Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Que recèlent les données textuelles issues du web ?

Adrien Barbaresi[1] Gaël Lejeune[2]
(1) Académie des Sciences de Berlin-Brandenbourg, Jägerstraße 22-23, 10117 Berlin, Allemagne
(2) Sorbonne Université, 1 rue Victor Cousin, 75005 Paris, France

RÉSUMÉ

La collecte et l'usage opportunistes de données textuelles tirées du web sont sujets à une série de problèmes éthiques, méthodologiques et épistémologiques qui méritent l'attention de la communauté scientifique. Nous présentons des études empiriques de leur impact en linguistique et TAL centrées sur la forme (méthodes d'extraction des données) ainsi que sur le fond (contenu des corpus).

ABSTRACT

What do text data from the Web have to hide ?

The opportunistic gathering and use of text data taken from the Web are subject to a whole series of ethical, methodological and epistemological problems which could benefit from the interest of the research community. We present empirical studies of their impact in linguistics and natural language processing, with respect to their form (extraction methods) and to their contents.

MOTS-CLÉS : Construction de corpus, Science du web, Extraction de texte, Méthodes d'évaluation.

KEYWORDS: Corpus construction, Web science, Text extraction, evaluation methods.

1 Introduction

Le web est fréquemment perçu comme un « réservoir indifférencié de textes à analyser » pour le TAL (Tanguy, 2013) et l'on peut affirmer sans risque que web et TAL poursuivent leur « histoire commune » (*op. cit.*). Les données issues du web y sont en effet omniprésentes, à la fois en tant qu'instantané d'un état de la langue, de données à analyser pour elles-mêmes, mais aussi de références destinées à construire des modèles de langue ou des ressources langagières. De la collecte au corpus, non seulement opportuniste (McEnery & Hardie, 2011) mais également « prêt-à-utiliser », il n'y a souvent qu'un pas. Le Common Crawl [1] notamment s'est imposé comme source majeure pour des tâches variées, de la traduction automatique neurale (Smith *et al.*, 2013) à la construction et (dans une situation optimale) à l'affinage de modèles de langue basés sur des techniques d'apprentissage profond nécessitant des données massives (Suárez *et al.*, 2019). Cette évolution a conduit à des problèmes récurrents d'ordre éthique, à l'image du robot conversationnel Tay lancé par Microsoft en 2016 sur Twitter et stoppé 16 heures après son entrée en fonction en raison de l'ampleur et de la gravité des messages racistes et sexistes « appris » et ensuite (re-)publiés par le robot [2]. De même, des modèles entraînés sur des données massives (d'origine contrôlée ou non) intègrent une série de biais sociétaux (Caliskan *et al.*, 2017). Malgré une certaine impression de facilité quant à la construction de corpus, les méthodes utilisant des corpus web nécessitent des dispositifs expérimentaux et des

1. https ://commoncrawl.org
2. https ://en.wikipedia.org/wiki/Tay_(bot)

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 19–28.
HAL : hal-02750223.

instruments ad hoc (Valette, 2008) afin d'estimer leur qualité et leur adéquation aux tâches proposées. D'un point de vue épistémologique, la simple accumulation de données textuelles ne rend pas pour autant ce terrain intelligible et un retour analytique sur ces données s'avère nécessaire [3].

L'objet de cet article est de documenter et de commenter des problèmes liés au contenu des corpus web ainsi qu'aux processus d'extraction qui permettent d'accéder aux textes. Nous considérons des corpus construits directement sans recours à des données pré-existantes et utilisons à cette fin des méthodes de parcours du web, *web crawling* (Olston & Najork, 2010), afin de découvrir des hôtes hébergeant des pages web d'une part (méthode « généraliste ») et des documents au sein de domaines déjà connus (méthode « ad hoc ») d'autre part (Barbaresi, 2015). La simple notion de texte comme unité langagière cohérente est sujette à caution sur ces données reprises et potentiellement déformées par les outils de traitement. Nous souhaitons apporter un retour quantitatif, avec un examen des méthodes d'extraction, ainsi qu'un examen qualitatif, avec l'exemple des discours haineux, afin de remettre en question l'usage sans garde-fous des textes tirés du web et des informations qu'ils recèlent.

