Emmanuel CARTIER, Julie VIAUX
Université Paris 13 Sorbonne Paris Cité
http://dx.doi.org/10.18778/8088-785-5.02

Étude de la pénétration des anglicismes de type N ou ADJ(-)V*ing* à partir d'un corpus contemporain journalistique : les exemples de *bashing* et *shaming* en français contemporain

Résumé
Cet article porte sur les emprunts à l'anglais dans la langue française journalistique contemporaine. À l'aide d'un outil de repérage et d'analyse des néologismes appelé Neoveille, nous avons observé l'apparition et la circulation des emprunts à l'anglais *bashing* et *shaming* ainsi que de leurs composés. La méthodologie d'analyse des néologismes s'appuie sur l'exploitation d'un corpus volumineux, d'un moteur de recherche à fonctionnalités étendues, de mesures statistiques croisées avec des paramètres diastratiques et diatopiques et de modules de visualisation facilitant l'interprétation. L'analyse linguistique des emprunts en (N ou ADJ)-V*ing* amène à faire l'hypothèse de l'existence d'un moule emprunté potentiellement très productif.

Mots-clés : néologie, emprunt, anglicisme, patron morphosyntaxique, N-V*ing*, ADJ-V*ing*, linguistique de corpus, plateforme d'analyse des néologismes, *bashing*, *shaming, diastratie, diatopie*

Summary
The aim of this article is to study English loanwords of the form N/ADJ-Ving in the contemporary French language focusing on written press. We use a new web platform called Neoveille, whose goal is to detect and track neologisms on huge web corpora. The adopted methodology is based on statistical analysis of neologisms through several diastratic and diatopic parameters, the use of a monitor corpus, an associated extended search engine, and several visualization modules easing the analysis. Our interpretation of neologisms on the form N/ADJ-Ving is that in French the borrowed construction has begun to disseminate.

Keywords: neology, loanwords, corpus linguistics, frequency, bashing, shaming, diatopy, diachrony, diastraty

Introduction

Dans cet article, nous voudrions détailler les propriétés linguistiques des anglicismes du type N/ADJ(-)V*ing*. Cette structure déverbale est un procédé de construction de mots par composition très fréquente et productive en anglais, et a donné lieu en français à plusieurs emprunts, comme *baby-sitting*, *dog-sitting*, *fact-checking*, *home-staging*, etc. Plus récemment, un certain nombre de néologismes sont apparus, notamment sur les bases *-bashing* et *-shaming* : *French-bashing, Sarkozy-bashing, syndicalisme-bashing, body-shaming, slut-shaming.* Des emprunts plus limités sont également apparus sur le schéma générique N/ADJ(-)V*ing* : *trainsurfing, book crossing, clickfunding…*

Pour étudier ces emprunts, nous nous appuierons sur les nombreux corpus dorénavant disponibles, qu'il s'agisse de ceux qui sont disponibles sur Google et de ses « produits dérivés » (Google Ngram, Google Books, Google Trends), ou d'autres bases (Gallica, Europresse et Neoveille).

Deux objectifs principaux seront poursuivis :

- proposer une méthodologie de l'analyse des néologismes qui permette de combiner l'analyse linguistique qualitative traditionnelle à l'analyse quantitative permise par l'accessibilité de corpus volumineux et un outillage dorénavant mûr ;
- caractériser linguistiquement les emprunts en *-shaming* et *-bashing* : description synchronique formelle des néologismes (Sablayrolles, 2016) en *-bashing* et *-shaming*, qui semblent être non pas simplement des emprunts lexicaux mais qui répondent à un *moule emprunté*, étant donné

la variabilité possible des noms et adjectifs comme premier élément, et même du second composant ; étude de la variation diatopique et diastratique (Coșeriu et Geckeler 1981 ; Gadet 2003) ; étude diachronique du cycle de vie de ce type de néologisme en prenant appui sur les trois phases proposées par (Traugott *et al.* 2013) : apparition, dissémination, conventionnalisation.

Pour ce faire, cette contribution comprendra deux parties principales : la première exposera la méthodologie adoptée, en détaillant les corpus disponibles et les différents outils d'analyse et d'exploration mobilisés pour la présente étude. Dans la seconde partie, nous appliquerons cette méthode aux emprunts en *-bashing* et *-shaming*.

1. Méthodologie adoptée

1.1. Principes et méthodologie générale

Deux évolutions majeures et récentes ont ouvert aux sciences humaines de nouvelles possibilités d'exploration et d'analyse des phénomènes linguistiques. D'une part, la disponibilité de corpus volumineux voire très volumineux, qui permettent aux linguistes d'appuyer leurs intuitions sur des occurrences avérées dans différents registres de langue, principalement à l'écrit mais de plus en plus également à l'oral, et dans un nombre sans cesse grandissant de langues. La linguistique dite de corpus a formalisé ce champ de la linguistique, en énonçant les grands principes de constitution d'un corpus de référence et en permettant la confection de corpus de référence pour un certain nombre de langues. Depuis la démocratisation de l'informatique puis d'Internet, ces corpus sont principalement actuels et dynamiques (*monitor corpora*), issus d'abord de la publication des différents organes de presse traditionnels, d'institutions et organisations de tout type, et d'individus. Cette tendance n'a fait que s'accentuer avec l'avènement des réseaux sociaux, aboutissant aujourd'hui à une situation où la communication numérique supplante tout autre support de communication. Ces corpus numériques sont et seront à la fois une mémoire des usages linguistiques, mais également la matière d'études linguistiques plus objectives, car pouvant s'appuyer sur des usages réels et permettant de relativiser les hypothèses linguistiques. D'autre part, et de manière complémentaire, la masse des données disponibles oblige au développement d'outils de traitements spécifiques, car il n'est plus possible de parcourir humainement tous

les discours émis ; d'où le développement tout d'abord des moteurs de recherche, permettant d'accéder au contenu à partir de requêtes, et plus précisément le développement des concordanciers qui, sur le même principe, permettent de voir les contextes de l'unité linguistique recherchée (KeyWord In Context, KWIC). Ce premier outil a été considérablement amélioré, dans plusieurs directions, avec des méta-informations associées à chacune des sources d'informations : date, titre de la source d'information, langue, pays, région, domaine, etc. Toutes ces méta-informations peuvent alors être ré-exploitées dans le moteur de recherche afin de filtrer les résultats, ou encore d'obtenir des répartitions/ agrégations/ visualisation de résultats (par date, période, aire géographique, domaine ou registre de langue, etc.), pour un phénomène linguistique donné. Il est donc possible d'obtenir une vision diatopique et diastratique du phénomène étudié, via les méta-informations liées à chaque unité textuelle du corpus : il s'agit aussi d'appliquer aux corpus bruts un traitement linguistique (automatique ou manuel) permettant d'ajouter aux unités lexicales et/ou syntagmatiques des informations linguistiques (partie du discours, informations de sous-catégorisation, rôle syntaxique pour un groupe, etc.). Ces informations peuvent ensuite être exploitées via des langages de requêtes, par exemple pour retrouver automatiquement toutes les séquences … N + V*ing*! (voir CQP (Hardie, 2012), actuellement le langage de requêtes « linguistiques » le plus puissant), avec des calculs statistiques ou probabilistes de plus en plus sophistiqués, aboutissant, sans que l'on ait aujourd'hui le recul nécessaire, à des résultats qualitatifs surpassant tous les systèmes à base de connaissances humaines explicites (voir par exemple Word2Vec (Mikolov et al., 2013) pour les relations sémantiques, ou les analyseurs morphosyntaxiques). Sans détailler ces différents calculs[1], il est utile de rappeler que les principes de l'analyse quantitative se trouvent exprimés très clairement dans les écrits du théoricien du distributionnalisme (Harris 1954 et 1988). Les unités linguistiques se regroupent en classes distributionnelles qui peuvent être repérées automatiquement parce qu'elles vont nécessairement être répétées, d'une part, et parce que les membres d'une même classe distributionnelle vont partager un grand nombre de contextes, d'autre part. Ces deux principes sont de nos jours exploités avec toute la puissance nécessaire, et avec l'appui de corpus suffisamment larges pour donner des résultats significatifs. Ces méthodes quantitatives peuvent aujourd'hui être mises en œuvre avec profit et constituent des outils d'exploration linguistique, étant donné la grande masse de documents accessibles. Ces différents développements doivent, à notre sens, être dorénavant assimilés par

