le moindre mot correspondant au schéma. La page a pu simplement disparaître ou son contenu a pu être totalement modifié par l'auteur. Certaines erreurs de segmentation en mots dans le processus d'analyse par le moteur ont pu également entraîner des erreurs (notamment lorsqu'une balise de formatage est utilisée en milieu de mot, par exemple pour des lettrines ou autres effets typographiques. Aucun traitement spécifique n'est à effectuer à ce stade.
- Le mot est en fait un **nom propre** (patronyme, toponyme, nom de marque, etc.). Pour éviter ces problèmes, seules les chaînes en minuscules sont retenues.
- **Erreur d'orthographe** ou de frappe. Cette source de bruit est une des plus importantes. L'éventail des types d'erreurs est très large, et une procédure de vérification automatique a été mise en place. Pour éviter de rejeter toute formation nouvelle (inconnue du lexique de référence utilisé par ce type d'outil), je me suis limité à la détection des erreurs suivantes :
  - les fautes d'accents, quelles qu'elles soient, et sans limite de nombre par mot, comme "*prêfèrable*" pour "*préférable*" ;
  - les dédoublements (ou pire) de lettres comme "*grottesque*" pour "*grotesque*" ou au contraire : "*décolage*" pour "*décollage*" ;
  - l'inversion de deux lettres consécutives comme "*obliagtion*" pour "*obligation*" ;
  - l'ajout ou la suppression d'une lettre comme "*adapatable*" pour "*adaptable*" ou bien "*abillage*" pour "*habillage*" ;
  - la modification d'une lettre comme "*mertion*" pour "*mention*".
- Vérification de la **segmentation des mots**. Il s'agit en fait d'une vérification orthographique en contexte destinée au traitement des mots collés ou mal découpés comme dans l'exemple ci-dessous où *avantageusesque* n'est pas un adjectif en *-esque* :

  > *les prestations obtenues sont moins* ***avantageusesque*** *celles dont bénéficie un salarié à revenu égal,*

  Dans ce cas, on rejette le mot s'il existe un découpage qui donne deux mots présents

dans le lexique de référence ("*avantageuses + que*") et qui a une fréquence sur Altavista supérieure à celle du mot suspect : "*avantageuses que*" est présent dans 785 pages alors que "*avantageusesque*" ne l'est que dans une seule.
Cette correction peut cependant être à l'origine de découpages abusifs comme dans le cas de l'adjectif "*lestable*" qui peut être découpé en "*les + table*" ; or il n'y a que 105 occurrences pour l'adjectif alors qu'il y a 257 occurrences de la séquence erronée "*les table*". En d'autres termes, certaines fautes de frappe et d'accord surviennent plus fréquemment que certains mots construits...
Le problème inverse se pose dans le cas de textes gardant des traces d'une mise en page préalable à leur formatage HTML, comme les césures dans l'exemple ci-dessous. Dans ce cas, la présence d'un tiret à gauche du candidat "*mentation*" permet de vérifier la pertinence du recollage sur les mêmes principes que précédemment.

> *créer une réserve d'eau pour l'ali- **mentation** en eau potable de la région...*

– **Contexte dans une autre langue**. Altavista, comme tous les moteurs de recherche généralistes sur le Web, effectue un diagnostic de langue sur les pages qui ne l'indiquent pas explicitement dans leurs en-têtes. Dans un premier temps, les requêtes générées par Webaffix indiquent qu'on limite la recherche aux pages rédigées en français, mais cela n'est pas suffisant et des erreurs subsistent. Le problème se pose d'autre part pour les pages multilingues. Altavista n'attribue en effet qu'une seule langue à chaque page Web, a priori en fonction du début du document ou de la langue majoritaire. En résultat, la forme candidate peut très bien apparaître dans un segment en anglais ou en espagnol au sein d'une page par ailleurs en français.
Au bout du compte, Webaffix vérifie systématiquement, dans une fenêtre de 100 caractères autour de chaque occurrence du mot-cible, qu'il n'y a pas plus d'un mot-outil emprunté aux autres langues romanes et germaniques (anglais, allemand, espagnol, italien). Les mots-outils ont été sélectionnés par leurs fréquences, en enlevant les cas de recouvrement avec le français. Par exemple, "*or*" n'appartient pas à l'antidictionnaire de l'anglais. Quelques problèmes résiduels demeurent, par exemple pour les segments trop courts comme ci-dessous, qui est un cas classique de citation d'un titre original :

> *il nous faut aller la trouver dans les pages de Sept jours pour expier (days of **atonement**) de WJ Williams.*

La méthode des mots-outils n'est pas non plus bien adaptée à la détection des langues trop proches, notamment l'ancien et le moyen français comme ici :

> *tant soit peu, diminue, Ny que ma foy descroisse **aulcunement**. Car ferme amour sans eulx est plus, que nue.*

– **Code informatique**. De nombreux contextes courants sur le Web sont les segments de code informatique, les URLs, les adresses mail, etc. qui peuvent contenir des chaînes de caractères correspondant au schéma recherché. La méthode de filtrage se fait simplement sur la base de certaines combinaisons de marques typographiques (notamment les slashes, les accolades, les soulignés...).

> *Method Summary (package private) void* ***actionaffichage****_détaillé() Méthode qui permet un affichage de...*
>
> *http ://www.****abacdepannage****.fr/*

A la suite de ces différents filtrages, les formes retenues présentent une précision autour de 40%. Le score est très variable entre les suffixes, tout comme varient énormément les sources de bruit (voir Tanguy et Hathout (2002); Hathout et Tanguy (2005) pour plus de détails). Les erreurs résiduelles concernent les cas les plus sévères des problèmes énumérés ci-dessus, des mots corrects mais ne correspondant pas au schéma dérivationnel visé, ou qui appartiennent à une autre catégorie grammaticale (adverbe au lieu de nom), mais aussi des situations plus complexes, qui correspondent cette fois au type de document rencontré plutôt qu'à la seule occurrence du mot visé. En voici les principales :