2 Examen des méthodes d'extraction

Étant donné le code source d'une page web, le processus d'extraction de contenu consiste à détourer le contenu textuel utile (c'est-à-dire notamment sans les éléments de structure, la publicité ou encore les commentaires) et à identifier les méta-données. Concrètement, cette tâche implique une conversion du format HTML vers un autre format (souvent plein texte ou XML). La construction de corpus à partir du web est devenue un élément si commun des chaînes de traitement de TAL que les détails techniques sur sa mise en œuvre sont souvent omis. Or, peut-on réellement s'abstenir de se demander ce qui est intégré dans des corpus ? Afin d'illustrer ce problème de contenu, nous comparons différents extracteurs récents et/ou populaires afin d'observer ce que des métriques d'évaluation spécialement conçues révèlent concernant leur efficacité. Nous laisserons de côté le choix des sources en elles-mêmes pour nous concentrer sur les résultats de l'extraction, qui forment la base de décisions quant à l'inclusion dans le corpus final (Schäfer *et al.*, 2013), et nous intéressons en particulier à la question multilingue. En effet, les outils mis à disposition de la communauté sont très souvent conçus pour la langue anglaise, l'applicabilité à d'autres langues étant souvent considérée de manière opportuniste comme allant de soi : si les extracteurs fonctionnent sur l'anglais, alors les mêmes ordres de grandeur de résultats seront obtenus dans d'autres langues.

Corpus et outils Nous reprenons un corpus proposé par Lejeune & Zhu (2018) qui comprend près de 1.700 documents en 5 langues (475 en anglais, 405 en chinois, 273 en grec, 274 en polonais et 267 en russe) avec la version HTML d'une part et une version de référence nettoyée manuellement d'autre part. La sélection des outils s'est faite sur trois critères : simplicité d'utilisation (existence d'une version ou *wrapper* PYTHON puisque ce langage est très répandu, en particulier en TAL) ; disponibilité sous licence libre (pour les mêmes raisons, nous tenons uniquement compte d'outils directement accessibles) ; popularité ou nouveauté (l'état de l'art fournit des informations, notamment pour des outils relativement anciens comme BOILERPIPE).

3. « La collecte et la mise en circulation des données dans des dispositifs adéquats aboutissent à une mise en ordre du monde qui relève d'une cosmétique : ces données sont triées, classées, archivées. Ces opérations textuelles permettent l'accumulation et donc l'archivage mais ne rendent pas pour autant le terrain intelligible. Cette intelligibilité du terrain est le résultat d'une deuxième opération, celle de mise en ordre des données accumulées, de leur traitement, de leur analyse et de leur restitution. » (Calberac, 2010, p. 104)

En conséquence, nous choisissons de comparer les outils suivants, classés en différentes catégories selon leur finalité : conversion du format HTML vers le format texte (I), intégration dans un contexte plus large d'extraction et d'analyse d'information (II), extraction proprement dite du texte principal d'une page web (*boilerplate removal*, III).

Cat.	Outil	Version	Adresse Github	Référence
I	HTML2TEXT	2020.1.16	Alir3z4/html2text	
I	INSCRIPTIS	1.0	weblyzard/inscriptis	
II	NEWSPAPER3K	0.2.8	codelucas/newspaper	
II	NEWS-PLEASE	1.4.25	fhamborg/news-please	(Hamborg *et al.*, 2017)
II	READABILITY	0.7.1	buriy/python-readability	
III	BOILERPY3	1.0.2	jmriebold/BoilerPy3	(Kohlschütter *et al.*, 2010)
III	DRAGNET	2.0.4	dragnet-org/dragnet	(Peters & Lecocq, 2013)
III	GOOSE3	3.1.6	goose3/goose3	
III	JUSTEXT	2.2.0	miso-belica/jusText	(Pomikálek, 2011)
III	TRAFILATURA	0.4.1	adbar/trafilatura	(Barbaresi, 2019)

Ce comparatif fait également l'objet d'une démonstration (Lejeune & Barbaresi, 2020), ces résultats peuvent être reproduits en utilisant les données et scripts mis à disposition [4]. Nous optons ici pour une version abrégée : une seule des configurations de BOILERPIPE (BP3_Article), configuration par défaut pour JUSTEXT et TRAFILATURA.