1 Nous renvoyons au travail de synthèse dans (Gries et Ellis 2015) et (Hilpert et Gries 2016) pour ce qui concerne les outils quantitatifs dans le cadre de l'analyse diachronique.

la linguistique elle-même. Il s'agit en effet non pas d'éliminer l'analyse linguistique qualitative, mais de la combiner avec l'analyse quantitative, qui permet d'asseoir les assertions sur des données attestables et quantifiables, voire de contredire les intuitions linguistiques. Cela nous conduit à établir un programme pour la linguistique outillée sur corpus :

- fonder les analyses linguistiques sur des **corpus de grande taille** étant donné leur disponibilité numérique ; ces corpus de grande taille sont également capitaux pour relativiser les phénomènes linguistiques étudiés ; il est à noter que les corpus oraux commencent également à être numérisés et exploitables ;
- fonder les analyses sur des **corpus contrôlés** (présence de méta-informations : date, source, type de documents, etc. permettant d'obtenir des paramètres d'étude variationnelle) ;
- fonder les analyses sur des **corpus annotés** si possible (analyse morphosyntaxique des corpus minimalement, mais également analyse syntaxique).

Et, du point de vue de l'exploration des corpus :

- utiliser les langages de requête « linguistique » (comme CQP par exemple) ;
- utiliser les outils de recherche et d'exploration quantitative disponibles (fréquence absolue et relative, évolution des fréquences, calcul des co-occurrences, etc.) ;
- utiliser les outils de visualisation des résultats disponibles (visualisation temporelle, répartition par type de documents, etc.).

Dans la pratique, l'outil idéal est encore à construire, mais il est clairement possible aujourd'hui d'aller plus loin que les moteurs de recherche généralistes. L'intuition et l'expertise linguistique restent fondamentales, mais peuvent et devraient s'appuyer de plus en plus sur les sources de données disponibles, et les outils automatiques d'analyse quantitative. Dans ce qui suit, nous passons en revue les sources de données sur lesquelles la présente étude linguistique s'est fondée, ainsi que les outils utilisés.

1.2. Sources des données

L'étude linguistique de néologismes actuels nécessite un accès à des sources de données contemporaines, en flux continu. Nous détaillons ci-dessous les différentes sources d'informations utilisées pour constituer ce corpus dynamique (*monitor corpora*). Des corpus moins récents ont également été utilisés pour mettre en perspective les phénomènes linguistiques.

1.2.1. Les applications Google

La première source, et la plus évidente, est le moteur de recherche *Google*, qui constitue encore aujourd'hui la source d'information textuelle la plus vaste. Cependant, ce moteur de recherche souffre de l'hétérogénéité des données qui sont proposées, de l'absence de méta-informations sur les données rendant parfois ardue la validation des occurrences, ainsi que de l'absence de dédoublonnage des textes. Il n'en reste pas moins que la recherche exacte permet de se faire une première idée de l'existence et de la circulation d'un néologisme. Le moteur de recherche propose également une recherche avancée permettant de restreindre les résultats à certaines langues, à certaines régions du monde ou encore de filtrer selon les dates. Notons également l'existence de deux autres moteurs de recherche Google plus spécialisés : *Google Actualités*[2], qui permet de restreindre les résultats à un corpus de presse généraliste et/ou spécialisée, et *Google Books*[3], qui permet d'accéder au corpus de livres numérisés le plus étendu (actuellement 7 % du stock mondial de livres imprimés, remontant pour le français à 1800). Ces deux derniers moteurs permettent d'avoir une idée de l'existence de néologismes récents dans des périodes antérieures de la langue, ainsi que de suivre l'évolution des néologismes récemment apparus. Pour ce qui concerne Google Books, il est également possible d'accéder aux données via une autre application, *Google Books Ngram Viewer*[4]. Ce moteur de recherche est construit sur la base des ngrams les plus fréquents dans cette immense base de données et il est possible d'obtenir le graphe de fréquence d'une lexie simple ou composée quelconque sur une période longue (depuis 1600 pour le français), puis d'accéder aux occurrences (Michel *et al.* 2011). *Google Ngram* est particulièrement utile pour se faire une idée du cycle de vie d'un néologisme et a donné et donne lieu à de multiples études en néologie (par exemple : Gulordava et Baroni 2011 ; Hamilton *et al.* 2016), cet outil étant librement accessible.

1.2.2. Les données Gallica[5] et BNF

La Bibliothèque nationale de France, depuis une vingtaine d'années, numérise, enrichit et donne accès à l'ensemble du fonds dont elle est dépositaire. Même si les données ne sont pas récentes, cette ressource est particulièrement utile étant

2 https://news.google.fr.

3 https://books.google.fr.

4 https://books.google.com/ngrams/.

5 http://gallica.bnf.fr/html/und/presse-et-revues/presse-et-revues.

donné la qualité des méta-informations et les fonctionnalités d'exploration de plus en plus étendues (voir par exemple l'accès proposé à la presse locale ancienne : http://presselocaleancienne.bnf.fr/accueil). Notons à ce propos que la BnF archive de manière systématique les données de la presse généraliste francophone depuis 1996 et donne accès à ces données sur les sites des bibliothèques nationales.

1.2.3. Europresse[6]

Cet agrégateur privé propose un accès payant à une base de données d'informations très riche en français : presse généraliste et spécialisée, réseaux sociaux, fils de presse, transcriptions d'émissions télévisées et radio, blogs, fiches biographiques. Sa principale limite est la qualité médiocre du moteur de recherche, qui génère du bruit dans les résultats, notamment parce qu'il est difficile d'exclure des sources anglo-saxonnes. Il est à noter que d'autres agrégateurs de presse[7] existent sur le marché, mais leurs solutions ne sont généralement pas accessibles aux universités.

