Le corpus WikiDisc : ressource pour la caractérisation des discussions en ligne

Lydia-Mai Ho-Dac, Veronika Laippala

LE CORPUS WIKIDISC : RESSOURCE POUR LA CARACTÉRISATION DES DISCUSSIONS EN LIGNE

Lydia-Mai HO-DAC, CLLE, University of Toulouse, CNRS, UT2J, France
Veronika LAIPPALA, TIAS, University of Turku, Finland

L'immense quantité de textes et de variation langagière disponible sur internet fait du Web une source de données importante pour la linguistique et pour le traitement automatique des langues (cf. Tanguy 2013, Hundt et al. (Eds.) 2007). De nombreux corpus collectés automatiquement voient le jour (e.g. Baroni et al. 2009) et d'énormes collections de données, composées de milliards de mots, sont utilisées, échangées et discutées par la communauté scientifique (cf. la série des *WaC– Web **as** Corpus Workshop*[1]). Cependant, ces données moissonnées sur le Web présentent un certain nombre d'inconvénients qui ne favorisent pas leur traitement et leur analyse en tant que « corpus » (cf. Tanguy 2013). Outre les problèmes liés à la présence massive de textes produits par des robots (génération et traduction automatique), l'inconvénient principal est la difficulté à caractériser les textes collectés. L'enjeu actuel est donc de proposer des techniques visant à mieux caractériser voire profiler les textes du Web afin d'en permettre la sélection et, à terme, la constitution de corpus représentatifs d'usages langagiers identifiés, dans l'idée du *Web **for** Corpus*. Dans ce contexte, Wikipédia offre une source de données qui semble combiner les avantages des grands ensembles moissonnés automatiquement et des collections de textes plus homogènes et restreintes : grande quantité de données et situations de communication limitées et largement renseignées, incluant des articles encyclopédiques, des discussions autour de la rédaction des articles, des forums de discussions autour du projet Wikipédia (la version française parle de « bistrots » et « cafés »), des Journaux et chats d'activité (e.g. « Bulletin des patrouilleurs »).

Cet article décrit le corpus WikiDisc qui fournit une version XML encodée selon la TEI-P5 de toutes les discussions associées à un article de Wikipédia version française en date du 12/05/2015. Ce corpus sera mis à disposition de la communauté à l'automne 2016 via le portail des ressources développées au laboratoire CLLE-ERSS, Unité Mixte de Recherche CNRS et Université de Toulouse Jean Jaurès[2]. Cet article propose un premier état des lieux de la ressource suivi d'une présentation des résultats d'un ensemble d'analyses contrastives mettant au jour un certain nombre de caractéristiques lexicales et syntaxiques des discussions Wikipédia.

1. WIKIPÉDIA « AS » ET « FOR » CORPUS

Wikipédia est une encyclopédie libre et coopérative qui existe depuis 2001 et à laquelle tout internaute peut contribuer en modifiant ou créant un article ou encore en postant un message dans une page de discussion portant sur la structure, la pertinence, le contenu de l'article. Cette communauté fonctionne par le travail des internautes actifs qui sont amenés à acquérir différents statuts comme celui de « patrouilleur » dont le rôle est d'alerter sur des actes de vandalisme, ou encore celui d'« administrateur » dont le rôle est de « protéger et maintenir la qualité des éditions du projet »[3]. Ces rôles participent tous à la modération de Wikipédia qui

[1] https://www.sigwac.org.uk/
[2] http://redac.univ-tlse2.fr
[3] La page http://fr.wikipedia.org/wiki/Aide:Statuts_des_utilisateurs fournit une liste détaillée des statuts des

consiste, comme pour tout forum de discussion, à décider de la publication ou non d'un ajout ou d'une modification que ce soit dans un article ou une discussion.

De nombreux travaux en linguistique et traitement automatique des langues (TAL) s'intéressent à Wikipédia avec une priorité donnée aux articles, notamment pour l'extraction de connaissance (Zech et al. 2008, Medelyna et al. 2009), l'accès à des ressources multilingues (Kilgariff et al. 2010) ou encore l'étude de la construction d'un article via l'historique des révisions (Ferschke et al. 2013). Plus récemment, Wikipédia intéresse la communauté scientifique pour ses pages de discussion qui se cachent derrière chaque article (Ferschke et al. 2012). Ferschke et al. (2013) listent plusieurs aspects qui attirent les chercheurs vers l'analyse de ces discussions : étude du travail collaboratif et de la résolution de conflits, mise au jour d'indicateurs pour qualifier la qualité de l'information ou la crédibilité de certains locuteurs, étude des réseaux sociaux qui se dessinent entre des contributeurs aux statuts délimités. Côté français, des travaux récents portent sur la détection de conflits dans les discussions de Wikipédia (Denis et al. 2012, Poudat et al. 2015). La section suivante revient sur les arguments qui justifient l'intérêt d'un corpus constitué de ces discussions pour les linguistiques de corpus.