- **Textes générés automatiquement**. Nombre de textes rencontrés par cette méthode sont clairement issus d'un processus informatique et ne sont pas écrits par un scripteur humain. C'est bien entendu le cas de la traduction automatique, très couramment utilisée pour rendre multilingue à peu de frais un site Web. Dans certains cas cela entraîne le transfert direct d'un mot étranger dans un contexte considéré comme du français, mais également la formation de mots construits par des mécanismes apparemment intégrés à ces outils. Je prendrai comme exemple le mot *conservatricement* [13]. L'extrait suivant est sans contestation possible une traduction automatique (l'original contenait a priori l'adverbe anglais *conservatively*) :

  > *Ils peuvent prévoir une réponse d'approximativement 35% des gens, à qui vous envoyez les E-MAILS. Mais nous laisser extrêmement* ***conservatricement*** *est tout d'abord et suppose que vous avez une quote-part moyenne de discours de seulement 10%. Si vous envoyez vos E-MAILS à 100 personnes différentes, vous pouvez prévoir que vous atteigniez au moins 10 de ces personnes, cela fait exactement cela, que vous avez fait.*

  Comme on le voit clairement dans la dernière phrase, il s'agit d'une de ces *chain-letters* qui égayaient nos boîtes aux lettres avant les mornes spams actuels, et qui a atterri sur un forum. D'autres contextes de cet adverbe existent dans des textes plus honnêtes, toujours comme traduction.
- **Listes de mots et autres textes métalinguistiques.** Le Web regorge également de pages qui ne contiennent pas de texte rédigé, mais parfois des listes de mots n'ayant pas vocation à former des énoncés. Certaines pages sont ainsi conçues pour remplir la fonction d'*attrape-moteur*, de façon plus ou moins sophistiquée et attirer des internautes vers des sites plus ou moins recommandables. Des contextes comme celui ci-dessous étaient assez courants :

  > *sites pirate sex passwords sexe argent blondes* ***naviguage*** *download proche orient palestine israel syrie naviguer*

13. Merci à Gilles Boyé pour cet exemple, ce type d'adverbes étant étudié dans (Plénat et Boyé, 2012, à paraître).

Généralement les mots contenus dans ce genre de listes étaient empruntés à des textes « normaux », et dans ce cas le mot découvert (ici *naviguage*) était confirmé par d'autres occurrences (on trouve actuellement des centaines de contextes légitimes de ce dérivé, comme *Bon naviguage sur mon site !*). De plus, les systèmes d'indexation des moteurs de recherche semblent avoir fait beaucoup de progrès et évitent soigneusement ce type de document.

D'autres contextes amusants sont ceux où l'on tombe sur des travaux de linguistes, notamment lorsque ceux-ci déclarent un dérivé impossible :

> ***ridiculage*** *est morphologiquement impossible car on y reconnaît la base adjectivale ridicule (ridicule, ridiculiser, ...) et le suffixe -age, mais ce dernier doit se fixer à des bases verbales (laver -> lavage, couper -> coupage, ...) et non des bases adjectivales. (On pourrait par contre imaginer ridiculisage : le suffixe -is rend l'ajout de -age possible en transformant la base adjectivale en base verbale.)*

Bien entendu, le mot *ridiculage* est maintenant employé dans des dizaines de contextes tout à fait acceptables :

> *Tu vas illico poser une main courante au commisariat du coin pour tentative de* ***ridiculage*** *en public*
>
> *Pour t'éviter le* ***ridiculage*** *voire la ridiculitude, je suggère que tu parles de groupement*
>
> *Bah en même temps, j'ai un peu participé au* ***ridiculage*** *de Lmiara. On était d'ailleurs trois à se délecter de proses blairiennes*

– **Niveau de langue et compétence du scripteur.** Si des débats intéressants peuvent concerner l'acceptabilité des exemples précédents, nous avons toujours décidé de les considérer comme tout à fait légitimes, mêmes si certains effets contextuels sont à prendre en compte (voir plus loin). Par contre, dans certains cas il est clair que la circonspection est de mise, par exemple pour cette occurrence de *conservatricement* :

> *Pour résumer, je pense que Sohane n'est pas contente avec sa vie et cela est la raison principale, pourquoi je pense qu'elle doit se changer. D'après moi, des gens peuvent vivre* ***conservatricement****, s'ils sont heureux, mais Sohane a l'air d'être un peu jalouse de la vie de Djelila, sa liberté, ses amitié et son plaisir d'être en vie.*

Il s'agit ici d'un extrait d'un message de forum entre des lycéens allemands qui étudient le français (ici le message indique que c'est explicitement une demande de correction d'une rédaction : le titre est « *Korrektur s'il te plaît !* »). Pour l'histoire, un locuteur natif lui a proposé de remplacer *conservatricement* par *de façon conservatrice*

Tous ces exemples montrent bien la nécessité d'observer précisément tous les contextes d'apparition. Ce travail minutieux est toujours très riche d'enseignement, notamment dans les exemples ci-dessus pour confirmer l'utilisation du procédé de dérivation (construire simplement l'adverbe sur la forme féminine de l'adjectif), que ce soit par des processus automatiques ou par des apprenants.

4.2.2 Analyse morphologique des dérivés

En plus du simple repérage de nouvelles formes suffixées, Webaffix comportait également un module permettant de calculer et de tester la forme de base supposée des candidats ainsi obtenus. Pour ce faire, deux phases supplémentaires étaient mises en place. La première correspond à un travail de calcul prédictif de la forme de base à partir du dérivé, en utilisant la méthode que Nabil Hathout avait mise au point pour la construction de ressources morphologiques (Hathout, 2000). La dernière étape consistait en une vérification du lien entre le candidat dérivé et la base prédite en recherchant des cas de cooccurrence des deux formes lexicales, là encore sur le Web. Ce module fonctionnait comme indiqué dans le schéma suivant :