Mesures Les mesures d'évaluation de la campagne Cleaneval (Baroni *et al.*, 2008) sont fondées sur la préservation des séquences de tokens. Bien qu'imparfaites, elles ont le mérite d'être globalement utilisées par la communauté scientifique (Weninger *et al.*, 2016). Nous ajoutons une mesure plus simple, fondée sur la préservation du vocabulaire, qui donne des résultats tout à fait comparables. Cette évaluation nécessite une vérité de terrain, que nous appellerons GT et GT_{tok} pour la séquence de tokens correspondante. Nous nommons RES le résultat de l'extraction automatique et RES_{tok} la séquence de tokens correspondante, en reprenant le tokeniseur fourni par Cleaneval. La mesure Cleaneval vérifie à quel point la séquence de tokens extraite automatiquement (RES_{tok}) est similaire à la séquence de référence (GT_{tok}). L'algorithme de Ratcliff/Obershelp (Ratcliff & Metzener, 1988) est utilisé pour détecter les plus longues séquences de tokens communes et non-redondantes, sa complexité quadratique est peu efficace et ses résultats ne sont pas immédiatement interprétables. Notre mesure plus simple (*occ_eval*) vérifie si le nombre d'occurrences des tokens correspond aux nombre d'occurrences attendues.

Quelques résultats Nous détaillons ici quelques résultats tirés d'une analyse des outils (Barbaresi & Lejeune, 2020) et nous concentrons sur leur variabilité. Le tableau 1a présente les résultats globaux avec la métrique `clean_eval`. La précision et le rappel sont des moyennes des précisions et rappels par document. La F-mesure est calculée à partir de ces moyennes [5]. Le tableau 1b présente les résultats obtenus avec `occ_eval`, ceux-ci diffèrent assez peu. L'ordre de grandeur des résultats et la « hiérarchie » entre les outils sont conservés à ceci près que JUSTEXT semble pénalisé par la mesure `clean_eval`.

Variation selon les langues D'un point de vue général, l'outil le plus fiable semble être BP3_ART, READABILITY TRAFILATURA et JUSTEXT se situant juste derrière. Toutefois, les moyennes (micro ou macro) masquent des différences entre les langues, comme nous le montrons dans les tableaux 2a

4. https ://www.github.com/run-dimeco/waddle
5. La moyenne des f-mesures donne un score peu intuitif car souvent inférieur à la macro-précision et au macro-rappel.

Outil	Macro F	Micro F	Micro P	Micro R
BP3_Art	72,73	**78,84**	82,80	75,24
READ	74,62	75,87	72,18	79,96
TRAF	**75,69**	75.71	68.33	84.87
JT	63,7	71,22	78,93	64,88
DRAGNET	58,21	69,66	**87,53**	57,85
NPLEASE	48,84	58,46	69,00	50,72
GOOSE	37,87	53,93	83,89	39,74
NPAPER	32,37	50,83	82,20	36,78
INSCRI	40,10	42,95	27,72	**95,28**
HTML2T	31,2	33,98	20,86	91,47

(a) Mesure `clean_eval`

Outil	Macro F	Micro F	Micro P	Micro R
BP3_Art	70,41	**76,38**	80,60	72,57
JT	67,7	74,13	81,36	68,08
READ	71,01	73,25	72,43	74,09
TRAF	68,63	72.89	65.02	82.93
DRAGNET	56,12	67,09	86,82	54,67
NPLEASE	50,92	66,64	**92,03**	52,23
GOOSE	41,72	57,74	89,42	42,64
NPAPER	36,18	54,78	88,68	39,63
INSCRI	34,98	37,10	23,22	**92,22**
HTML2T	30,95	33,45	20,56	89,80

(b) Mesure `occ_eval`

TABLE 1: Evaluation sur le corpus multilingue, F-mesure calculé à partir des micro-moyennes de la Précision et du Rappel (sur fond grisé la différence d'ordre entre les deux mesures)

Outil	F-mes.	Préc.	Rap.
NPAPER	**91,32**	91,34	91,31
GOOSE	90,69	**92,94**	88,54
NPLEASE	88,91	87,89	89,96
DRAGNET	88,78	88,52	89,04
READ	87,16	84,31	90,21
BP3_Art	87,00	87,50	86,51
JT	84,86	83,16	86,62
TRAF	82.58	74.28	92.97
INSCRI	45,84	29,88	**98,46**
HTML2T	44,61	28,98	96,84

(a) `occ_eval` (Anglais)

Outil	F-mes.	Préc.	Rap.
JT	**76.29**	71.64	81.59
READ	74.27	72.29	76.36
TRAF	71.20	64.80	79.02
BP3_Art	69.31	70.11	68.53
DRAGNET	50.94	85.13	36.34
NPLEASE	42.64	93.16	27.64
GOOSE	40.24	90.96	25.83
INSCRI	32.53	19.77	**91.75**
HTML2T	29.55	17.63	91.35
NPAPER	5.14	**92.34**	2.64

