Université de Montréal

Département de linguistique et de traduction, Faculté des arts et des sciences

---

Ce mémoire intitulé

La détection automatique multilingue d'énoncés biaisés dans Wikipédia

Présenté par

Desislava Aleksandrova

A été évalué par un jury composé des personnes suivantes

Patrick Drouin

Président-rapporteur

François Lareau

Directeur de recherche

Antoine Venant

Membre du jury

Novembre 2020

Résumé

Nous proposons une méthode multilingue pour l'extraction de phrases biaisées de Wikipédia, et l'utilisons pour créer des corpus en bulgare, en français et en anglais. En parcourant l'historique des révisions des articles, nous cherchons ceux qui, à un moment donné, avaient été considérés en violation de la politique de neutralité de Wikipédia (et corrigés par la suite). Pour chacun de ces articles, nous récupérons la révision signalée comme biaisée et la révision qui semble avoir corrigé le biais. Ensuite, nous extrayons les phrases qui ont été supprimées ou réécrites dans cette révision. Cette approche permet d'obtenir suffisamment de données même dans le cas de Wikipédias relativement petites, comme celle en bulgare, où de 62 000 articles nous avons extrait 5 000 phrases biaisées. Nous évaluons notre méthode en annotant manuellement 520 phrases pour le bulgare et le français, et 744 pour l'anglais. Nous évaluons le niveau de bruit, ses sources et analysons les formes d'expression de biais. Enfin, nous utilisons les données pour entrainer et évaluer la performance d'algorithmes de classification bien connus afin d'estimer la qualité et le potentiel des corpus.

Mots-clés : biais, neutralité, classification, multilingue, corpus, Wikipédia.

Abstract

We propose a multilingual method for the extraction of biased sentences from Wikipedia, and use it to create corpora in Bulgarian, French and English. Sifting through the revision history of the articles that at some point had been considered biased and later corrected, we retrieve the last tagged and the first untagged revisions as the before/after snapshots of what was deemed a violation of Wikipedia's neutral point of view policy. We extract the sentences that were removed or rewritten in that edit. The approach yields sufficient data even in the case of relatively small Wikipedias, such as the Bulgarian one, where 62k articles produced 5 thousand biased sentences. We evaluate our method by manually annotating 520 sentences for Bulgarian and French, and 744 for English. We assess the level of noise and analyze its sources. Finally, we exploit the data with well-known classification methods to detect biased sentences.

Keywords : bias, neutrality, classification, multilingual, corpora, Wikipedia.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations et des sigles

NPV	Neutralité de point de vue
ssi	si et seulement si
moy.	moyenne (d'échantillon)
s	écart type (d'échantillon)
κ	kappa de Fleiss
SVM	machine à vecteurs de support
RL	régression logistique
BG	bulgare
FR	français
EN	anglais

Remerciements

Mes premiers remerciements sont dus à François Lareau, un directeur brillant, rigoureux et stimulant qui a non seulement assuré la qualité de ce travail, mais il a su rendre le processus très agréable.

Je voudrais aussi remercier Pierre-André Ménard avec qui j'ai eu la chance de collaborer durant ma maîtrise pour son aide, sa compétence et la volonté de la transmettre. Un grand merci également à Lise Rebout, Hans Bherer et l'ancienne équipe PATX du CRIM pour l'opportunité de travailler avec eux.

Je tiens également à remercier les membres du jury Patrick Drouin et Antoine Venant pour la lecture approfondie et pour leurs commentaires pertinents.

Mes remerciements les plus sincères vont à Mme Xiabo Ren pour la généreuse bourse de maîtrise.

Un grand merci à mes collègues de l'Observatoire Linguistique Sens-Texte pour les nombreux moments partagés (prépandémie).

Enfin, je remercie mes familles bulgare et canadienne pour leurs encouragements, mes amis de près et de loin pour leur écoute, et avant tout, je remercie Patrice Fiset pour son support, patience, amour et surtout pour les excellents repas tous les jours!

Chapitre 1 Introduction

Dans le contexte actuel de fausses nouvelles, partisanerie et propagande largement diffusées dans l'espace numérique, l'utilité d'un outil de détection de biais est évidente. Sa conception l'est beaucoup moins. Une première complication découle de l'ambiguïté du terme biais, qui représente un manque d'équité ou de neutralité dans des domaines aussi variés que la cognition humaine (Tversky & Kahneman, 1974), la société (Perez, 2019; Ross et al., 1977; Wagner et al., 2015), les médias (Entman, 2007), Internet (Baeza-Yates, 2018; Bonart et al., 2019; Pitoura et al., 2018) ou les modèles et algorithmes statistiques (O'Neil, 2016; Shadowen, 2019) pour n'en nommer que quelques-uns. Compte tenu de la diversité des types de biais et de leurs définitions, il n'est pas anodin d'établir la portée d'une étude de détection des biais.