1.2.4. Corpus spécifiques pour la linguistique de corpus

D'autres corpus contemporains du français sont disponibles par le biais d'interfaces de recherche spécialement consacrées à la linguistique de corpus. Les principaux corpus contemporains généralistes constitués pour le français sont FrWAC[8] (Baroni *et al.* 2009) qui est le résultat de la récupération automatique d'une partie du web francophone, et la version française de *Wikipedia*. Citons également le Corpus français, ressource proposée par l'université de Leipzig[9]. Il s'agit d'une base de données composée de près de 37 millions de phrases, soit environ 700 millions de mots. Elle a été constituée par le groupe de recherche TAL de l'université de Leipzig en Allemagne, et aménagée avec le concours de Daniel Elmiger et Alain Kamber (Université de Neuchâtel, Suisse). Les interfaces et outils de recherche ont été développés dans le cadre

6 http://www.europresse.com

7 Notamment : Factiva, LexisNexis, PressEDD, Cedrom-SNI, Argus de la Presse, Pickanews.

8 http://wacky.sslmit.unibo.it/doku.php?id=corpora . Un accès via l'interface NoSketchEngine est disponible ici : http://nl.ijs.si/noske/wacs.cgi/corp_info?corpname=frwac.

9 http://wortschatz.uni-leipzig.de/ws_fra/.

du projet Leipzig Corpora Collection of Computer Science de l'Université de Leipzig. Le corpus, destiné à l'étude du français contemporain écrit, est composé de trois parties : informations tirées de journaux francophones (plus de 19 millions de phrases), pages web (plus de 11 millions de phrases), *Wikipédia* (près de 6 millions de phrases). En dehors du fait de permettre une recherche par mot-clé, l'interface donne accès aux « voisins » de gauche et de droite les plus significatifs pour chaque terme, ce qui permet d'accéder à certaines locutions et constructions représentatives : on apprend ainsi que les termes suivants sont les plus significatifs à gauche de *bashing* : Québec, *monster*, *french*, *paki*, *China*, *French*. La limite principale de ce corpus est de s'achever en 2011. Plusieurs autres corpus du français contemporain ont été constitués, mais ne sont pas librement accessibles : les archives du *Monde* qui comprennent l'ensemble des articles publiés par le quotidien depuis 1989. Les sources précédentes ne sont pas dynamiques, mais couvrent des périodes plus ou moins longues jusqu'à 2011.

Une dernière ressource, particulièrement utile car couvrant la période actuelle, permettant de gérer le corpus et dotée de fonctionnalités sophistiquées d'analyse est liée à la plateforme Neoveille (Cartier, 2016). Il s'agit d'une plateforme, dont nous présenterons dans la section suivante les fonctionnalités, qui, du point de vue du corpus exploitable, a les caractéristiques suivantes :

- corpus en sept langues (français, russe, chinois, tchèque, grec, portugais du Brésil, polonais). Le corpus visé concerne les sites web dotés de flux RSS, protocole permettant de délivrer de l'information dans un format XML spécifique, à même de récupérer des méta-informations sur chaque article ainsi que les articles de presse complets ;
- corpus de presse généraliste et spécialisée. L'objectif étant d'étudier les innovations lexicales et d'usage, le corpus vise à représenter la langue générale, en prenant pour base le corpus de presse généraliste publié sur Internet. Au 1er janvier 2017, en français, le système a 227 sources d'informations différentes, réparties en 108 sites de presse généraliste et 119 sites de presse spécialisée ou de vulgarisation. On trouvera dans le tableau 1 une vision globale des sources d'informations par langue ;
- récupération continue depuis septembre 2015, trois fois par jour étant donné la « prolixité » des informations publiées sur internet, cela aboutit aux volumes d'articles détaillés dans le tableau 1 ;
- richesse des méta-informations. Chaque source d'information est dotée de méta-informations détaillées : langue source, pays source, nom du journal, domaine ; chaque article emporte également avec lui d'autres

méta-informations récupérées dans le flux RSS : titre, date, auteur, thématique(s). Ces informations sont très utiles pour filtrer les résultats de recherche pour étudier en synchronie les variations diatopiques, diastratiques et diaphasiques, et en synchronie notamment afin d'identifier les sociolectes influents (voir section suivante) ;

– extensibilité du corpus. Le système est évolutif, puisqu'il est possible, via l'interface web, d'éditer les sources d'informations : ajout, édition, suppression.

Tableau 1. Synthèse des sources d'information par pays, articles récupérés (Neoveille)

Langue	Nombre de fils de presse (Général / Spécialisé)	Nombre total d'articles récupérés (au 01/01/2017)
Chinois	10 (10/0)	56 552 (depuis le 07/04/2016)
Français	227 (108/119)	425 346 articles (depuis le 07/07/2015)
Grec	37 (33/4)	173 799 (depuis le 07/04/2016)
Polonais	26 (25/1)	78 781 (depuis le 07/04/2016)
Portugais du Brésil	22 (21/1)	115 337 (depuis le 07/04/2016)
Russe	40 (33/7)	318 380 (depuis le 07/04/2016)
Tchèque	41 (28/13)	98 459 (depuis le 07/04/2016)

Dans cet article, nous fonderons principalement nos analyses sur les données fournies par Neoveille, les autres corpus permettant de compléter nos hypothèses et les données. Nous utiliserons également deux grands corpus de l'anglais, étant donné que l'étude portera sur des anglicismes. Il s'agit de corpus contemporains qui permettront de figer l'état de langue concernant les emprunts que nous étudierons dans la suite de l'article :

– NOW[10] (corpus anglais) : il s'agit d'un corpus dynamique comprenant à ce jour plus de 4 milliards de mots extraits de la presse anglophone depuis 2010.
– WebCorp[11] qui donne accès à un corpus d'environ 128 millions de mots extraits du web de 2000 à 2010.

10 http://corpus.byu.edu/now/.

11 http://wse1.webcorp.org.uk/cgi-bin/DIA/index.cgi.

1.3. Outils de recherche, d'analyse et de visualisation des corpus pour la linguistique outillée

Comme indiqué dans l'introduction à cette section, l'accès à de grandes masses de documents textuels nécessite des outils de recherche et d'analyse appropriés. Nous présentons ci-après les différents modules utilisés dans le cadre de la plateforme Neoveille, ainsi que d'autres fonctionnalités présentes dans d'autres plateformes, qui ont pu être mobilisés dans la présente étude.

1.3.1. Fonctionnalités de recherche

La première fonctionnalité incontournable concerne la possibilité d'interroger le corpus, au moyen d'un langage de requête évolué. Il s'agit tout d'abord non pas d'obtenir comme résultat des textes, comme dans les moteurs de recherche généralistes, mais de donner des concordances, c'est-à-dire des extraits comprenant l'unité lexicale recherchée. Concernant Neoveille, qui se base sur le moteur de recherche Open Source Apache Solr[12], la recherche d'une unité lexicale peut se faire sous forme exacte (*binge-drinking* entre guillemets), de manière approximative (*binge-drinking* sans guillemets, qui donnera alors des extraits comprenant *binge-drinking*, mais aussi *binge* et *drinking*), approximative par stemmatisation (auquel cas *drinking* sera stemmatisé en *drink*), par expressions régulières (« .+ing » renverra par exemple tous les termes se terminant par *ing*). Nous renvoyons à la documentation officielle de la recherche sous Apache Solr[13], qui offre une panoplie très étendue d'options. Les résultats dans Neoveille sont présentés actuellement via l'interface Hue (voir figure 1).