(b) `occ_eval` (Russe)

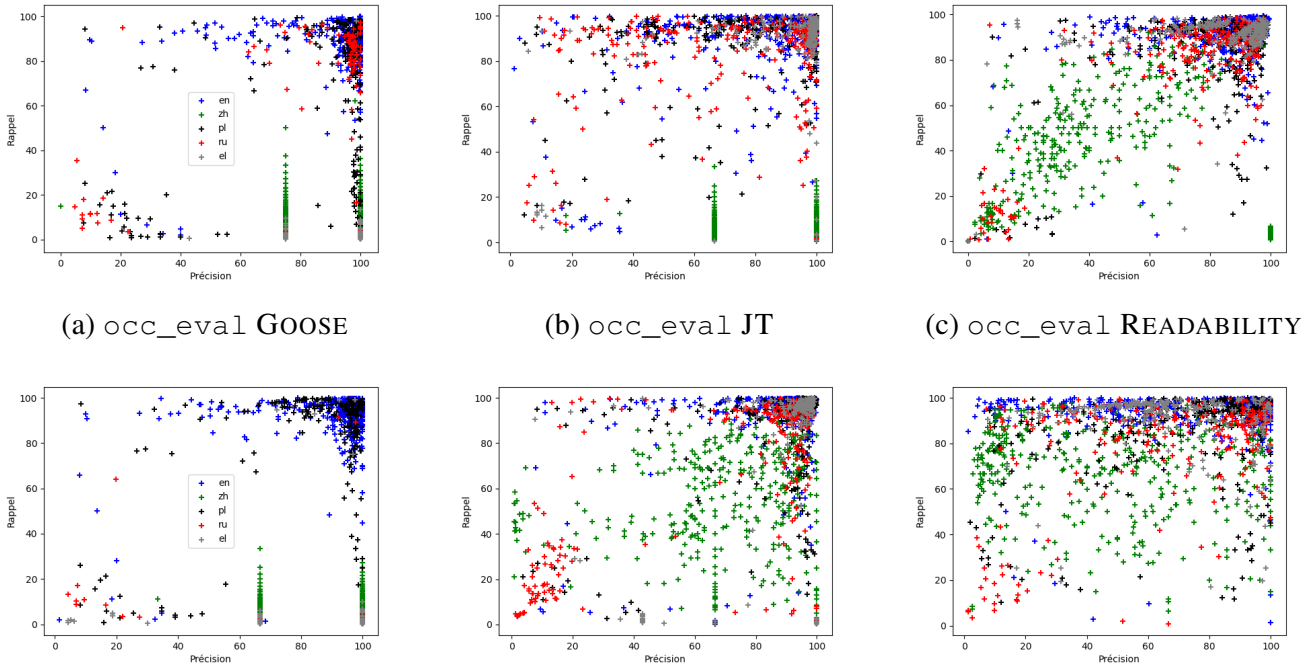
Outil	F-mes.	Préc.	Rap.
BP3_Art	**63.30**	71.28	56.93
TRAF	55.48	46.81	68.09
DRAGNET	44.53	81.81	30.59
READ	42.36	48.00	37.91
GOOSE	20.60	82.54	11.77
JT	19.19	82.32	10.86
NPAPER	19.17	82.72	10.84
HTML2T	13.83	7.62	74.87
NPLEASE	13.31	**97.52**	7.14
INSCRI	12.97	7.06	**79.52**

(c) `occ_eval` (Chinois)

TABLE 2: `occ_eval` par langue, sur fond gris les systèmes les plus performants au global

à 2c qui présentent les résultats sur les sous-corpus anglais, russe et chinois pour une sélection d'outils parmi les plus efficaces. Nous avons marqué en grisé les performances des 4 outils les plus efficaces sur le corpus multilingue, ce qui permet de voir qu'ils sont bien placés, sauf sur le sous-corpus anglais où des outils très spécialisés sont plus performants. L'anglais est évidemment la langue la mieux traitée puisque 9 des 11 systèmes testés ont une F-mesure au dessus de 80% (2 en Grec et 3 en polonais). En ce qui concerne les performances par outil, BP3_ART, le meilleur outil selon la micro-moyenne générale, est inégal selon la langue traitée : très efficace comparativement aux autres sur le chinois mais en-dessous de ses concurrents sur le russe. JUSTEXT s'impose sur cette langue, ce qui semble valider la robustesse de son approche multilingue fondée sur les mots outils. Ses résultats sont compétitifs sur l'anglais et c'est sans doute sur le chinois qu'il perd la confrontation à distance avec BP3_ART. En effet les modèles langagiers de JUSTEXT sur les mots outils ne sont pas applicables à la langue chinoise. Si l'on partait des résultats sur l'anglais pour choisir un outil d'extraction de contenu, nous pourrions être tentés de choisir NEWSPAPER ou encore GOOSE. Mais leurs performances sont très variables, en particulier pour le grec (moins de 6% de F-mesure avec un rappel très faible). NEWSPAPER et NEWSPLEASE apparaissent véritablement spécialisés sur l'anglais.