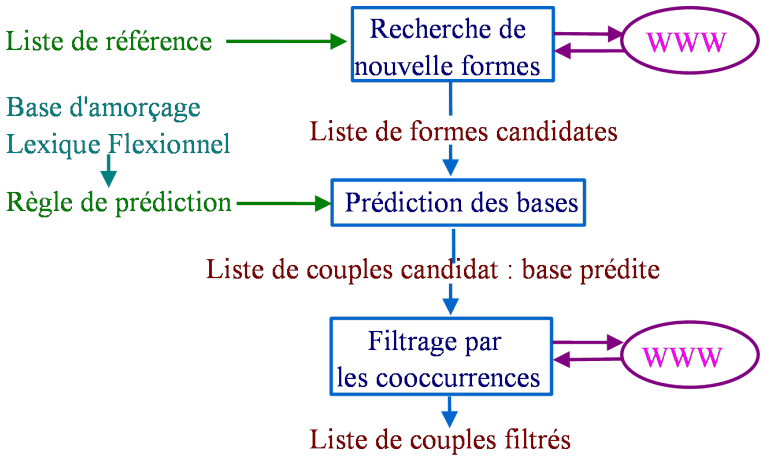

FIGURE 2 – Schéma de fonctionnement de Webaffix pour l'extraction de couples base-dérivé

Ainsi, l'analyse d'un nom déverbal comme *naviguage* arrivait à la prédiction du verbe *naviguer* comme base potentielle. Le couple *naviguage - naviguer* (en utilisant bien entendu des formes fléchies du verbe) était ensuite soumis comme requête, et les résultats vérifiés avec les mêmes procédures que pour le repérage initial. Le principe de validation d'un lien de dérivation par recherche de cooccurrence avait été découvert par Baayen et Neijt (1997), et mis en pratique immédiatement par Xu et Croft (1998) dans une application de recherche d'information.

Bien que constituant une contrainte exigeante, la procédure se révéla tout à fait efficace, et une quantité importante de bruit fut ainsi éliminée. Pour certains schémas dérivationnels (i.e. pour un suffixe comme *-age* et une catégorie de base donnée, par exemple le verbe) la précision finale pour les couples obtenus atteignait un très respectable 85%.

Ce type d'approche est, à mon avis, assez unique dans l'histoire du Web comme corpus, notamment par le soin apporté au filtrage des contextes. Les efforts généralement fournis dans cette mouvance concernaient, comme on l'a vu, la constitution d'un corpus le plus utilisable possible pour des études ultérieures, et ne pouvaient donc chercher à filtrer aussi précisément. De plus, le groupe Wacky s'est rapidement focalisé sur une tâche de nettoyage qui ne présentait pas d'intérêt pour notre approche. La campagne *Cleaneval* (Baroni *et al.*, 2008) visait principalement à rechercher des duplications dans les pages accumulées, et surtout à supprimer d'un

document complexe comme une page Web les segments comme les barres de navigation, les en-têtes et pieds de page, etc. On comprend bien l'intérêt de ces travaux, puisqu'ils sont nécessaires si l'on veut obtenir des fréquences fiables, ce qui reste une approche majoritaire dans la linguistique de corpus. Malheureusement, cela ne concerne que très peu nos besoins, focalisés sur des phénomènes rares et sans aucun appel à la notion de fréquence.

4.3 Principaux résultats obtenus

Pendant les quelques années de son existence, Webaffix a été fortement mis à contribution. Je récapitulerai ici les principales campagnes de récolte et les questions scientifiques abordées (un aperçu global est disponible dans Hathout *et al.* (2008)).

4.3.1 Suffixes *-esque* et *-este*

Il était logique d'utiliser Webaffix pour répondre en détail à la demande initiale qui a entraîné sa création. La base d'adjectifs en *-esque* de Marc Plénat a donc été complétée par le biais de cette méthode. Mais le fait le plus marquant de ce travail, et sans doute un des plus beaux travaux scientifiques auxquels j'ai pu participer est l'étude des dérivés en *-este*, présentée dans Plénat *et al.* (2002). En 1940, Edouard Pichon avait repéré dans un texte de Verlaine le dérivé *Silvio-pelliqueste* (construit sur Silvio Pellico), dont il avait supposé qu'il se substituait à l'attendu *silvio-pelliquesque* à cause de la dissonance entraînée par le redoublement du phonème /k/. Les recherches de Marc Plénat lui avaient permis de repérer 3 autres dérivés de ce type (et vérifiant l'hypothèse), malheureusement tous trois du même auteur, Frédéric Dard, dans ses romans de la série San-Antonio que M. Plénat avait repérée depuis longtemps comme une source abondante de phénomènes de ce type. Une fois Webaffix opérationnel, il a suffi d'une vingtaine d'heures pour repérer 7 autres dérivés du même type. De plus, en relançant la même recherche (cumulative) quelques mois plus tard, de nouveaux dérivés furent repérés grâce à l'évolution constante et exponentielle du Web. Le tableau 1 répertorie l'évolution de l'inventaire de ces formes. Comme on le voit, la conjecture énoncée par Pichon se voit totalement confirmée, sans contestation possible au vu de la quantité de dérivés répondant tout à fait à la règle qu'il avait induite d'un seul exemple. Nul doute qu'un nouvelle campagne ajouterait encore de nouveaux dérivés, mais l'objectif était atteint.

4.3.2 Suffixe *-able*

Une autre étude initiée par Marc Plénat et Nabil Hathout concernait cette fois le suffixe *-able* et est détaillée dans Hathout *et al.* (2004). L'objectif de cette étude était de confronter les théories existantes sur le fonctionnement de ce suffixe à une quantité de lexèmes plus importante que celle utilisée classiquement, à savoir les seuls dérivés répertoriés dans les dictionnaires. La phase de récolte automatisée par Webaffix a permis de former une liste de plus de 5 000 adjectifs alors que les dictionnaires les plus exhaustifs cumulés n'en contenaient que 1400. Ce travail de dépouillement très important (malgré les méthodes de filtrage mentionnées plus haut) a ainsi permis de revisiter la description de ce suffixe.