Une recherche plus sophistiquée peut être effectuée via le langage de requête CQL. Cet outil, qui est embarqué dans les principaux outils de fouille de corpus (NoSkechEngine, IMS Corpus WorkBench pour ne citer que les plus utilisés), permet d'effectuer des recherches structurées, en utilisant l'annotation morpho-syntaxique des documents. Ainsi, pour extraire toutes les séquences Nom/Adj + *bashing*, il faut saisir [pos=N|A] [word="bashing"]. Cela donne les résultats de la figure 2.

12 http://lucene.apache.org/solr/.

13 http://apache.crihan.fr/dist/lucene/solr/ref-guide/apache-solr-ref-guide-6.3.pdf (section Search).

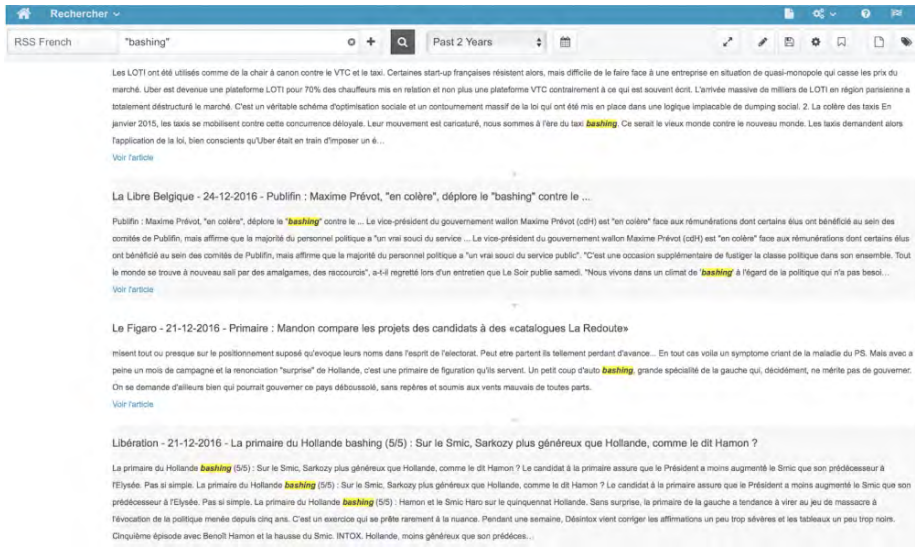

Figure 1. Extrait de résultats pour la requête *bashing* dans Neoveille (Hue)

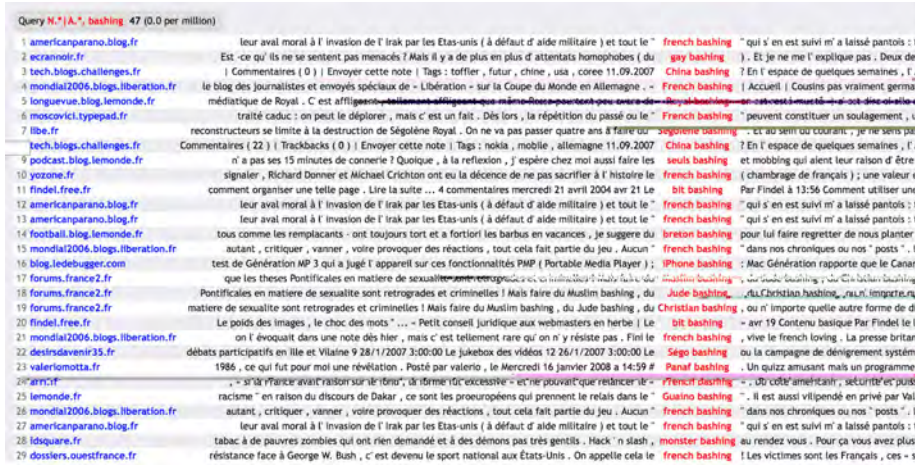

Figure 2. Extrait de résultats pour la requête [pos="N|A"] [word="bashing"] (NoSketchEngine, FrWac)

1.3.2. Filtrage des résultats

Les occurrences des néologismes sont caractérisées par un ou plusieurs mécanismes de formation (Sablayrolles, 2016), mais elles doivent également être caractérisées par les propriétés des sources dont ils sont issus. C'est le moyen le plus sûr :

– d'une part de s'assurer qu'un néologisme, d'un point de vue diastratique et/ou diaphasique, appartient à la langue générale ou à une langue

spécifique (presse féminine, langue des jeunes, etc.), voire à un domaine d'activité particulier.

– d'autre part de suivre l'évolution d'un néologisme qui peut rester cantonné à la presse féminine par exemple, mais peut également ensuite se répandre dans d'autres sociolectes, voire passer dans la langue générale.

Les méta-données explicitées pour chaque source d'informations dans Neoveille permettent de ce point de vue d'effectuer des filtres de façon dynamique sur ces paramètres de la variation linguistique, afin de ne visualiser par exemple que les attestations dans la presse féminine, que dans la presse généraliste, dans telle ou telle variante régionale, etc. ; d'un point de vue diachronique, de suivre la répartition des attestations selon les différents sociolectes, ou encore les différents topolectes. Ces filtres permettent non seulement de restreindre les occurrences visibles, mais également de visualiser les répartitions, synchroniquement ou diachroniquement. La figure 3 présente à titre d'illustration la répartition des occurrences de *bashing* selon quatre paramètres : la thématique des articles telle qu'elle est informée par le journal (répartition par thématique), le type de presse tel qu'il est informé pour chaque source d'information (répartition par type de presse), le pays d'origine de la source d'information (répartition par pays), enfin l'évolution temporelle depuis 2015. On constate aisément que *bashing* n'est plus cantonné ni à un type de presse particulier, ni à une thématique, et que sa fréquence d'apparition est très régulière.