Visualisation de la variation à l'échelle des documents Afin de mieux visualiser ces variations nous présentons dans la Figure 1 les résultats par document pour le corpus global pour les 6 outils les plus performants au point de vue monolingue ou multilingue. Les écarts types sont plutôt élevés en général (± 24 sur la précision pour JUSTEXT par exemple, ou sur des sous-corpus particuliers, ± 17 sur la précision de GOOSE sur l'anglais). Les graphiques permettent de saisir d'un coup d'œil l'importance de cette variabilité. On peut ainsi observer que les documents en anglais sont mieux



(a) `occ_eval` GOOSE (b) `occ_eval` JT (c) `occ_eval` READABILITY

(d) `occ_eval` NEWSPAPER (e) `occ_eval` BP3_ARTICLE (f) `occ_eval` TRAFILATURA

FIGURE 1: Visualisation mettant en rapport la précision (abscisse) et le rappel (ordonnée) pour chaque document du corpus (*el* = grec, *en* = anglais, *pl* = polonais, *ru* = russe, *zh* = chinois)

traités (en bleu). Les points correspondant au grec (en gris) sont peu nombreux, ce qui correspond à un plus faible nombre de sources (les points groupés sur la même abscisse). La dispersion des points et les codes couleur permettent ainsi de saisir des informations sur la composition du corpus lui-même, et d'en déduire ici qu'il peut être intéressant de tenir compte du nombre de documents par source ou de retenir la macro-moyenne sur les sources.

En regard de ces graphiques, le comportement des outils peut être classé en trois catégories distinctes. GOOSE est plutôt efficace en termes de précision, d'où le grand nombre de points situés à droite du graphique mais aussi les problèmes afférents bien visibles par des pics de faible rappel. Au contraire, JUSTEXT offre un bon rappel pour plus de documents, d'où le grand nombre de points en haut de la courbe. Enfin, READABILITY laisse apparaître une diagonale, ce qui suggère plus de résultats équilibrés entre la précision et le rappel et explique pour partie les bons résultats en F-mesure, et notamment en chinois (en rouge dans la courbe). TRAFILATURA se manifeste par une dispersion des points plus homogène que les autres outils, signe que l'outil n'a pas de réel point faible mais pas non plus de point fort. Cette performance est mesurée par le meilleur résultat en macro-moyenne, tandis que les problèmes de précision en anglais notamment visibles ici permettent d'expliquer la performance plus faible capturée par la micro-moyenne.

3 Quelques problèmes liés au contenu

Nous avons montré que l'extraction est en elle-même problématique et souhaitons à présent examiner à titre d'exemple des problèmes liés aux textes trouvés par des méthodes de parcours du web. Ces dernières créent un phénomène de « pêche au chalut » en ce que le taux de remplissage ainsi que la

qualité des spécimens capturés ne se voit qu'a posteriori, pouvant nécessiter un tri. La recherche et du suivi de liens de manière indiscriminée ou non-supervisée implique que des sites comportant une forte concentration d'interliens seront « (re)pêchés » plus rapidement que d'autres. La question de l'échantillonnage renvoie à des problèmes existants qui ne feront qu'indirectement l'objet de cette étude, à travers des textes et communautés problématiques. De même, les méthodes d'optimisation des résultats d'une page sur les moteurs de recherche (*search engine optimization*) avec des textes modifiés ou même générés automatiquement ne sont pas examinés mais entrent en résonance avec les problématiques que nous décrivons. Enfin, la publicité, dissimulée ou non (sur les blogs de mode par exemple), fait également figure de limitation à l'approche opportuniste. Afin de fournir des exemples concrets des biais que l'on peut observer, nous proposons une typologie des discours de haine et des potentielles infractions en termes légaux, qui peuvent s'appliquer par extension aux textes republiés par un projet de recherche. Les exemples cités ci-dessous sont problématiques et pourraient heurter la sensibilité des lecteurs et lectrices.