Source	Date	Dérivés	Nb
Pichon	1940	*silvio-pelliqueste*	1
Plénat	1997	*astraqueste, grandiloqueste, dingueste*	4
Webaffix (1re campagne)	2002	*dingueste, hageste, iagueste, langueste, marqueste, pragmatiqueste, punkeste, titaniqueste*	12
Webaffix (2e campagne) complété par Plénat	2004	*algueste, bangueste, big-bangueste, blagueste, blogueste, borgueste, bouledogueste, bouygueste, cirqueste, darkeste, dukeste, fiasqueste, fliqueste, gagueste, gangueste, jack-langueste, haddockeste, loqueste, luna-parkeste, mandrakeste, orqueste, pâqueste, pétanqueste, planckeste, ringueste, rockeste, stringueste, swingueste, tankeste, tongueste, turqueste, zingueste, zoukeste*	44

TABLE 1 – Évolution de la liste de dérivés en *-este*

Notamment, la recherche d'attestations sur le Web a permis de mettre au jour des utilisations plus libres de ce suffixe, appliqué à des bases verbales comme c'est normalement le cas, mais avec une signification bien différente de la glose traditionnelle de ces dérivés (un nom est *V-able* si on peut le *V*). Ainsi, pour reprendre l'exemple emblématique de *pêchable*, ont été trouvés comme noms recteurs (en plus des différents poissons) :

- des jours, saisons, des conditions météorologiques et autres périodes de temps (*pendant lesquelles* on peut pêcher) :

  Ne parlons pas du mois d'août... Impêchable !

- des rivières, étangs et autre plans d'eaux ou lieux (*dans lesquels* on peut pêcher) :

  3 km de rives pêchables, bien aménagées pour le lancer

- du fil, des cannes à pêches ou tout autre matériel (*avec lesquels* on peut pêcher) :

  je remarque après quelques lancers (je pêche généralement à 40 mètres en étang) que mon nylon se met à vriller et devient impêchable.

- des tailles de poisson (*que ceux-ci doivent atteindre* pour qu'on puisse les pêcher) :

  La sur-pêche et le non respect de la taille pêchable en Guadeloupe a entraîné une forte régression de la population.

De plus, de nombreux cas de constructions dénominales ont pu être repérés, élargissant grandement les séries isolées qui avaient pu être repérées. On voit donc que la plasticité du suffixe est bien plus grande que ce qu'un échantillon plus réduit de données permettait d'observer. Dans certains cas également cette enquête a permis de trouver des attestations déclarées préalablement impossibles.

4.3.3 Concurrence suffixale : le projet Wesconva

Dans le même esprit de vérification des hypothèses établies concernant le fonctionnement des suffixes, j'ai participé avec Georgette Dal, Nabil Hathout, Stéphanie Lignon, et Fiammetta Namer (qui l'a dirigé) à un projet (financé par l'ILF) nommé Wesconva (pour Web, Suffixation, Concurrence de déVerbaux d'Action) présenté dans Dal *et al.* (2004). Ce projet se proposait d'étudier la concurrence suffixale des déverbaux, en étudiant notamment le cas où un même verbe donne lieu à deux noms d'actions, l'un en *-age* et l'autre en *-ment*, et d'étudier les cas où l'un ou l'autre était présent dans un dictionnaire ou n'apparaissait au contraire que sur le Web.

La méthode utilisée à été légèrement différente, puisque nous sommes cette fois partis exclusivement des verbes répertoriés dans le TLFi, avons généré les dérivés possibles en utilisant la méthode WaliM de Namer (2003). L'approche de Fiammetta Namer est en fait hypothético-déductive et fonctionne en sens inverse de Webaffix : à partir d'une liste de bases et d'un processus de dérivation, des candidats dérivés sont générés et recherchés tels quels via des requêtes automatiques sur le Web. Nous avons donc décidé de mélanger les deux approches, et d'utiliser les fonctions de filtrage de Webaffix pour améliorer cette phase de vérification.

Au final, nous avons retenu 1 150 verbes présentant cette concurrence, et pour lesquels seul un des deux dérivés est répertorié dans un dictionnaire par exemple :
amocher : *amochage* (dict.) et *amochement* (Web)
estropier : *estropiement* (dict.) et *estropiage* (Web)

D'un point de vue quantitatif, le suffixe *-age* est majoritaire dans les dérivés absents des dictionnaires (65% des couples). D'un point de vue qualitatif, on peut en déduire soit qu'en les créant, le locuteur entend instituer une différence par rapport aux dérivés en *-ment* correspondants, s'il les connaît, soit, s'il ne les connaît pas, que le suffixe *-age* fonctionne tendanciellement comme suffixe par défaut.

Dans 30% des cas, le nouveau dérivé apparaît dans un domaine que ne couvre pas le dérivé du lexique conventionnel (par ex., *décagement* : /ornithologie/, /mine/ vs *décageage* : /agro-alimentaire/). Dans 70% des cas, le nouveau dérivé n'est donc pas motivé par une différence de domaine d'emploi (ex. *gravillonnage*, *gravillonnement* : /équipement/).

Du point de vue des théories préalablement posées pour expliquer la distinction entre les deux suffixes, nous avons voulu tester les propositions de Kelling (2003). Son hypothèse est que la concurrence *-age*/*-ment* à base verbale constante est passible d'une explication en termes de proto-rôles : selon elle, le suffixe *-age* se combinerait avec des bases verbales dont le premier argument est proto-agent, tandis que *-ment* serait sensible, lui, à sa proto-patience. Par exemple, *battage* (*battage de tapis*) suppose un sujet qui effectue volontairement l'action, contrairement à *battement* (*battement de cœur*).

Sur nos données, les contextes révélateurs de la (proto-)agentivité du dérivé ont montré que les distinctions instituées par C. Kelling pour expliquer la concurrence *-age*/*-ment* ne correspondent pas à un phénomène nettement repérable en contexte. En effet, si 65% des déverbaux en *-age* ont au moins un emploi de type agentif, cela vaut aussi pour 46% des déverbaux en *-ment*.

Bien que les conclusions soient moins marquantes que dans les précédentes études, ce travail

a montré que de tels phénomènes ne pouvaient plus être étudiés sans prendre en compte une plus grande variété de données. C'est notamment ce qu'a fait plus récemment Fabienne Martin dans (Martin, 2008), lorsqu'elle a étudié de nombreux dérivés repérés sur le Web pour proposer une distinction plus précise. F. Martin a d'ailleurs depuis utilisé des données obtenues par une adaptation de Webaffix dans (Martin, 2013).