1.1. UN CORPUS DE DISCUSSIONS WIKIPÉDIA

Un corpus constitué de discussions Wikipédia représente un nombre important d'intérêts pour les linguistiques de corpus. Premièrement, ces discussions diffusées sous licence Creative Commons by-sa sont libres, contrairement à la majorité de forums internet. Ensuite, leur contexte de production est beaucoup plus renseigné que pour tout autre forum de discussion, étant donné que chaque discussion porte sur une thématique explicite et détaillée dans l'article associé. Les discussions Wikipédia peuvent ainsi être systématiquement définies par un nombre important de méta-données portant à la fois sur la thématique (portail thématique, article associé), le caractère subjectif de la discussion (caractère polémique, etc.) et le statut du locuteur (informations sur ses contributions et sur son statut dans la communauté). Ces statuts ont fait l'objet d'un certain nombre d'études sur la corrélation entre le statut et le style langagier utilisé (Danescu-Niculescu-Mizli et al. 2012, 2013, Burke & Kraut 2008 inter alia).

Autre point important, notamment pour l'application de techniques en TAL, les discussions Wikipédia semblent présenter relativement peu de déviance par rapport à la norme langagière (cf. Baldwin et al. 2013). Les messages y seraient écrits de manière plus « normée » avec un langage plus formel que dans les traditionnels forums de discussion avec relativement peu de fautes d'orthographe et de grammaire, peu de recours à des modes de rédaction particuliers (lettres capitales PLUS, répétées ASSSSEEEZ, suite de ponctuation répétées !!!!, émoticônes, écriture non accentuées) et peu de vulgarité.

Ces trois avantages (textes libres, relativement bien renseignés, présentant un niveau d'écriture correct) font des discussions Wikipédia un terrain

utilisateurs de Wikipédia version française.

d'expérimentation indéniable pour, entre autres, caractériser un certain type de communication médiée par les réseaux (CMR).

1.2. CONSTITUTION DU CORPUS WIKIDISC

Afin de constituer le corpus WikiDisc, plusieurs procédures automatiques ont été mises en place pour extraire et formater les discussions. L'extraction consiste à traiter le *dump*[4], d'y sélectionner les discussions présentant un contenu textuel et de les transformer en fichiers XML normés selon la TEI-P5. Le *dump* utilisé contient en tout 3 487 480discussions (repérables par l'expression régulière /<title>Discussion/) parmi lesquelles seules 1 496 553 portent sur un article[5]. 76 % de ces discussions ont été écartées du corpus WikiDisc. Le tableau 1 donne le détail de la sélection effectuée.

Discussions portant sur un article	**1 496 553**	
Discussions redirigées vers une autre discussion	116 432	8 %
Discussions vides ou contenant moins de 2 mots	1 013 791	68 %
Discussions retenues	**365 612**	**24 %**

Tableau 1 : Sélection des discussions composant le corpus WikiDisc

Chaque discussion a été automatiquement formatée en XML selon la norme TEI-P5 et associée à un certain nombre de méta-données. Les méta-données collectées sont l'identifiant Wikipédia de la discussion tel qu'indiqué dans le *dump* et les informations contextuelles données dans le bandeau de la discussion. Ces informations, encodées dans l'élément `classDecl` du `teiHeader`, concernent le(s) portail(s) associés à la discussion et l'article (`category type='discipline'`), le niveau d'avancement de l'article associé (`category type='avancement'`), l'indication d'un appel au calme (`category type='interaction'`) et d'autres informations potentielles comme l'indication d'une traduction (`category type='autre'`).

La délimitation des différents fils, messages et contributeurs s'appuie sur un ensemble de règles, notamment la présence nécessaire de la date de publication du message et la détection d'une différence de niveau. Les éléments de la TEI-P5 retenus pour structurer le corps des discussions sont `<div>`pour les fils, `<head>` pour les titres des fils, et pour chaque message :

```
<post who='nom d'utilisateur' bot='yes/no' when='date de plublication' interactionalLevel='#'>
```

`@bot` indique si le nom de l'utilisateur fait partie ou pas de la liste des robots Wikipédia (scripts mis en place pour des procédures de maintenance). `@interactionalLevel` indique le niveau d'enchâssement du message dans le fil

[4]Un *dump* est une sauvegarde globale de Wikipédia. La version utilisée pour le corpus WikiDisc est la version contenant toutes les pages courantes (sans les historiques de révision) à la date du 12 mai 2015 (frwiki-20150512-pages-meta-current#.xml.bz2) diffusée librement sur la page http://dumps.wikimedia.org/frwiki/20140331/

[5]1 990 927 sont des discussions portant sur un utilisateur.

de discussion. Chaque message est découpé en paragraphe <p> ou éléments de liste <item>.