Notre examen qualitatif s'appuie sur un corpus de documents en allemand établi comme base empirique au sein du projet de lexicographie DWDS (Geyken *et al.*, 2017) [6]. Ces documents web proviennent majoritairement d'Allemagne, mais aussi d'Autriche et de Suisse. Dans ce contexte, les interliens font sens pour des communautés distinctes du reste du web (par exemple les bloggeu·r·se·s de mode en Autriche) mais exposent également à des risques en déformant ou en rattachant le contenu du corpus à un genre ou un groupe précis. La propagande d'extrême-droite représente d'après nos observations une portion significative des documents problématiques. Nous n'avons pas trouvé de cas avérés contraires à la loi sur le reste du spectre politique, si ce n'est un exemple unique en son genre : une page de propagande nord-coréenne rédigée en allemand. Dans la plupart des cas nous ne proposerons pas de liens vers les documents, d'une part à cause d'un problème strictement légal (apologie de crimes) mais également pour des raisons éthiques (ne pas servir de florilège). Les documents restant après filtrage peuvent être interrogés en ligne.

3.1 Description

Appareil législatif Confronté à des abus ou irrégularités, le législateur désire souvent réguler les discours de haine. Nous pouvons évoquer les cas suivants, tous formalisés en droit : (1) par référence à l'anti-constitutionnalité de certains groupes et ce qui en découle pour les documents et les textes produits, par exemple les lois contre la propagande fasciste ; (2) en raison d'une incitation caractérisée à la haine raciale, infraction pénale depuis 1972 en droit français, loi revue en 2015 en Allemagne ; (3) des cas clairement établis en droit, comme la négation de la Shoah, la négation de la culpabilité dans les crimes commis pendant la période nationale-socialiste ou le révisionnisme ; (4) la description et apologie de la violence, notamment sous l'angle de la protection de la jeunesse : déclarations et slogans contraires aux droits de l'Homme, bellicisme.

Remises en cause systémiques Outre des slogans exploités par le personnel politique d'extrême-droite (« on n'est plus chez soi »), la remise en cause d'un soi-disant « système » est un élément central des textes problématiques du corpus, qui se font parfois injurieux, par exemple dans le cas de la critique d'un soi-disant consensus tourné vers les médias mainstream (*Mainstream-Medien*), où des « putes journaleuses » (*Medienhuren*) censureraient certaines idées.

6. https ://www.dwds.de

Sexisme et racisme Sexisme et racisme sont des corollaires parfois sous-jacents mais omniprésents des théories complotistes et des remises en cause systémiques, comme à travers des discours ouvertement xénophobes sur le soi-disant afflux ou trop-plein d'étrangers ainsi que la dénonciation d'une soi-disant « fémocratie », pouvoir féministe et injustement castrateur qu'il conviendrait d'identifier, de brider, voire d'éliminer.

La pornographie constitue un cas à part, tant il s'agit d'une industrie majeure de production et de publication de contenu, qui génère un trafic considérable. Par conséquent, toute collecte de données va trouver des hôtes hébergeant des annonces ou des vidéos, comportant nettement plus de pages web que d'autres ainsi que des liens pour favoriser le référencement. L'impact sur des corpus est réel, dans une collecte ciblant des sites utilisant WordPress, décrite dans Barbaresi (2016), *mydirtyhobby* (nom de marque) figure parmi les catégories et tags les plus fréquents des sites en `.at` (Autriche). La présence de descriptions de vidéos pornographiques dans un corpus, si elle paraît logique dans une certaine proportion au vu de la large diffusion de ces sites, a un fort impact au niveau lexical, avec un vocabulaire mais aussi un imaginaire issus du règne animal (docilité attendue des « femelles », impétuosité des « mâles ») et des concepts souvent hétérosexuels et dominateurs comme le mot composé *Dreilochstute* (« jument à trois trous »), mot quasiment absent des corpus de référence et plusieurs centaines de fois plus fréquent dans les corpus web non-filtrés.

Apologie du fascisme ou du national-socialisme L'apologie de crimes passés se fait notamment à travers des mots d'ordre de la période national-socialiste, qui servent de moyen d'identification (« *Deutschland erwache* », « *Allemagne réveille-toi* »), ce qui modifie en conséquence la teneur du corpus. Les documents connexes utilisent parfois une iconographie nationale-socialiste glorifiant les principales figures du régime qui pose un problème d'identification puisqu'elle est parfois introuvable dans le texte. En effet, l'idolâtrie par le biais d'images ne semble pas aussi strictement poursuivie.