4.3.4 Lexique Verbaction

Dans une optique plus TAL que linguistique, Webaffix a également été utilisé pour constituer des ressources lexicales à large couverture. En l'occurrence, Nabil Hathout et moi-même avons lancé plusieurs campagnes de récolte de noms déverbaux d'action suffixés en (*-ade*, *-age*, *-ance*, *-erie*, *-ement* et *-tion*) pour étendre le lexique Verbaction [14]. Ce lexique a été initialement conçu par Nabil Hathout à l'ATILF à partir des données extraites du TLFi et en utilisant la méthode de (Hathout *et al.*, 2002) et validées manuellement. Il contenait à ce stade 6 471 couples noms/verbes tels que le nom dénote l'action ou l'événement exprimé par le verbe. Par exemple, *élection/élire*. Après deux campagnes de collecte par Webaffix, en utilisant le module de prédiction et de vérification des bases (l'idée de cette technique est d'ailleurs issue de cet objectif précis), Verbaction contient actuellement 9 393 couples (donc une augmentation de 50% par rapport aux données initiales).

La couverture ainsi élargie permet de prendre en compte des couples correspondant à de nouveaux référents (*pacsage/pacser*), des termes techniques (*aquamarquage/aquamarquer*) ou des niveaux de langue (*baisage/baiser*) non couverts par la source d'origine.

Ce lexique est très utilisé pour un ensemble d'applications de TAL, comme l'analyse syntaxique ou l'extraction d'information.

4.4 Avatars de Webaffix dans l'adversité

Comme on l'a vu, les travaux d'exploitation du Web comme corpus sont soumis aux aléas des moteurs de recherche qui restent incontournables pour accéder aux données. Webaffix a donc subi comme d'autres un terrible coup lors du rachat du moteur Altavista par Yahoo ce qui a entraîné (au 1erAvril 2003) l'arrêt des requêtes par troncation : l'approche inductive utilisée jusque là ne pouvait donc plus être utilisée. Passée la période de deuil, nous avons envisagé deux types de réactions.

La première était de revoir à la baisse la couverture de la méthode, et de remplacer le premier niveau de Webaffix (voir figure 2) par une approche hypothético-déductive comme celle précédemment utilisée par Fiammetta Namer (Namer, 2003), c'est-à-dire partir de bases connues, construire des dérivés potentiels (en utilisant la technique de N. Hathout dans l'autre sens) et en interrogeant les moteurs de recherche avec ces formes complètes. Bien que rapidement opérationnelle (les deux autres niveaux restaient inchangés), cette méthode ne permettait absolument plus de répondre à des besoins comme l'extension de Verbaction, ni d'observer des créations simultanées de bases et de dérivés (comme par exemple le couple *wapiser/wapisable* construits tous deux sur l'acronyme *Wap*).

14. Téléchargeable sur `http://www.univ-tlse2.fr/erss/ressources/verbaction/`

L'autre solution envisagée était, comme le prônait avec l'enthousiasme qui les caractérisent Kilgarriff et Grefenstette (2003) :

> « *This suggests a better solution : Do it ourselves. Then the kinds of processing and querying would be designed explicitly to meet linguists' desiderata, without any conflict of interest or "poor relation" role. Large numbers of possibilities open up. All those processes of linguistic enrichment that have been applied with impressive effect to smaller corpora could be applied to the Web. We could parse the Web. Web searches could be specified in terms of lemmas, constituents (e.g., noun phrase), and grammatical relations rather than strings. The way would be open for further anatomizing of Web text types and domains. Thesauruses and lexicons could be developed directly from the Web. And all for a multiplicity of languages.* »

C'est donc surtout Franck Sajous, alors récemment arrivé à l'ERSS qui a décidé de créer un moteur dédié au moissonnage de créations lexicales, nommé Trifouillette [15]. En lieu et place d'un moteur de recherche, Trifouillette était un *crawler* qui parcourait le Web (en suivant les liens hypertextes), mais en n'indexant que les pages contenant des mots nouveaux, et en mettant en place un système d'interrogation, d'alerte et de validation dédié aux travaux menés dans l'équipe. Malgré les louables efforts et la compétence de F. Sajous, ce projet a malheureusement dû être abandonné, face à la complexité de l'opération de parcours du Web (et de la ruse dont il faut faire preuve pour échapper aux multiples pièges tendus par des sites Web peu conformes aux normes édictées), des ressources matérielles exigées (notamment en termes d'accès réseau) pour arriver à atteindre un rythme de moissonnage suffisant.

La dernière solution était donc d'utiliser des corpus tout faits, comme ceux fournis par les travaux mentionnés en 2.3, ou encore mis à disposition (en fonction de leur bonne volonté) par des moteurs de recherche. Dans Hathout *et al.* (2009) nous montrons ainsi comment nous avons pu relancer une dernière campagne d'extension de Verbaction en utilisant un corpus fourni par la société Exalead (que je remercie au passage). Si là encore la plupart des procédures développées pour Webaffix peuvent être recyclées à peu près telles quelles, les résultats obtenus sur un corpus statique ont du mal à justifier l'effort fait pour s'y adapter : une fois exploité le corpus n'est plus utilisable. Par contre, cela nous a permis de mesurer précisément le volume nécessaire à l'acquisition des données du type de celles présentées dans la section précédente : le volume est absolument décisif, et rien ne peut remplacer le Web pour y chercher des phénomènes rares.

5 Quelques pistes à explorer

Malgré les déconvenues récentes et l'actuel manque de moyens permettant de relancer de grandes campagnes d'acquisition de données, les nombreuses expériences menées lorsque cela était possible ont ouvert un ensemble de questions de recherche que je vais exposer ici.

C'est sans doute moins le travail mené pour effectuer les opérations de filtrage que les dépouillements (ou dépouillages) manuels des données plus ou moins brutes qui ont permis

15. `http://w3.erss.univ-tlse2.fr/membre/fsajous/trifouillette/`

d'identifier des phénomènes que j'estime intéressants à examiner.