La figure 1 fournit un extrait de discussion selon cet encodage.

```
    </post>
  </div>
 −<div id="3" level="1">
    <head>au sujet du miel artificiel</head>
  −<post id="4" who="anonyme" bot="no" when="unknown" interactionalLevel="0">
    −<p id="1">
      http://www.uni-regensburg.de/Fakultaeten/nat_Fak_IV/Organische_Chemie/Didaktik
      /Keusch/D-art_honey-e.htm
     </p>
    −<p id="2">
      Il y est expliqué comment produire une sorte de miel artificiel c-à-d par hydrolyse
      en milieux acide acide citrique) du sucre de bettrave (saccharose):
     </p>
     <p id="3"> *"'Hydrolyse"':</p>
    −<p id="4">
      saccharose +H3O+ milieu acide ---- fructose+ glucose (pseudo-miel)
     </p>
    </post>
  </div>
 −<div id="4" level="1">
  −<head>
     À quoi sert aux abeilles le miel qu'elles fabriquent ?
    </head>
  −<post id="5" who="Aeleftherios" bot="no" when="04-07-2006-22:13"
    interactionalLevel="1">
    −<p id="1">
      D'après le premier paragraphe (à la fin), à avoir une réserve de nourriture pour
      l'hiver. --Aeleftherios 4 juillet 2006 à 22:13 (CEST)
     </p>
    </post>
  −<post id="6" who="Gyp" bot="no" when="06-07-2006-18:06"
    interactionalLevel="2">
    −<p id="1">
      Cette information ne devrait-elle pas se trouver en début d'article ? C'est la base,
      non ? --Gyp 6 juillet 2006 à 18:06 (CEST)
     </p>
    </post>
  </div>
```

Figure 1 : Extrait du corpus WikiDisc encodé selon la norme TEI-P5 (discussion associée à l'article « Miel »)

Une évaluation manuelle de 8 discussions comptabilisant 413 messages et 47 284 mots a montré une précision de 0,92 (3 messages vides ; 5 messages scindés en 2 ; 25 messages fusionnant 2 ou 3 messages) et un rappel de 0,95 (23 messages absents). Les discussions retenues pour cette évaluation ont été sélectionnées pour représenter des discussions de taille variée en termes de nombre de fils, nombre de messages et nombre de mots.

La valeur des attributs @who, @bot, @when et @interactionalLevel n'a été vérifiée que pour la discussion portant sur l'article « L'affaire bogdanoff » par comparaison avec la version présente dans le corpus WikiConflits (Poudat et al. 2015) pour laquelle une correction manuelle de ces informations a été réalisée.

2. CARACTÉRISTIQUES GLOBALES DU CORPUS

Avant de proposer les analyses permettant de mettre au jour des caractéristiques lexicales et syntaxiques du corpus WikiDisc, nous proposons un premier état des lieux de la ressource. Cet état des lieux consiste à dresser un panorama quantitatif du contenu des discussions et à proposer des points de

comparaison pour situer le corpus constitué par rapport à d'autres types de CMR.

Le corpus WikiDisc compte 365 612 discussions contenant 1 023 841 fils, 2 406 514 messages et 161 833 298 mots. Sur ce large ensemble, seul un tiers pourrait se révéler pertinent ; en effet, seules 30 % des discussions contiennent plus de 3 messages et plus de 215 mots (54% ne contiennent qu'un message et 50% moins de 54 mots). Une même tendance se dessine quant au nombre de locuteurs en interaction avec 62 % des discussions qui semblent n'impliquer pas plus de deux locuteurs différents (150 603 discussions mono-contributeurs où un locuteur lance un sujet auquel personne ne répond et 77 766 dialogues)[6]. A l'opposé, certaines discussions apparaissent très prolifiques, avec des records atteignant les 1 220 messages (discussion liée à l'article « D. Strauss-Kahn », archive 2, idno 5486756), 150 421 mots (discussion liée à l'article « Opposition au mariage homosexuel en France », partie 1, idno 7298738) et 225 locuteurs différents (discussion autour de l'admissibilité de la page « Mickaël Vendetta »[7]). La proposition de messages postés par des locuteurs anonymes (i.e. qui n'ont pas signés avec un nom d'utilisateur Wikipédia) est très élevée, avec une moyenne de 80 % des contributions.

Afin de situer le corpus WikiDisc parmi les CMR, nous proposons une première étude contrastive qui oppose les discussions (1) aux articles Wikipédia, textes expositifs présentant un niveau d'écriture *a priori* plus normé et formel ; et (2) à des forums de santé (CancerDuSein.org et Enceinte.com collectés dans le cadre du projet inter-MSH *Patients' mind*[8]) qui représentent un type de CMR plus proche des discussions, mais avec un caractère sans doute moins normé en terme de niveau de langage et d'écriture.