Négationnisme Révisionnisme et négationnisme portent notamment sur une remise en cause ou une discussion du nombre de personnes enfermées et exécutées pendant la période nazie. Face à la répression de ces discours qui tombent sous le coup de la loi, les groupes concernés semblent opérer par mots-clés, comme celui de température des chambres à gaz (*Gaskammertemperatur*), concept désignant une théorie (absolument fausse) visant à exonérer la hiérarchie des camps de concentration de toute responsabilité dans l'extermination de millions de personnes.

Théories conspirationnistes Un bon indicateur de théories conspirationnistes consiste à chercher le néologisme *Reptiloiden* / reptiliens ainsi que des thématiques connexes pour débusquer des documents problématiques. Si le concept de créatures imaginaires (ici des reptiles à figure humaine) ayant pris le pouvoir ou le contrôle des gouvernants peut sembler inoffensive, il s'agit bien d'opérer une distinction entre des êtres humains et des nuisibles, catégories qui en recouvrent d'autres ou opèrent de manière souple pour rappeler d'autres théories du complot, par exemple à travers le syllogisme « X (forme attestée dans le corpus : Angela Merkel) est un reptile », « X poursuit une œuvre secrète de destruction ou d'accaparement des ressources », « les reptiles sont des juifs ».

3.2 Conséquences

Filtrer les documents problématiques sans fausser l'échantillon prélevé sur le web représente un triple problème :

éthique corpus et outils de recherche se transforment en un raccourci vers des discours d'extrême-droite, ils peuvent ou doivent être utilisés pour effectuer des signalements et constituer des corpus spécialisés utiles notamment en sciences politiques ou en sociologie ;

légal certaines pages tombent clairement sous le coup de la loi et nécessitent un dépistage et une intervention immédiate, d'autres pas nécessairement tout en présentant un risque difficilement appréciable pour des néophytes ;

linguistique un compromis en forme de « ni-ni » paraît adéquat : ni conserver en l'état, ni supprimer tous les documents ou les occurrences. Malgré la distorsion des corpus, garder des échantillons portant trace de différents types de discours permet d'offrir une perspective large et à l'image de l'époque sur les modes d'expression en ligne.

4 Conclusions

La collecte et l'usage de données web sont sujets à une série de problèmes éthiques, méthodologiques et épistémologiques qui méritent l'attention de la communauté scientifique. Il appert que les approches opportunistes présidant à l'établissement de grands corpus tirés du web ne sont pas sans poser un certain nombre de difficultés. Nous avons apporté des preuves empiriques de leur impact, tout d'abord en étudiant la forme des documents obtenus à travers la comparaison de méthodes d'extraction des données et ensuite en recensant des problèmes centrés sur le contenu des corpus et liés aux méthodes d'acquisition opportuniste des données. La faible supervision conduit à un *far west*, *« Wild West Web Crawling »* selon Jo & Gebru (2020), tandis qu'une approche plus supervisée et maîtrisée ne suffit pas à résoudre des problèmes posés par l'extraction de texte.

Au phénomène de dispersion des segments textuels visible sur les graphiques d'évaluation répond une probabilité élevée de cerner certaines communautés (hobby précis ou frange politique) et genres textuels (petites annonces et annuaires). Sur la forme, les corpus web peuvent receler des documents incomplets et tronqués ainsi que des doublons et des segments génériques, dans une proportion variable qui pourrait bien être inconnue ou mal estimée par la communauté scientifique. Par ailleurs, des problèmes de fond substantiels peuvent surgir. Des textes ou éléments indésirables se trouvent dans des corpus destinés à la recherche en linguistique et en TAL, d'une part à cause de l'impossible contrôle des sources et adaptation à certains types de pages dès que la taille du corpus atteint un certain ordre de grandeur, et d'autre part en raison de l'application d'outils génériques et supposés adéquats sans vérification de leur efficacité pour des textes, langues ou sujets divergents, problématique connue en apprentissage artificiel par la notion d'adaptation de domaine.