Le premier point concerne la question du genre des pages Web, dont on a vu à travers notamment les travaux de Santini (2007) qu'elle était loin d'être résolue, puisque la diversité et la nature des genres du Web ne fait pas l'objet d'un consensus. Il est pourtant évident que des progrès dans cette direction pourraient bénéficier directement aux études des phénomènes cités dans cet article, et à l'exploitation du Web comme corpus en général.

Il est par exemple clair que certains genres du Web sont plus productifs que d'autres, et se concentrer sur ces documents ferait gagner en efficacité (on a vu que M. Plénat avait repéré la grande productivité de San Antonio, et les morphologues de l'ERSS ont depuis quelques années identifié des forums Web comme *doctissimo* ou *aufeminin.com* comme étant des mines sans fond de créations suffixales innovantes).

Certains genres sont également plus pertinents que d'autres en termes de stabilité des lexèmes qui y sont découverts, et certains contextes repérés à la volée sont vus avec moins de circonspection que d'autres par les linguistes qui les utilisent comme simples exemples. En ceci, nos travaux sur le repérage des néologismes se distinguent de ceux de Valette (2010) qui ont une visée lexicographique, et de ce fait tendent à rejeter certains genres du Web qui ne garantissent pas la moindre autorité éditoriale.

Pour l'instant les traits utilisés pour les approches en classification automatique (par apprentissage automatique supervisé) sont :

- des traits structurels concernant l'organisation logique du document : sans doute les traits les plus productifs pour séparer les grandes catégories, comme on l'a vu dans les toutes premières explorations de la question, par exemple le projet TypWeb (Beaudouin *et al.*, 2001) ;
- des traits lexico-syntaxiques comme ceux utilisés dans le travail fondateur de Biber (Biber, 1988) ;
- des configurations ponctuationnelles qui ont été identifiés comme pertinents pour des distinctions spécifiques, comme par exemple la distinction entre sites Web racistes et antiracistes (Valette et Grabar, 2004).

Un phénomène qui mérite à nos yeux d'être examiné à large échelle est celui des *rafales suffixales*, autrement dit des séquences contenant des séries de termes suffixés (généralement hors dictionnaire). Ces rafales suffixales avaient été repérées initialement lors du projet Wesconva, et nous nous étions posé la question de la recevabilité de dérivés dont la création semblait répondre à un besoin très local et stylistique, en grande partie humoristique. Nous les avons au final écartés des données utilisées pour quantifier les phénomènes étudiés, mais précieusement gardés. En voici quelques exemples :

> *J'ai testé pour vous... le* ***visionnage*** *juste après* ***lecturage*** *!*
> `http://www.melonthecake.com/page/10/`
>
> *Hobbies :* ***Cuisinage, véloballadage , lecturage , cinéphage*** *...*
> `http://www.viadeo.com/fr/profile/`
>
> *Jeudi, nous débutons une journée classique dans une famille :* ***levage, douchage, mangeage*** *(pancake au miel).*
> `http://amelieetyoann.over-blog.com/article-j-93-a-j-99-79192223.html`

Ce type de rafales en *-age* est très fréquent dans les textes dont les auteurs racontent une série d'actions en insistant sur l'accumulation et/ou le côté rituel.

> *Puis **rangeage**, **nettoyage**, **vidage**, **goutage**, **siestage**, **douchage**, **mangeage**, **dormage**....*
> http://www.sentiersnomades.com/spip.php?article73

> ***Ramenage** d'enfant chez son père. **Douchage**. **Mangeage**. Et **attendage** de turquoise.*
> http://inkr3dible.canalblog.com/archives/monsieur_turquoise_/index.html

L'utilisation de déverbaux est normale pour une telle énumération d'actions, mais le recours à des dérivés hors dictionnaire est très frappante. Si le suffixe *-age* semble majoritaire, il n'est pas exclusif :

> ***douchement**, **mangement**, **filmement**...*
> http://maison.et.travaux.du.nefast.over-blog.com/[...]carrelage.html

> *je vaque à les petites occupations du matin (discussion avec Filip, **douchation**, **maquillation**, **habillation**, **coiffation**... bref que des choses follement intéressantes ...)*
> http://irlandetrip.canalblog.com/archives/2007/07/16/5780915.html

Dans certains cas, les suffixes sont mélangés :

> *10h :retour chez Laura dans un état à peine croyable, **douchation** par groupe de 2, **séchage** puis **grattage** d'habits salubres et de maillots de bain(ce fut un combat difficile).*
> *10h45 :**Raccompagnage** de Julie chez Margot, puis **faisage** de pâtes pour le miam-miam.*
> http://la-jettatura.blogspot.com/

La volonté ludique est évidente, et parfois explicite :

> *En fin d'après-midi retour au camping. Puis : **arrivage**, **douchage**, **préparage**, **mangeage** et départ pour THE CONCERT (j'ai pas trouvé de rime...)*
> http://salsahora.free.fr/La-Seyne-sur-Mer-Festival-Cubain.html

En fait, de telles créations lexicales sont très souvent commentées par les scripteurs :

> *Toutes ces tragédies, en plus de vous donner une expérience assez solide en matière de rupture en tout genre, de **pleurnichage** dans les bras de vos proches et de **ridiculage** (comment ça, ca n'existe pas ce mot)(m'en fiche !)*
> http://uneautrequemoi.20six.fr/uneautrequemoi/art/1091791/Ainsi-va-la-vie

Une hypothèse à creuser serait le rôle de l'amorçage dans ce type d'énumération, lorsqu'une séquence naturelle de déverbaux (présents dans les dictionnaires) semble entraîner une contagion et la création de nouveaux dérivés pour compléter la série :

> *mes besoins :*
> ***nettoyage** de mes planches après scann*
> ***rattrapage** de dessin*
> ***lettrage** (mais là ya pas besoin de palette)*
> *colorimétrie de planche mais simple hein faut pas pousser !*

***essayage** de dessin direct à la palette*

*Après **farfouillage** et **lecturage** j'opterais (?) pour : Bamboo Fun Medium Pen & Touch*
`http://www.bdamateur.com/forum2/viewtopic.php?id=10500`