Notre comparaison s'est effectuée relativement à deux types de caractéristiques généralement associées aux CMR. Le premier type relève d'indices qui peuvent être interprétés comme révélateurs du niveau de dégradation de l'écriture de textes issus du Web (Baldwin et al. 2008). Nous utilisons ici le taux de mots inconnus, la longueur moyenne des phrases et des mots. La deuxième série d'indices concerne les traces de subjectivité traditionnellement utilisées pour détecter automatiquement des textes contenant des opinions. Nous comptabilisons pour cette étape le pourcentage de noms, adjectifs et verbes inclus dans le lexique des *affects* développé par Augustyn et al. (2008). Cette mesure est complétée par la fréquence relative (sur 100 mots) des pronoms de première personne - Pro1 (*nous, je, j', me, m'*) en position sujet ou objet. Le tableau2 donne les résultats pour les trois corpus envisagés. L'ensemble des résultats a été calculé sur la base de l'analyse syntaxique réalisée par le logiciel Talismane (Urieli 2013, Urieli & Tanguy 2013)[9].

[6]Le nombre de locuteurs différents se base sur une détection automatique des pseudonymes des utilisateurs avec un pseudo « anonyme » attribué uniformément à tous les utilisateurs non déclarés. Cette détection automatique n'a pas été évaluée ce qui explique le terme 'tendance' utilisé ici.

[7]**Nous faisons référence ici à une discussion archivée dont la thématique portait sur la légitimité d'un article Wikipedia sur Mickaël Vendetta, participant d'une émission de télé-réalité. L'enjeu du débat était de décider si l'article proposé avait un but encyclopédique ou commercial. La discussion peut être consultée en ligne à l'url : https://fr.wikipedia.org/wiki/Discussion:Micka%C3%ABl_Vendetta/Suppression**

[8]https://www.lirmm.fr/patient-mind

[9]**L'analyse a été réalisée en utilisant la plateforme OSIRIM qui est administrée par l'IRIT et soutenue par**

	Nb Mots	mots inconnus (%)	Longueur moyenne phrases	mots	Lexique d'« *affect* » (%)	Pro1 (%)
Forum de santé	236 368 151	22	10,3	5,2	4,7	3,48
WikiDisc	161 833 298	5	18,2	5,4	2,1	0,05
Articles Wikipédia	226 207 672	5	14,6	5,5	1,8	1,34
	622 154 102	**12**	**13,4**	**5,4**	**2,8**	**1,53**

Tableau 2 : Caractérisation contrastive du corpus WikiDisc

Ce premier état des lieux montre une certaine stabilité en termes de longueur de mots mais des variations au niveau de la proportion de mots « inconnus » et de la longueur des phrases. Les mots « inconnus » regroupent tous les noms communs, verbes, adjectifs et adverbes que Talismane n'a pas réussi à lemmatiser. Comme attendu, cette proportion est très importante pour les forums de santé qui présentent *a priori* beaucoup de mots mal orthographiés au contraire des articles et discussions Wikipédia. Ajoutons à ces chiffres que sur les 560 841 formes inconnues repérées dans les discussions, 104 687 se retrouvent également dans les articles Wikipédia, ce qui suppose que ces 19 % de mots inconnus ne soient pas nécessairement des erreurs d'orthographe. De plus, en observant la fréquence relative de deux fautes d'orthographe courantes (*jai* et *j* comptés dans les inconnus), la différence de niveau d'écriture apparaît clairement, avec 11,3 *jai* et *j* sur 10 000 mots dans les forums de santé contre 0,2 dans les discussions Wikipédia.

Ces résultats vont dans le sens de l'hypothèse d'un certain niveau d'écriture chez les Wikipédiens, tant dans la rédaction des articles que des discussions, ce qui se retrouve également dans les résultats de Baldwin et al. (2013) qui, pour l'anglais, ont mesuré ces mêmes indices dans des posts de forums et de blogs, des commentaires « youtube », des *tweets*, des articles Wikipédia et des extraits du BNC (British National Corpus). Concernant la longueur des phrases, Baldwin et al. (2013) associe également cet indice à un langage plutôt formel, ce qui conforte là encore les différences observées au niveau du pourcentage de mots inconnus. Le taux de termes exprimant un *affect* montre également un record dans les forums de santé. Notons cependant que cette mesure nécessite une analyse plus fine des indices projetés qui paraissent encore très « brutaux » si l'on observe par exemple les *affect* les plus fréquents pour chaque sous-corpus :

- *attendre* et *beau* dans les forums de santé,
- *croire* et *demander* dans les discussions Wikipédia,
- *jouer* et *considérer* dans les articles Wikipédia.