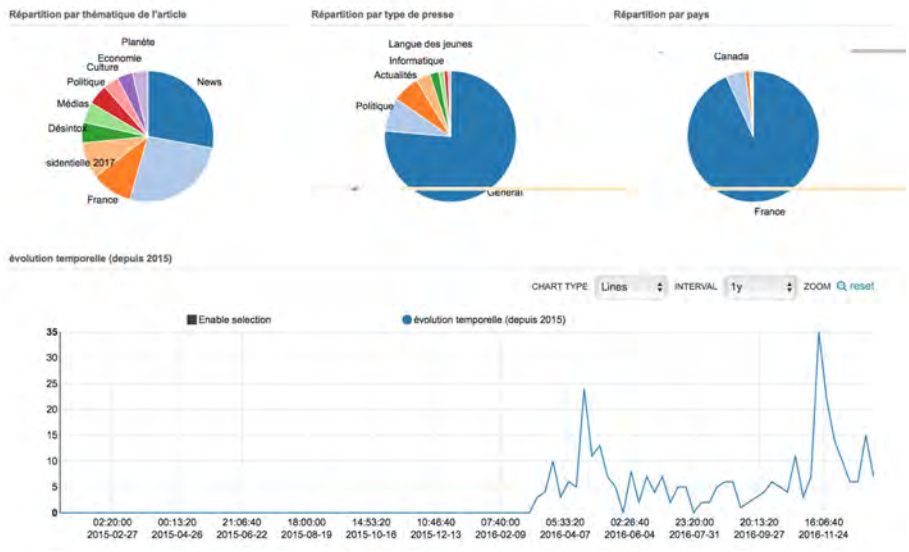

Figure 3. Répartition des occurrences de *bashing* selon différents points de vue (Neoveille)

1.3.3. Autres fonctionnalités

Développée depuis 2015, la plateforme Neoveille sera prochainement dotée de nouvelles fonctionnalités, notamment concernant un repérage fin des contextes d'apparition des néologismes et la possibilité de suivre l'évolution sémantique des lexies (néologismes de forme ou non) par les méthodes distributionnelles. Mais cela sort du cadre de la présente étude.

En guise de synthèse sur la méthodologie adoptée

Le système Neoveille, et d'autres outils, permettent de proposer un canevas général d'analyse des néologismes en corpus, dont nous pensons qu'il permet de manière précise, systématique et scientifique d'étudier ces phénomènes. D'un point de vue théorique, nous proposons un modèle qui reprend et étend les trois paramètres d'analyse proposés par (Gévaudan et Koch, 2010), en distinguant d'une part le paramètre formel (décrit le plus exhaustivement par Sablayrolles, 2016), d'autre part le paramètre sémantique (décrivant les modifications ou apparition de sens et les mécanismes sous-jacents), d'autre part le paramètre diastratique, que nous proposons, à la suite de (Coşeriu et Geckeler 1981) de scinder en quatre composantes (diatopie, diastratie, diaphasie, diamésie). L'étude diachronique consiste alors à décrire les évolutions de ces différents paramètres. L'outillage de Neoveille tient compte des différents aspects des néologismes et, étant donné le caractère dynamique des corpus, permet de détecter automatiquement de nouvelles formes (et à terme de nouveaux sens), de décrire leurs propriétés formelles et leurs propriétés sociolinguistiques, et de suivre l'évolution de ces propriétés au cours du temps. La plateforme, dans son état actuel, propose un outil permettant d'associer l'expertise linguistique humaine et les capacités de traitement automatique de l'information textuelle. Le dernier mot revient au linguiste, qui dispose ainsi de relevés précis et variés, et qui garde la main pour effectuer l'analyse linguistique précise.

Le reste de cette communication tentera d'appliquer ces principes afin d'avoir une idée de l'importance relative des emprunts à l'anglais de la forme N/ADJ-V*ing* parmi l'ensemble des néologismes, du cycle de vie de ce type de néologisme (émergence, évolution), du ou des types de discours/domaines impactés, de ses formes et de sa productivité, de l'ancrage de ses manifestations en français notamment.

2. Étude linguistique des emprunts en N ou ADJ(-)*Ving*

2.1. Situation du schéma en anglais

Le verbe *to bash* en anglais est ancien et signifie 'porter un coup violent', dès le XVII^e^ siècle selon le *Concise Oxford Dictionary of English Etymology (CODEE)*. En anglais, il est davantage utilisé avec un sens figuré depuis le milieu du XX^e^ siècle ('critiquer vertement, dénigrer'). Depuis lors, la fréquence du sens figuré dépasse celle du sens premier. Le statut de la construction N ou ADJ(-) *bashing* en anglais est extrêmement courant et productif depuis 2000, comme en témoignent des recherches faites dans les corpus NOW et dans WebCorp. L'évolution temporelle dans Google Ngram (figure 4) montre l'apparition progressive du sens figuré, ainsi qu'une explosion des occurrences à partir des années 80.

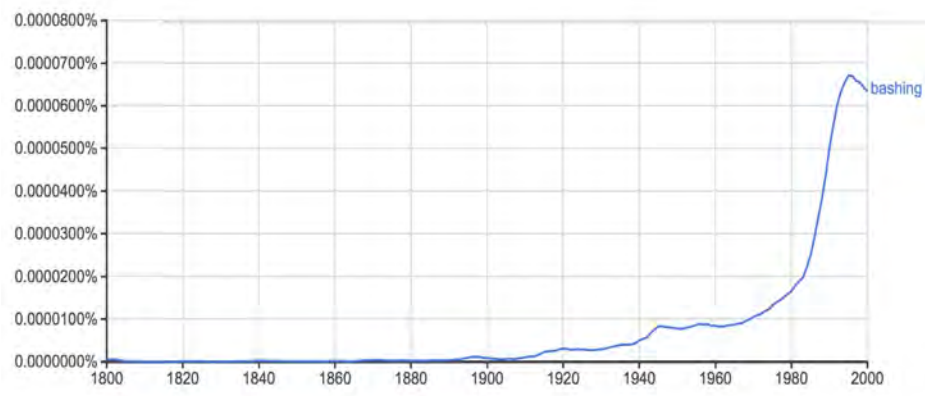

Figure 4. Évolution des occurrences de *bashing* dans le corpus Google Books depuis 1800

Concernant le verbe *to shame*, autre transitif direct, le *CODEE* le date du XIV^e^ siècle, avec une très grande stabilité sémantique, puisque le sens n'a pas évolué depuis lors. On note cependant dans Google Ngram une explosion des occurrences dans les années 80 également, correspondant à l'application du sens initial à différents objets, et dans un emploi déverbal.

2.2. Situation du schéma en français

2.2.1. Description formelle

On mettra de côté les emprunts directs que sont *bashing* et *shaming* que l'on trouve seuls pour s'intéresser de plus près aux constructions qui présentent un intérêt morphologique du point de vue de la langue d'accueil. Par exemple, depuis les années 1970, des scandales de diverses natures ont été rapprochés du *Watergate*, en anglais comme en français, par l'utilisation du formant *gate* dont on peut dire qu'il est devenu un suffixe : *Irangate*, *Monicagate*, et plus récemment *DSKgate* ou *dieselgate*, également appelé *Volkswagengate*. Pour les occurrences relevées dans la presse de langue française, les lexies contenant *bashing* et *shaming* donnent l'impression d'obéir à un petit nombre de patrons, puisqu'ils aboutissent à la création de lexies :

Tableau 2. Patrons attestés pour *bashing*

Patron	Exemples
Nom commun en français toujours au singulier + trait d'union + *bashing*	*syndicalisme-bashing*
Nom propre (toponyme) en français + trait d'union + *bashing*	*Chine-bashing*
Nom propre (toponyme) en anglais + trait d'union + *bashing*	*China-bashing*
Adjectif de nationalité en anglais + trait d'union + *bashing*	*French-bashing*
Nom propre (personne) + trait d'union + *bashing*	*Sarkozy-bashing, Hollande-bashing*
Apocope + trait d'union + *bashing*	*pédago-bashing, écolo-bashing*
Sigle + trait d'union + *bashing*	*SNCF-bashing, PS-bashing*

Tableau 3. Patrons attestés pour *shaming*

Patron morpho-sémantique	Exemples
Adjectif en anglais + trait d'union + *shaming*	*skinny-shaming*
Nom commun en anglais + trait d'union + *shaming*	*passenger-shaming*

Remarque : on a également relevé quelques occurrences sans trait d'union, pour *bashing* comme pour *shaming*.