*Niveau **élégance prestance classance** et **distinctance**, je reste sur mes positions*
`http://ashleyandr.blogspot.com/2009/05/moda-vant-tout-le-monde.html`
(exemple emprunté à Dal et Namer (2010))

Certains cas sont également clairement parodiques, comme ceux qui se basent sur des créations très médiatisées. Par exemple le célèbre *bravitude* de Ségolène Royal a généré une quantité importante de dérivés parodiques en *-itude* [16] :

Ignioritude
7 décervelage, Sainte Forçats, pollorcètes
*Hier, notre **candidateuse socialistique** à la **présidenciation** de la **Républiquitude** était en **baladage** chez les Chinetoques.*
`http://www.melfrid.net/index.php?post/2007/01/07/130-ignioritude`

Si la plupart des travaux sur ces dérivés éphémères se concentrent sur leur intérêt pour l'étude des mécanismes de la suffixation, ils n'utilisent le contexte que pour expliciter leur interprétation. Une étude à large spectre de ces contextes permettrait sans doute d'observer le phénomène plus globalement, et de délimiter les conditions de leurs formations.

On rejoindrait donc ainsi les travaux de caractérisation des écrits du Web évoqués en première partie..

On voit donc à travers ces différentes approches la grande diversité des études possibles sur le matériau langagier issu du Web. La place des outils y est par contre différente de celle qu'ils prennent pour l'interrogation de corpus classiques. Certes, l'outillage informatique sert encore une fois à gérer la masse, mais cette fois il a un rôle plus proche de la nature des données, en s'attaquant au filtrage et à la caractérisation des contextes. Ce sont donc les données elles-mêmes qui posent au final le plus de problème, et cet état de fait a des implications bien au-delà des pratiques des linguistes (voir notamment les écrits collectifs de Pédauque (2003) sur les évolutions qu'a entraînées le développement du document numérique). Pour une discipline si attachée à son matériau, le manque d'informations, de stabilité, de fiabilité voire de matérialité des documents trouvés sur le Web sont autant de problèmes majeurs que la linguistique doit affronter de face.

Pour les écrits du Web comme pour les corpus plus traditionnels, les besoins sont spécifiques à l'étude envisagée, et le développement ne peut se faire indépendamment d'une compréhension fine des objectifs scientifiques, tout comme il ne peut reposer sur une solution logicielle générique préconçue. Même les gros corpus issus du Web ne répondent pas à tous les besoins que le Web a su faire émerger : aussi volumineux soient-ils, ils restent statiques.

16. Le site `http://www.echolalie.org/wiki/index.php?ListedItude` en répertorie un très grand nombre.

Une des pistes les plus intéressantes concernerait à mon avis la mise en place de procédures automatiques de caractérisation à la volée, ne visant pas à la catégorisation en genres, mais permettant par contre de donner des informations utiles sur une page Web pour une exploitation linguistique (par exemple l'identification de contextes considérés comme licites). De telles procédures couvriraient des besoins en googleologie, et seraient insérables dans des approches plus lourdement outillées.

Références

Atkins, B. S. et M. Rundell (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Baayen, R. H. et A. Neijt (1997). Productivity in context : a case study of a dutch suffix. *Linguistics*, **35** :565–587.

Baroni, M. et S. Bernardini (2004). Bootcat : bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon.

Baroni, M., S. Bernardini, A. Ferraresi, et E. Zanchetta (2009). The wacky wide Web : A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, **43**(3) :209–226.

Baroni, M., F. Chantree, A. Kilgarriff, et S. Sharoff (2008). Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*. Marrakech.

Beaudouin, V., S. Fleury, B. Habert, G. Illiouz, C. Licoppe, et M. Pasquier (2001). TypWeb : Décrire la toile pour mieux comprendre les parcours. In *CIUST'01 : Colloque International sur les Usages et les Services de Télécommunications*. Paris.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Brants, T. et A. Franz (2006). Web 1t 5-gram corpus version 1.1. Linguistic Data Consortium.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings 3rd Conference on Computational Lexicography and Text Research (COMPLEX '94)*. Budapest, Hungary.

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

Dal, G. et F. Namer (2010). Les noms en -ance/-ence du français : quel(s) patron(s) constructionnel(s) ? In *Actes du 2e Congrès Mondial de Linguistique Française*, pp. 893–907. Nouvelle Orléans, Etats-Unis.

De Schryver, G.-M. (2002). Web for / as corpus : a perspective for the african languages. *Nordic Journal of African Studies*, **11**(2) :266–282.

Duclaye, A., , F. Yvon, et O. Collin (2002). Using the Web as a linguistic resource for learning reformulations. In *Proceedings of the third international conference on language resources and evaluation (LREC'02*. Citeseer.

Fairon, C., K. Macé, et H. Naets (2008). Glossanet 2 : A linguistic search engine for rss-based corpora. In *Proceedings of the 4th web as corpus workshop (WAC-4)*, pp. 34–39.

Fletcher, W. (2006). Concordancing the web : promise and problems, tools and techniques. *Language and Computers*, **59**(1) :25–45.

Gala, N. (2003). Une méthode non supervisée d'apprentissage sur le web pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel. In *Actes de TALN*. Batz-sur-Mer.

Grefenstette, G. (1998). The world wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*. London.

Hathout, N. (2000). Morphological pairing based on the network model. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, pp. 35–38. Pyrgos, Grèce.

Hathout, N., F. Montermini, et L. Tanguy (2008). Extensive data for morphology : using the World Wide Web. *Journal of French Language Studies*, **18**(1) :67–85.

Hathout, N., F. Namer, et G. Dal (2002). An experimental constructional database : The mortal project. In P. Boucher, éd., *Many Morphologies*. Cascadilla, Somerville, Mass.

Hathout, N., M. Plénat, et L. Tanguy (2004). Enquête sur les dérivés en -able. *Cahiers de Grammaire*, **28** :49–90.