Malgré ces réserves, la tendance observée est confortée par la fréquence relative des pronoms de première personne relevés.

Afin de compléter ce premier état des lieux, nous proposons dans la section suivante une caractérisation qui adopte une approche complémentaire capable de

CNRS, la région Midi-Pyrénées, le gouvernement français, et le FEDER (voir http://osirim.irit.fr/site/fr).

mettre au jour de façon inductive les caractéristiques lexicales et surtout syntaxiques des discussions Wikipédia.

3. CARACTÉRISTIQUES LEXICALES ET SYNTAXIQUES

Cette section propose une description complémentaire du corpus WikiDisc qui s'inscrit dans une approche inductive (*data-driven*),sans connaissances ou décisions *a priori* des traits examinés. Cette démarche qui permet de faire émerger les traits lexicaux et syntaxiques caractéristiques du corpus se justifie également par l'objectif à plus long terme de notre étude qui vise la mise en place d'outils pour profiler et classifier les textes du Web.

La caractérisation lexicale d'un corpus relève généralement du calcul des spécificités lexicales, méthode classique pour l'examen du style et des thèmes d'un corpus (Scott & Tribble 2006). Le calcul des spécificités présente l'avantage de faire émerger rapidement et facilement les mots individuels significativement sur- et sous-représentés dans un corpus par rapport à un corpus de comparaison. Un des inconvénients de cette méthode est qu'elle permet difficilement d'aller au delà de la description thématique d'un corpus, ce qui pose problème dans le cas où le corpus à décrire présente une variété de thématiques (Laippala et al. 2015a, 2015b). Pour cette raison, afin de faire émerger des caractéristiques plus génériques, nous proposons de nous intéresser à la fois aux caractéristiques lexicales et syntaxiques.

Afin de mettre au jour les caractéristiques lexicales et syntaxiques typiques des discussions Wikipédia, nous proposons de tirer partie des résultats d'une tâche de classification automatique avec analyse des traits spécifiques jugés utiles par le classifieur. Même si le but de cette étude n'est pas de développer un classifieur, nous avons choisi cette méthode pour plusieurs raisons. Premièrement, c'est une méthode qui permet de mettre au jour les "caractéristiques" des différents sous-corpus, c'est-à-dire les traits qui permettent de classifier automatiquement les différents sous-corpus. Par conséquent et à l'inverse des méthodes classiques de calcul des spécificités (Paquot & Bestgen 2009), les traits inutilisables par le classifieur car trop rares et peu partagés entre les textes du corpus ne ressortent pas. Par ailleurs, notre objectif étant à plus long terme la classification du Web, il était important de s'assurer que les discussions Wikipédia pouvaient être distingués automatiquement des articles et des forums de santé.

Les caractéristiques décrites ici sont observées en comparant les discussions Wikipédia aux deux corpus de comparaison présentés dans la section précédente. Cette double comparaison – articles encyclopédiques d’une part et discussions en ligne dans un autre domaine d'autre part – permet un regard croisé sur les discussions Wikipédia et réduit de ce fait l'influence de la nature des corpus de comparaison sur l'interprétation des résultats (Scott & Tribble 2006 : 63-65).

La classification a été réalisée deux par deux (discussions vs. articles Wikipédia ; discussions Wikipédia vs. forums de santé) avec le classifieur linéaire

Vowpal Wabbit (Agarwal et al. 2011), entraîné sur la moitié des corpus. La mission du classifieur est d'attribuer à chaque segment de trois phrases la catégorie « wikiDisc » ou « autre ». Le choix de segmentation répond uniquement à des critères techniques, une prochaine expérimentation en cours vise à tester l'unité paragraphe, plus justifiée linguistiquement. Les sections suivantes présentent, pour la section « caractéristiques lexicales » les résultats du classifieur entraîné sur les corpus non analysés syntaxiquement et réduits par l'outil à de simples « sacs de mots » ; pour la section « caractéristiques syntaxiques » les résultats du classifieur entraîné sur les corpus analysés syntaxiquement et considérés par l'outil comme des « sacs de n-grams syntaxiques ».

3.1. CARACTÉRISTIQUES LEXICALES

Le tableau 3 présente les résultats du classifieur paramétré « sac de mots » pour les tâches de classification WikiDisc vs. articles Wikipédia et WikiDisc vs. forums de santé. Pour chaque comparaison, la F-mesure (F1, moyenne harmonique du rappel et de la précision) indique la performance de la classification et la colonne « caractéristiques lexicales » liste les traits lexicaux les plus typiques, c'est-à-dire les traits les plus utiles au classifieur pour distinguer les discussions des autres textes.