Dans l'un et dans l'autre cas, le résultat final en français est un nom masculin. On peut noter dans un premier temps que *shaming* paraît pour l'instant moins productif.

L'effet de série n'est pas limité à ces deux déverbaux puisque dans Neoveille, si nous tapons la requête .+?*ing*, beaucoup d'autres résultats apparaissent : *problem-solving*, *happy-slapping*, *micro-targeting*, *shitposting* … Il s'agit d'un patron très courant en anglais, mais qui est en train d'être adopté par d'autres langues.

2.2.2. Analyse au moyen des mécanismes décrits dans (Sablayrolles, 2016)

L'analyse des mécanismes sous-jacents à la création de ces néologismes, en utilisant la typologie proposée par (Sablayrolles, 2016) fait apparaître trois mécanismes : tout d'abord, évidemment, l'emprunt, puisque les lexies sont empruntées. Cependant, les exemples montrent qu'il peut s'agir de deux types d'emprunts : soit un emprunt lexical direct (*China-Bashing* ou *bashing* lui-même), ou bien, et c'est là qu'il semble que ce qui est emprunté va au-delà de l'emprunt d'une ou plusieurs lexies, un moule de composition morphologique tout à fait étranger au français, auquel cas il s'agit donc d'un calque. Cette dernière hypothèse s'appuie d'une part sur le fait que de très nombreux néologismes en *bashing* comportent en premier membre des lexies du français, et les moules du tableau 1 montrent suffisamment que tout type de nom commun peut prendre la première place du patron. Pour ce qui concerne les adjectifs, on notera cependant que sémantiquement il n'est pas possible d'avoir autre chose que des gentilés. D'autre part, le schéma N-V*ing* est fréquent également en dehors des instances *bashing* et *shaming*. Nous serions donc dans une situation où un *moule de formation morphologique emprunté* s'intégrerait progressivement en français. Cependant, il est à prévoir que ce moule, étant donné la structure linguistique dont il provient, restera bloqué, car en français, il n'est a priori pas possible de construire des structures N-Participe présent.

Dans certains exemples, notons également la présence de fracto-compositions, comme dans *pédago-bashing, écolo-bashing*, etc. Ici, il faut considérer que la composition est seconde et s'applique sur le moule N-*bashing*. Une telle analyse n'est pas possible pour *shaming*, puisqu'il n'y a aucune attestation de lexie française en première position. Il s'agit donc purement d'emprunts lexicaux. On peut cependant se demander si l'extension aperçue pour *bashing* ne va pas s'étendre à *shaming*.

2.2.3. Description sociolinguistique : variations en diatopie et en diastratie

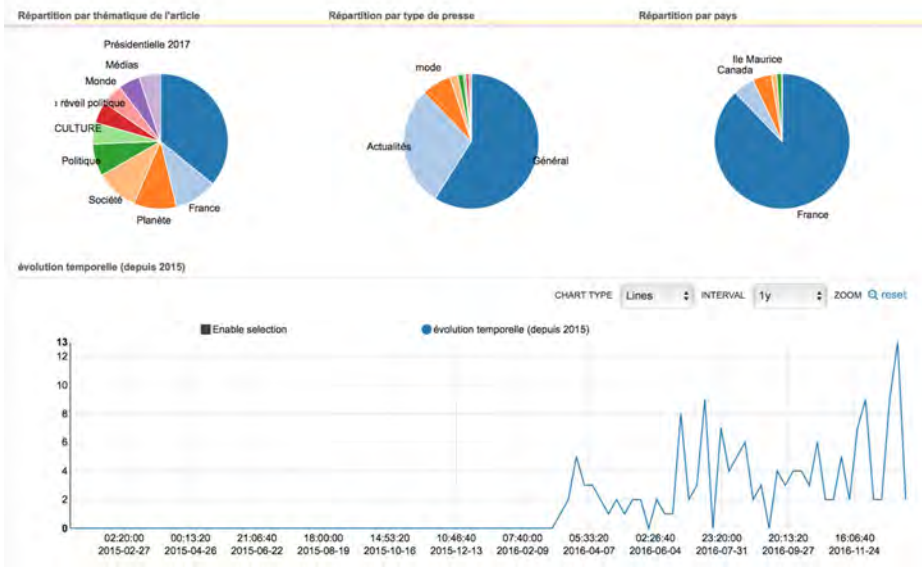

Figure 5. Répartition des occurrences de *shaming* selon différents points de vue (Neoveille)

L'étude diastratique de *bashing* montre, sur le corpus Neoveille (voir figure 3), qu'il n'appartient pas ni n'est cantonné à un sociolecte particulier, puisqu'il se rencontre à la fois dans la presse généraliste et dans plusieurs titres plus spécifiques. Sa sémantique spécifique reste évidemment liée à une activité sur internet, ce qui explique son contexte d'apparition. Du point de vue de la diatopie, même si des attestations sont présentes dans chacun des topolectes récupérés par Neoveille, il est difficile aujourd'hui de conclure quoi que ce soit, car ces titres ne sont récupérés que depuis peu.

Shaming est pour sa part beaucoup plus marqué (figure 5) : les premières attestations proviennent de la presse féminine, qui marque encore la répartition par type de presse (en additionnant les actualités et les articles de mode de la presse généraliste, et la presse féminine proprement dite). Du point de vue diatopique, mêmes réserves sur la quantité de données non métropolitaines, avec toutefois, là encore, des attestations régulières. Dans Europresse, entre le premier janvier 2015 et le 31 décembre 2016, 308 occurrences de *shaming* ont été repérées dans la presse de langue française, dont 42 dans le site de *Madame Figaro*, soit 14 %. La presse généraliste a repris certains composés et les utilise désormais couramment, sans l'assortir de remarques du type "phénomène venu

d'Outre-Atlantique". Le mot connaît une courbe ascendante, avec un essaimage dans la presse généraliste. Du fait de l'existence des réseaux sociaux depuis une dizaine d'années, faire honte à quelqu'un pour un défaut ou une action répréhensible est désormais public, certes, mais à une échelle mondiale.