Hathout, N., F. Sajous, et L. Tanguy (2009). Looking for French deverbal nouns in an evolving Web (a short history of WAC). In *Proceedings of the Fifth Workshop on Web As Corpus (WAC)*, pp. 37–44. San-Sebastian, Spain.

Hathout, N. et L. Tanguy (2002). Webaffix : a tool for finding and validating morphological links on the WWW. In *Proceedings of LREC*. Las Palmas, Spain.

Hathout, N. et L. Tanguy (2005). WEBAFFIX : une boîte à outils d'acquisition lexicale à partir du Web. *Revue Québécoise de Linguistique*, **32**(1) :61–84.

Hundt, M., N. Nesselhauf, et C. Biewer, éds. (2007). *Corpus linguistics and the Web*. Rodopi, Amsterdam.

Jacquemin, C. et C. Bush (2000). Fouille du Web pour la collecte d'entités nommées. In *Actes de TALN*. EPFL, Lausanne.

Kehoe, A. et A. Renouf (2002). Webcorp : applying the web to linguistics and linguistics to the web. In *Proceedings of the WWW 2002 Conference*. Honolulu.

Keller, F. et M. Lapata (2003). Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, **29**(3) :459–484.

Kelling, C. (2003). The role of agentivity for suffix selection. In *Proceedings of the Third Mediterranean Meeting on Morphology*.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, **33**(1).

Kilgarriff, A. et G. Grefenstette (2003). Introduction to the special issue on web as corpus. *Computational Linguistics*, **29**(3) :333–347.

Lamiroy, B. et M. Charolles (2010). Les clitiques accusatifs versus datifs dans les constructions causatives en faire. In *actes du 2ème Congrès Mondial de Linguistique Française*.

Lin, D., K. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, *et al.* (2010). New tools for web-scale n-grams. In *Proceedings of LREC*.

Lüdeling, A., S. Evert, et M. Baroni (2007). Using Web data for linguistic purposes. In Hundt *et al.* (2007).

Martin, F. (2008). The semantics of eventive suffixes in french. In F. Schäfer, éd., *'SinSpec', Working Papers of the SFB 732*. University of Stuttgart.

Martin, F. (2013). Stage level and individual level readings of quality nouns. In N. Hathout, F. Montermini, et J. Tseng, éds., *Morphology in Toulouse : selected Proceedings of Décembrettes 7*. Lincom Europa.

Mourlhon-Dallies, F., F. Rakotonoelina, et S. Reboul-Touré (2004). Les discours de l'internet : quels enjeux pour la recherche ? *Les Carnets du Cediscor*, **8**.

Namer, F. (2003). Valider les unités morphologiques par le Web. In *Silexicale3, Actes du 3e Forum de Morphologie*.

Pédauque, R. T. (2003). Document : forme, signe et médium, les re-formulations du numérique. `http://archivesic.ccsd.cnrs.fr/sic_00000511`.

Plénat, M., S. Lignon, N. Serna, et L. Tanguy (2002). La conjecture de Pichon. *Corpus*, **1** :105–150.

Plénat, M. (1997). Analyse morphophonologique d'un corpus d'adjectifs dérivés en *-esque*. *Journal of French Language Studies*, **7** :163–179.

Plénat, M. et G. Boyé (2012, à paraître). Le choix des thèmes dans les dérivés désadjectivaux en français. In B. Tranel, éd., *Understanding Allomorphy. Perspectives from Optimality Theory*. Equinox.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. In G. Williams, éd., *La Linguistique de corpus*, pp. 31–46. Presses Universitaires de Rennes.

Rebeyrolle, J., D. Bourigault, C. Fabre, A. Josselin Leray, et L. Tanguy (2007). Un laboratoire d'observation de l'usage du vocabulaire recommandé par les instances officielles françaises. In *Colloque Presciptions En Langue*. Paris.

Resnik, P. (1999). Mining the Web for bilingual text. In *37th Meeting of ACL*, pp. 527–534. Maryland, USA.

Resnik, P., A. Elkiss, E. Lau, et H. Taylor (2005). The Web in theoretical linguistics research : Two case studies using the linguist's search engine. In *proceedings of the 31st Meeting of the Berkeley Linguistics Society*, pp. 265–276.

Rundell, M. (2000). The biggest corpus of all. *Humanising Language Teaching*, **2**(3).

Santini, M. (2007). Characterizing genres of web pages : Genre hybridism and individualization. In *Proceedings of the 40th Hawaii International Conference on System Sciences*.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni et Bernardini, éds., *Wacky ! Working Papers on the Web as Corpus*. GEDIT.

Sinclair, J. (2004). Corpus and text : basic principles. In M. Wynne, éd., *Developing Linguistic Corpora. A guide to Good Practice*, pp. 1–16. Oxbow.

Tanguy, L. (2012). *Complexification des données et des techniques en linguistique : contributions du traitement automatique des langues aux solutions et aux problèmes*. Mémoire d'habilitation à diriger des recherches, Université de Toulouse le Mirail.

Tanguy, L. et N. Hathout (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Actes de TALN'2002*, p. 254. Nancy France.

Valette, M. (2010). Méthodes pour la veille lexicale. In *Actes de la journée d'étude sur Le dictionnaire électronique : quelles perspectives pour les sciences humaines et sociales ?*, pp. 15–29.

Valette, M. et N. Grabar (2004). Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP. In *Actes des 7èmes Journées Internationale d'Analyse statistique des Données Textuelles (JADT)*, pp. 1106–1116. UCL-Presses Universitaires de Louvain, Louvain-la-Neuve, Belgique.

Valette, M. et F. Rastier (2006). Prévenir le racisme et la xénophobie – propositions de linguistes. *Les langues modernes*, **2** :68–77.

Wang, K., C. Thrasher, E. Viegas, X. Li, et B. june Hsu (2010). An overview of Microsoft Web N-gram corpus and applications. In *Proceedings of the NAACL-HLT demonstration session)*, pp. 45–48.

Wooldridge, R. (2004). Le web comme corpus d'usages linguistiques. *Cahiers de lexicologie*, **85** :209–225.

Xu, J. et W. B. Croft (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems*, **16**(1) :61–81.