WikiDisc vs.	**Caractéristiqueslexicales**
articles Wikipédia **F1**=0,94	*cest, bonjour, merci, paragraphe, faudrait, !!, cordialement, bloquer, scientologie, wallonie, ???, article, wikipédia, qatari, œuvre*
forums de santé **F1**=1	*article, Wikipédia, sources, seigneur, fusion, source, peuple, supprimé, habitants, références, anonyme, CET, articles, bibliographie, définition*

Tableau 3 : Traits lexicaux typiques pour la classification « sac de mots » discussions vs. articles Wikipédia et discussions Wikipédia vs. forums de santé

Les F-mesures(F1) indiquent que la tâche de classification est de façon globale « facile » pour Vowpal Wabbit et plus performante pour la classification discussions Wikipédia vs. forum de santé que discussions vs. articles. Concernant les traits lexicaux relevés, les deux groupes de mots reflètent des aspects typiques du processus de rédaction des articles (e.g. *article, wikipédia, sources, supprimé, fusion, bloquer*). La comparaison aux forums de santé, met également en avant des termes relatifs à la politique éditoriale de Wikipédia pour assurer la rédaction d'articles de qualité : *références, définition, sources.* On voit apparaître des thèmes que l'on trouve également dans les articles encyclopédiques : *qatari, scientologie, habitants.*

La classification discussions vs. articles fait apparaître des traits lexicaux plus en lien avec le genre « discussion » : des salutations (*bonjour, merci, cordialement*), des jeux de ponctuation (*!!, ???*) caractéristiques des discussions en ligne (Herring 2012). Leur présence, qui contredit quelque peu les observations sur le langage plutôt formel des discussions Wikipédia, s'explique par le point de comparaison : les jeux de ponctuation ne se retrouvent absolument pas dans les articles Wikipédia.

3.2. CARACTÉRISTIQUES SYNTAXIQUES

Pour analyser les caractéristiques syntaxiques de notre corpus de discussion Wikipédia, nous avons transformé les données en n-grams syntaxiques délexicalisés qui correspondent à des sous-arbres des analyses en dépendances dont l'information lexicale est supprimée (Goldberg & Orwant 2013, Kanerva et al. 2014). Plus précisément, nous avons utilisé des tri-arcscomposés de quatre nœuds i.e. quatre mots, et trois arcs i.e. trois relations de dépendance (Laippala et al. 2015b). Nos travaux antérieurs ([deleted for review]) montrent que ces constructions sont plus robustes que les mots individuels pour caractériser des genres et qu'elles s'adaptent mieux à la variation thématique et situationnelle des textes.

Les tableaux 4 et 5 fournissent pour chaque comparaison la F-mesure et une sélection des caractéristiques syntaxiques des discussions Wikipédia. Vu que les n-grams incluent une analyse morphologique très détaillée, plusieurs décrivent des constructions syntaxiques similaires. La sélection proposée dans les tableaux a été obtenue en analysant les 25 n-grams les plus typiques selon le classifieur[10] et concerne les cas les plus fréquents qui nous ont paru les plus motivés linguistiquement. Pour chaque n-gram, un exemple lexicalisé permet un aperçu du type de structure que le n-gram représente.

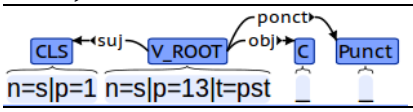
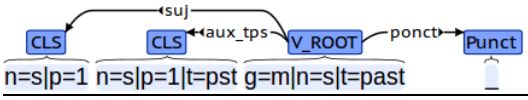
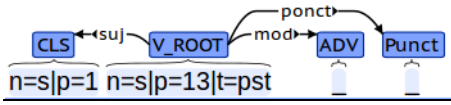
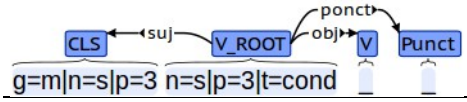
Caractéristiques syntaxiques **F1 = 0,86**	**Exemples lexicalisés**
	Je + verbe + que + ponct *Je pense que […].* *J'ajoute que […] .*
	Je + aux + verbe + ponct *J'ai profité […] .* *J'ai retiré […] .*
	Je + verbe + adverbe + ponct *Je ne pense pas […] .* *Je me retrouve donc […].*
	Il + V. conditionnel + ponct *Il vaudrait mieux[…] .* *Il faudrait expliciter […].*

Tableau 4 : Traits syntaxiques typiques pour la classification « sac de n-grams syntaxiques » des discussions vs. articles Wikipédia

Le premier constat concerne la performance du classifieur qui est moins bonne qu'avec l'approche « sac de mots » : F-mesure de 0,86 et 0,91 contre 0,94 et 1 précédemment. Malgré ce score qui reste néanmoins correct, les traits syntaxiques retenus apportent un nouveau regard sur la caractérisation des discussions

[10]Comme précédemment, les traits « typiques » correspondent aux traits les plus utiles au classifieur pour distinguer les discussions des autres textes.