Il faudrait pouvoir non seulement décrire ces problèmes mais également les circonscrire, ce qui implique de trouver des méthodes de mesure ainsi que des heuristiques de limitation. Alors que le détourage peut être évalué et résolu par des approches quantitatives (comprenant étalons et métriques), les difficultés d'ordre qualitatif sont plus difficile à cerner et à étalonner, alors même que leurs potentielles conséquences éthiques voire pénales sont plus graves. La « corne d'abondance » représentée par la collecte de données massives à coût moindre semble bien réelle mais est en réalité assujettie à un examen approfondi en termes de calibrage et d'équilibrage afin de constituer une nécessaire assise scientifique et de dépasser les logiques opportunistes.

Références

BARBARESI A. (2015). *Ad hoc and general-purpose corpus construction from web sources*. Thèse de doctorat, École Normale Supérieure de Lyon.

BARBARESI A. (2016). Efficient construction of metadata-enhanced web corpora. In P. COOK, S. EVERT, R. SCHÄFER & E. STEMLE, Éds., *Proceedings of the 10th Web as Corpus Workshop*, p. 7–16 : Association for Computational Linguistics.

BARBARESI A. (2019). Generic Web Content Extraction with Open-Source Software. In *Proceedings of KONVENS 2019, Kaleidoscope Abstracts*, p. 267–268 : GSCL.

BARBARESI A. & LEJEUNE G. (2020). Out-of-the-Box and Into the Ditch ? Multilingual Evaluation of Generic Text Extraction Tools. In *Proceedings of the 12th Web as Corpus workshop (WAC-XII)* : ELRA. à paraître.

BARONI M., CHANTREE F., KILGARRIFF A. & SHAROFF S. (2008). Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*, p. 638–643 : ELRA.

CALBERAC Y. (2010). *Terrains de géographes, géographes de terrain. Communauté et imaginaire disciplinaires au miroir des pratiques de terrain des géographes français du XXe siècle*. Thèse de doctorat, Université Lumière Lyon 2.

CALISKAN A., BRYSON J. J. & NARAYANAN A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

GEYKEN A., BARBARESI A., DIDAKOWSKI J., JURISH B., WIEGAND F. & LEMNITZER L. (2017). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, **45**(2), 327–344.

HAMBORG F., MEUSCHKE N., BREITINGER C. & GIPP B. (2017). news-please : A generic news crawler and extractor. In M. GAEDE, V. TRKULJA & V. PETRA, Éds., *Proceedings of the 15th International Symposium of Information Science*, p. 218–223.

JO E. S. & GEBRU T. (2020). Lessons from Archives : Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 306–316.

KOHLSCHÜTTER C., FANKHAUSER P. & NEJDL W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, p. 441–450.

LEJEUNE G. & BARBARESI A. (2020). Bien choisir son outil d'extraction de contenu à partir du Web. In *Actes de la conférence JEP-TALN-RECITAL 2020, Démonstrations* : ATALA. à paraître.

LEJEUNE G. & ZHU L. (2018). A New Proposal for Evaluating Web Page Cleaning Tools. *Computación y Sistemas*, **22**(4).

MCENERY T. & HARDIE A. (2011). *Corpus linguistics : Method, theory and practice*. Cambridge University Press.

OLSTON C. & NAJORK M. (2010). Web Crawling. *Foundations and Trends in Information Retrieval*, **4**(3), 175–246.

PETERS M. E. & LECOCQ D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22nd International Conference on World Wide Web*, p. 89–90.

POMIKÁLEK J. (2011). *Removing boilerplate and duplicate content from web corpora*. Thèse de doctorat, Masaryk University.

RATCLIFF J. W. & METZENER D. E. (1988). Pattern Matching : The Gestalt Approach. *Dr. Dobb's Journal*, **13**(7), 46.

SCHÄFER R., BARBARESI A. & BILDHAUER F. (2013). The Good, the Bad, and the Hazy : Design Decisions in Web Corpus Construction. In *Proceedings of the 8th Web as Corpus Workshop*, p. 7–15.

SMITH J., SAINT-AMAND H., PLAMADĂ M., KOEHN P., CALLISON-BURCH C. & LOPEZ A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, p. 1374–1383.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p. 9–16.

TANGUY L. (2013). La ruée linguistique vers le Web. *Texto ! Textes et Cultures*, **18**(4).

VALETTE M. (2008). Pour une science des textes instrumentée. *Syntaxe et sémantique*, **9**, 9–14.

WENINGER T., PALACIOS R., CRESCENZI V., GOTTRON T. & MERIALDO P. (2016). Web Content Extraction : A Meta-Analysis of Its Past and Thoughts on Its Future. *SIGKDD Explorations Newsletter*, **17**(2), 17–23. DOI : 10.1145/2897350.2897353.