2.2.4. Description diachronique récente

2.2.4.1. Évolution fréquentielle

D'après la base de données Gallica de la Bibliothèque nationale de France, la première occurrence de N ou ADJ-*bashing* en France date de 1974 avec *Paki-bashing* entre parenthèses et en italique, un xénisme qui illustre le concept d'agression physique à caractère xénophobe dans l'Angleterre des années 1970 (*Bulletin de la Société d'histoire moderne*, une publication de la Société d'histoire moderne et contemporaine, 1974). D'après la base de données Europresse, avec comme unique filtre la langue française, c'est *Le monde diplomatique* qui a employé, en 1981, le premier l'expression dans la phrase : « ils [les skinheads] clament haut leur volonté de "tabasser les Pakistanais" (*Paki-bashing*) ». Toujours dans Europresse, pour la totalité des années 1980, on ne trouve que quatre occurrences, puis 124 pour les années 1990 avec beaucoup d'articles québécois. L'utilisation s'est nettement répandue dans les années 2000 avec plus de mille occurrences dans des articles de presse rédigés en français, pour arriver à presque 9 000 occurrences pour la période qui va du premier janvier 2010 à la date de rédaction de notre article. Les premiers pics d'emploi répondent à des circonstances particulières, comme le *French-bashing* qui date du refus de la France de prendre part à la seconde guerre du Golfe en mars 2003. Un écho s'est récemment fait entendre en octobre 2014 avec un prix Nobel d'économie attribué à un chercheur français : le *French-bashing* et le *France-bashing* ont alors connu un second pic, justement parce qu'ils n'étaient plus de mise (voir figure 6).

Si les noms de pays ou les nationalités écrits en anglais sont nombreux, on relève une multiplication des structures de type Nom de famille / trait d'union / *bashing* : *Chirac-bashing* dès 2005, puis Sarkozy et *Ségolène-bashing* dès 2008, *Hollande-bashing* à partir de fin 2011, *Morano-bashing* début 2012, *Taubira-bashing* à partir de mars 2014, etc. On a affaire bien entendu à des personnes publiques et il semble que la particularité du formant *-bashing* est qu'il soit employé avec des noms de politiques.

Au gré des actualités et des éventuels scandales vont fleurir des *fonctionnaire-bashing* (première occurrence au Québec dans les années 1990 puis multiplication dans la presse de France à partir de 2013), *football-bashing*, *Sénat-bashing*, *SNCF-bashing* ou *pédago-bashing*.

Figure 6. Évolution des occurrences de N/ADJ(-)*bashing* dans Europresse

La toute première occurrence de *shaming* dans la presse de langue française d'après Europresse date de 1997 dans *Le Figaro* : il s'agissait de donner le nom anglo-américain d'un type de peine pour des délits mineurs (*shaming penalty* aux États-Unis, soit se faire homme-sandwich dans sa propre ville avec une pancarte indiquant « J'ai volé à la supérette »). Un autre châtiment par la honte, le *naming and shaming*, est né au début des années 2000. Il s'agit de rendre public le nom d'une entreprise se livrant à des pratiques discutables du point de vue éthique, en utilisant en particulier internet. Les *naming and shaming* apparaissent une à quatre fois par an dans la presse francophone entre 2000 et 2009 d'après Europresse (voir ces maigres volumes dans la figure 7). L'année 2012 marque une étape : les *naming and shaming* vont laisser la place à de plus nombreux *slut-shaming*, *fat-shaming* et *body-shaming* eux aussi importés des États-Unis. Si les magazines féminins francophones ont repris ces trois lexies, la construction avec *-shaming* semble assez peu productive par comparaison avec *-bashing*. On a vu dans la presse française des *tax-shaming* puis des *passenger-shaming* en 2014, mais en considérant de près la presse de langue anglaise à la même période, on ne peut que conclure que la langue française a simplement opéré des emprunts directs.

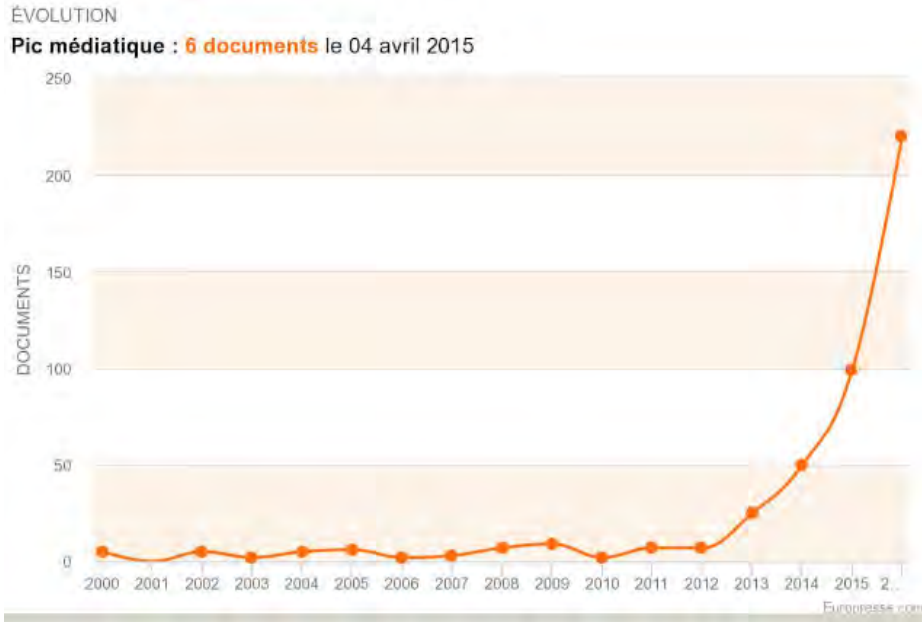

Figure 7. Évolution des occurrences de N/ADJ(-)*shaming* dans Europresse

Alors que de nouveaux composés de *bashing* sont créés dans les pays francophones et largement diffusés, *shaming* n'a pas de réelle productivité en langue française et a en outre moins de succès dans la langue journalistique contemporaine.

Le premier peut être considéré comme faisant partie d'une construction néologique productive. Le second est un emprunt nouvellement en circulation mais peine à sortir du cadre de la presse féminine. À un foisonnement de créations autochtones s'oppose une circulation plus limitée d'un ensemble de mots précis. L'ascension des deux courbes ci-dessus semble similaire mais les deux lexies ne fonctionnent finalement pas du tout de la même façon.

2.2.4.2. Signes supplémentaires de l'intégration de *bashing*

D'autres signes d'intégration morphologique peuvent nous éclairer : la dérivation nominale, verbale, adjectivale commencent à exister pour les composés de *bashing*. Le verbe du premier groupe *basher* est apparu en octobre 2011 dans le *Wiktionnaire*-français, avec le sens « Critiquer vertement », assorti de quelques exemples tirés de la presse gratuite. Dans la version française du *Huffington Post*, un journal en ligne, une chronique écrite par le journaliste Birenbaum pendant les années 2013–2014 s'appelait « Birenbaum bashe ».

Sur le site www.senscritique.com, des particuliers écrivent des critiques d'œuvres, notamment de films. Le titre de l'une d'entre elles est "Un bashage extrême et démesuré" à propos des mauvaises critiques récoltées par le film *Aladin* à sa sortie en novembre 2015. Le nom *basheur* a été relevé de manière anecdotique dans la presse (*Le Monde,* novembre 2014, par exemple) mais surtout en ligne sur des forums et comptes Twitter publics. On observe en ligne quelques exemples du verbe *shamer*, mais plus difficiles à repérer. Si *bashing* semble en train de se fondre dans la langue française avec des locuteurs qui créent des dérivés spontanés, *shaming* est moins dynamique.