Wikipédia qui se distinguent des articles Wikipédia par une plus forte présence de verbes à la première personne et de constructions impersonnelles avec verbe au conditionnel (tableau 4). Ces aspects reflètent surtout la situation communicationnelle des discussions Wikipédia : les scripteurs écrivent pour exprimer leur opinion sur les modifications à faire dans l'article et pour rapporter des changements effectués. Les n-grams retenus montrent également une tendance dans les discussions aux phrases complètes avec un verbe fini, ce qui indiquerait que le style télégraphique souvent associé aux discussions en ligne informelles (e.g. omissions des auxiliaires et verbes conjugués, cf. Herring 2012) ne se retrouve pas dans les discussions Wikipédia, ce qui rejoint notre premier état des lieux.

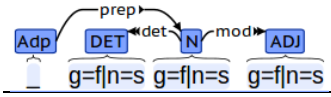
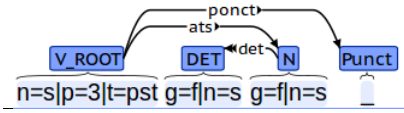
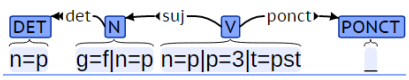
Spécificités syntaxiques **F1= 0,91**	**Exemples lexicalisés**
	adposition + det+ nom + adjectif *[...] dans leur version initiale [...]* *[...] de la musique occidentale [...]*
	verbe + det + nom + ponct *[...] c'est une légende [...]* *[...] ce n'est pas une bonne idée [...]*
	det + nom + verbe + ponct *[...] les erreurs rapportées sont :[...]* *Les races descendent [...] .*

Tableau 5 : Traits syntaxiques typiques pour la classification « sac de n-grams syntaxiques » des discussions Wikipédia vs. forums de santé

Concernant les caractéristiques syntaxiques permettant de distinguer les discussions Wikipédia des discussions de forums de santé (tableau 5), les éléments relevés reflètent également une syntaxe complète (syntagmes prépositionnels, déterminants complexes, sujet+verbe, constructions attributives complètes). Contrairement aux n-grams observés lors de la comparaison avec les articles Wikipédia, nous ne retrouvons pas les indices de première personne, également très fréquents dans les forums de santé, comme nous l'avions également observé dans notre premier état des lieux.

4. CONCLUSION

Dans cet article, nous avons présenté le corpus WikiDisc, une collection de toutes les discussions associées à un article de Wikipédia dans sa version française, comportant près de 161 millions de mots rassemblés dans 365 612 discussions et 2 406 514 messages. Le corpus sera mis à disposition de la communauté dès l'automne 2016. Chaque discussion est formatée en XML selon la norme TEI-P5 et associée à un certain nombre de méta-données facilitant son analyse et traitement automatique (portail thématique, état d'avancement de l'article associé, indication d'un appel au calme si la discussion a été jugée trop polémique).

Afin de profiler les discussions, nous les avons comparées à deux corpus : des discussions en ligne issues des forums de santé et des articles encyclopédiques de Wikipédia. Notre méthode de comparaison repose à la fois sur une analyse orientée par des caractéristiques linguistiques généralement associées aux CMR (approche déductive) et sur une analyse guidée par les données (approche inductive)tirant partie des résultats d'un classifieur automatique et mettant en avant les caractéristiques lexicales et syntaxiques des discussions. Nos résultats vont dans le même sens que ceux obtenus par les études antérieures portant sur Wikipédia, et ce quelle que soit la langue étudiée. Les discussions Wikipédia présentent un certain niveau d'écriture et de langage qui contraste avec les caractéristiques souvent associées aux CMR (forte proportion de termes subjectifs, déviances orthographiques et style télégraphique). Seuls les jeux de ponctuation typiques des CMR semblent persister dans les discussions de notre corpus WikiDisc. Concernant les caractéristiques lexicales, outre celles associées aux thématiques des articles encyclopédiques, nos résultats indiquent l'utilisation d'un lexique propre au processus de rédaction des articles et de la politique éditoriale de Wikipédia ; et également des traces de l'expression de l'opinion du scripteur et la description des changements effectués dans les articles.

RÉFÉRENCES BIBLIOGRAPHIQUES

Agarwal, Alekh, Chappelle, Olivier, Dudik, Miroslav & Langford, John (2011): "A Reliable Effective Terascale Linear Learning System", *JMLR* 15, 1111-1133.

Baldwin, Thimoty, Cook, Paul, Lui, Marco, MacKinlay, Andrew & Wang, Li. (2013): "How Noisy Social Media Text, How Different Social Media Sources?", in *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan, 356-364.

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), 209-226.