2.2.4.3. Équivalents / concurrents autochtones

La Délégation générale à la langue française et aux langues de France (DGLFLF) et sa Commission générale de terminologie et de néologie devenue, depuis 2015, Commission d'enrichissement de la langue française, propose, depuis la loi Toubon, des équivalents autochtones aux termes empruntés qui apparaissent dans la langue française. C'est ainsi que pour *bashing* et *shaming*, ont été recommandés, consultables sur la base France*Terme* :

- *shaming* : il faut préférer *mise au pilori* (JO de janvier 2010),
- *bashing* : il faut préférer *éreintage* (JO de septembre 2013).

On peut vérifier, via Google Trends, si ces équivalents ont eu le succès espéré. Il s'avère (voir figures 8 et 9) qu'ils ne sont absolument pas utilisés.

Pour *bashing* / *éreintage*

Figure 8. Graphe d'évolution fréquentielle de *bashing* et *éreintage* (Google Trends)

Pour *shaming* / *mise au pilori*

Figure 9. Graphe d'évolution fréquentielle de *shaming* et *mise au pilori* (Google Trends)

Conclusion et perspectives

Dans cette dernière section, nous voudrions insister à la fois sur la méthodologie suivie pour effectuer l'analyse des néologismes et sur nos conclusions sur la pénétration en français du patron N ou ADJ-V*ing*. Du point de vue méthodologique, nous croyons avoir montré l'intérêt, voire la nécessité, d'un travail linguistique sur corpus et sur corpus outillé : moteur de recherche spécifique, outils d'analyse quantitative et de visualisation des données textuelles. La plateforme Neoveille, encore en développement, permet d'analyser les lexies néologiques en synchronie, en permettant d'une part de décrire les procédés formels de construction (à partir du modèle proposé par Sablayrolles, 2016), mais aussi de rendre compte des paramètres diatopique, diastratique et diaphasique. La diachronie (courte pour l'instant) est également prise en compte, permettant de suivre le cycle de vie des nouvelles formes ; très prochainement, il sera également possible d'obtenir des informations sur les changements linguistiques des formes lexicales existantes. La plateforme Neoveille est actuellement utilisée par un certain nombre de chercheurs dans sept langues pour traiter les néologismes de forme. Étant donné la généricité des processus développés, la possibilité d'ajouter de nouvelles sources d'information, et des traitements qui sont identiques d'une langue à l'autre, une multiplicité d'exploitation sont en-

visageables. Du point de vue des emprunts en N ou ADJ(-)V*ing*, notre intérêt s'est porté sur les formes en *bashing* et *shaming*, qui sont actuellement en pleine expansion, mais qui ne doivent pas occulter que le schéma a donné et donne lieu aussi à d'autres réalisations.

Au moins pour les formes en *-bashing*, l'intégration dans l'usage semble être fait, avec comme particularité que ce n'est pas seulement la lexie qui est empruntée mais la construction N ou ADJ-*bashing*, qui peut être rapprochée des modèles *-gate* ou encore *-sitting* (*baby-sitting*, *dog-sitting*, etc.). Il s'agit donc, comme dans le cas de *gate*, de l'importation d'un moule de construction morphologique emprunté à l'anglais qui semble se répandre en français. Deux pistes nous paraissent intéressantes à creuser pour approfondir la genèse de ce moule emprunté : il conviendrait de retracer l'histoire depuis le début à partir des années 1940, on trouve des mentions de *baby-sitting* en français, qui ne connaîtra pas une extension aussi fulgurante que *-bashing* mais lui est antérieur. Y aurait-il d'autres occurrences encore plus anciennes? Et n'y a-t-il pas aujourd'hui une multiplication des emprunts sur ce moule, maintenant assimilé dans son interprétation par les locuteurs français ? Il serait par ailleurs intéressant d'étudier le destin de *-bashing* et *shaming* dans d'autres langues.

Références bibliographiques

Baroni Marco, Bernardini Silvia, Ferraresi Adriano, & Zanchetta Eros, 2009, « The WaCky wide web : a collection of very large linguistically processed web-crawled corpora », *Language resources and evaluation*, 43(3), p. 209–226.

Cartier Emmanuel, 2016, « Neoveille, système de repérage et de suivi des néologismes en sept langues », *Neologica*, 10, Paris, Garnier, p. 101–131.

Coşeriu Eugeniu and Geckeler Horst, 1981, *Trends in structural semantics*. (Tübinger Beiträge zur Linguistik, 158). Tübingen : Narr.

Gadet Françoise, 2003, *La variation sociale en français*, Paris, Ophrys, Coll. L'essentiel.

Gévaudan Paul et Koch Peter, 2010, « Sémantique cognitive et changement sémantique », *Grandes voies et chemins de traverse de la sémantique cognitive*, Mémoire de la Société de linguistique de Paris, XVIII, p. 103–145.

Gries Stefan Th. & Nick C. Ellis, 2015, « Statistical measures for usage-based linguistics », *Language Learning* 65 (Supplement 1), p. 1–28.

Gulordava Kristina, Baroni Marco, 2011, « A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus »,

Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011, p. 67–71.

Hamilton William L., Leskovec Jure, and Jurafsky Dan, 2016, Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, *Proceedings of ACL 2016*, Berlin.

Hardie Andrew, 2012, CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409.

Harris Zellig Sabbetai, 1954, *Distributional structure.* Word, 10(23), 146–162. [Traduction française 1970 : La structure distributionnelle. Analyse distributionnelle et structurale ed. by Jean Dubois & Françoise Dubois-Charlier Langages, nº. 20, 14–34. Paris : Didier / Larousse.]

Harris Zellig Sabbetai, 1988, *Language and Information*. New York: Columbia University Press, ix, 120 pp. [Revised version of the Bampton Lectures given at Columbia University, New York City, in Oct. 1986].

Hilpert Martin et Stefan Th. Gries, 2016, « Quantitative approaches to diachronic corpus linguistics », In Merja Kytö & Päivi Pahta (éds.), *The Cambridge Handbook of English Historical Linguistics*, p. 36–53. Cambridge : Cambridge University Press.

Michel Jean-Baptiste, Shen Yuan Kui, Aiden Aviva Presser, Veres Adrian, Gray Matthew K., Brockman William, Pickett Joseph P., Hoiberg Dale, Clancy Dan, Norvig Peter, Orwant Jon, Pinker Steven, Nowak Martin A., and Aiden Erez Lieberman, [est-il dans l'usage d'indiquer tous les euteurs quand ils sont si nombreux ?] 2011, « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, 14, vol. 331, Issue 6014, p. 176–182.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg et Dean, Jeffrey, 2013, Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (p. 3111–3119).

Pruvost Jean, Sablayrolles Jean-François, 2016, *Les néologismes*, Paris, Presses Universitaires de France, Coll. Que sais-je ?

Traugott Elizabeth Closs and Trousdale Graeme, 2013, *Constructionalization and Constructional Changes*, Oxford, Oxford University Press.