Burke, Moira & Kraut, Robert (2008): "Mopping up: modeling wikipedia promotion decisions", in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 27-36. ACM.

Danescu-Niculescu-Mizil, Christian, West, Robert, Jurafsky, Dan, Leskovec, Jure & Potts, Christopher (2013): "No country for old members: User lifecycle and linguistic change in online communities", in *Proceedings of the 22nd international conference on World Wide Web*, 307-318, international World Wide Web Conferences Steering Committee.

Danescu-Niculescu-Mizil, Christian, Lee, Lillian, Pang, Bo & Kleinberg, Jon (2012): "Echoes of power: Language effects and power differences in social interaction", in *Proceedings of the 21st international conference on World Wide Web*, 699-708. ACM.

Denis, Alexandre, Quignard, Matthieu, Fréard, Dominique, Détienne, Françoise, Baker, Michael & Barcellini, Flore (2012) : « Détection de conflits dans les communautés épistémiques en ligne », in *Actes de la Conférence sur*

le Traitement Automatique des Langues Naturelles(TALN 2012), Grenoble, France, 351-358.

Ferschke, Olivier, Gurevych, Iryna & Chebotar, Yevgen (2012):. "Behind the article: Recognizing dialog acts in Wikipedia talk pages", in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 777-786. Association for Computational Linguistics.

Goldberg, Yoav & Orwant, Jon (2013): "A dataset of syntactic-ngrams over time from a very large corpus of English books", in *Proceedings of the Main Conference and the Share d Task: Semantic Textual Similarity. Association for Computational Linguistics*, 241–247.

Ferschke, Oliver, Daxenberger, Johannes & Gurevych, Iryna (2013): "A survey of NLP methods and resources for analyzing the collaborative writing process in Wikipédia", in *The People's Web Meets NLP*, Chapter 5, 121-160. Springer :Berlin, Heidelberg.

Herring, Susan C. (2012): "Grammar and Electronic Communication", in C. Chapelle (dir) *Encyclopedia of applied linguistics*. Wiley Blackwell.

Hundt, Marianne, Nesselhauf, Nadja & Biewer, Carolin (Eds.). (2007): *Corpus linguistics and the Web* (No. 59). Rodopi

Kanerva, Jenna, Luotolahti, M. Juhani, Laippala, Veronika & Ginter, Filip (2014): "Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish", in *Proceedings of the Sixth International Conference Baltic HLT2014*, 184-191.

Kilgarriff, Adam, Reddy, Siva, Pomikálek, Jan & Avinesh, P.V.S. (2010): "A Corpus Factory for Many Languages", in *Proceedings of the 7th international conference on Language Resources and Evaluation* (LREC 2010), 904-910.

Laippala, Veronika, Kanerva, Jenna, Pyysalo, Sampo, Missilä, Anna, Salakoski, Tapio & Ginter, Filip (2015a): "Towards the classification of the Finnish Internet Parsebank: Detecting Translations and Informality", in *Proceedings of the 20th Nordic Conference of Computational Linguistics* (Nodalida 2015), May 11-3, 2015, Vilnius.

Laippala, Veronika, Kanerva, Jenna & Ginter, Filip (2015b): "Syntactic Ngrams as Keystructures Reflecting Typical Syntactic Patterns of Corpora in Finnish", in *Procedia – Social and Behavioral Sciences. Current Work in Corpus Linguistics*. 198, 233–241.

Medelyan, Olena, Milne, David, Legg, Catherine & Witten, Ian H. (2009): "Mining meaning from Wikipédia", *International Journal of Human-Computer Studies*. 67(9), 716-754.

Paquot, Magali & Bestgen, Yves (2009): "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction", *Language and Computers*. 68(1), 247-269.

Poudat, Céline, Grabar, Nathalie, Kun, Jin & Paloque-Berges, Camille (2015) : « Corpus wikiconflits, conflits dans le Wikipédia francophone », in Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy.

Scott, Michael& Tribble, Christopher (2006): *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia, PA, USA: John Benjamins Publishing Company.

Tanguy, Ludovic (2013) : « La ruée linguistique vers le Web ». *Texto! Textes et Cultures*. 18(4). Publié en ligne: <halshs-00953760> (consulté le 1 juin 2016)

Urieli Assaf (2013) :*Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. thèse de doctorat, Université de Toulouse 2 - Jean Jaurès, France.

Urieli Assaf & Tanguy Ludovic (2013) : « L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane », in *Actes de la conférence Traitement Automatique des Langues Naturelles* (TALN 2013). Les Sables d'Olonne, France, 188-201.

Zesch, Torsten, Müller, Christof & Gurevych, Iryna (2008): "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary", in *Proceedings of the 8th international conference on Language Resources and Evaluation* (LREC 2012), 1646-1652.