La deuxième grande difficulté devant la conception d'un algorithme de détection de biais est la disponibilité de données d'entrainement. La majorité des travaux dans ce domaine portent sur des articles de presse (Baly et al., 2018; Hirning et al., 2017; Hutto et al., 2015) et des blogues politiques (Iyyer et al., 2014; Yano et al., 2010) plutôt que sur Wikipédia, en raison de la rareté relative des exemples fournis par une encyclopédie.

L'approche que ce travail décrit propose des solutions directes à ces problèmes. Premièrement, Wikipédia fournit une définition de la partialité dans son principe fondateur de neutralité de point de vue[1] (NPV), contrairement aux autres sources de données. Deuxièmement, l'historique de révisions des articles offre la possibilité d'extraire sans frais des exemples de contenu biaisé étiquetés par des contributeurs de l'encyclopédie. Enfin, ses versions multilingues permettent la reproduction de l'expérience en plusieurs langues.

[1] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Neutralit%C3%A9_de_point_de_vue

1.1 Définition de la tâche

La détection automatique de biais est un type d'analyse de sentiment très similaire à la détection de subjectivité (Al Khatib et al., 2012; C. Lin et al., 2011; Murray & Carenini, 2009; Riloff & Wiebe, 2003; Wiebe & Riloff, 2005; Wilson & Raaijmakers, 2008), où l'on considère une opinion comme subjective et un fait comme objectif. Dans ces travaux, on apprend à discriminer entre des phrases subjectives (le plus souvent tirées d'articles d'opinion ou de commentaires) et objectives (habituellement tirées d'articles de presse). Cependant, nous verrons plus loin que la partialité se manifeste autant par des opinions subjectives que par des faits dits objectifs (cf. §6).

Plusieurs approches d'analyse ou de détection de partialité ont fait valoir les versions multilingues des articles (Al Khatib et al., 2012; Massa & Scrinzi, 2012; Zhou et al., 2015, 2016). Par exemple, Massa and Scrinzi (2012) se sont penchés sur les degrés d'objectivité que présentent les versions multilingues d'un même article. Contrairement à ces démarches, la nôtre examine les manifestations de biais dans trois langues isolément, notamment en anglais, en français et en bulgare. Ces versions langagières, non pas choisies au hasard, sont représentatives à la fois de trois familles de langues indo-européennes distinctes (germanique, romane, slave) et de trois tailles de Wikipédia en termes de nombres d'articles (large, moyenne, petite). Finalement, nous avons choisi de classifier des phrases plutôt que des articles (Zhou et al., 2015) afin d'assurer une quantité suffisante d'exemples dans un plus grand nombre de langues. Pour la même raison, nous ne nous limitons pas à une classification de mots (Recasens et al., 2013), car cette approche ne fournit d'exemples que pour les wikis les plus larges.

Notre objectif est de détecter automatiquement des phrases en violation de la politique de neutralité de point de vue de Wikipédia au moyen d'une procédure applicable à plusieurs langues. Dans ce qui suit, nous utiliserons le terme « biais » comme synonyme de « partialité » pour désigner l'absence de neutralité; autrement dit, la présence d'aspects subjectifs et non neutres dans des propos à l'intérieur d'articles de l'encyclopédie publique.

1.5 Organisation et contributions

Ce travail représente un compte détaillé et augmenté de l'article *Multilingual Sentence-Level Bias Detection* in Wikipedia (Aleksandrova et al., 2019) développé en collaboration avec François Lareau[2] et Pierre-André Ménard[3] et présenté lors de la conférence *Recent Advancements in Natural Language Processing* à Varna (Bulgarie) en 2019. Le tableau 1 résume les contributions initiales des différents auteurs organisées en fonction des étapes de développement du projet. Puisque l'organisation du mémoire suit le même ordre, la première colonne fait référence aux chapitres correspondant aux étapes respectives, s'il y a lieu.

§	CONTRIBUTION	AUTEUR(S)
1-2	Conception du projet et définition de biais	*Aleksandrova D*
3	Revue de la littérature	*Aleksandrova D*
4	Création des corpus	
—	Logique et procédure	*Aleksandrova D, Lareau F*
4.1	Listes de balises	*Aleksandrova D, Ménard PA*
4.2	Extraction des paires de révisions + parallélisation du traitement des données en anglais	*Aleksandrova D, Lareau F* *Ménard PA*
4.3	Traitement et filtrage des paires de révisions	*Aleksandrova D*
4.4	Segmentation et filtrage des phrases	*Aleksandrova D*
—	Équilibrage des classes; division des corpus	*Aleksandrova D*
5	Évaluation des corpus	
5.1	Analyse préliminaire	*Aleksandrova D, Lareau F*
5.2	Protocole d'annotation	*Aleksandrova D*
5.3	Calcul des résultats	*Lareau F*
5.4	Analyse des sources de bruits	*Aleksandrova D*

[2] Professeur agrégé au département de linguistique et de traduction à l'Université de Montréal
[3] Chercheur en traitement automatique du langage au centre de recherche en informatique de Montréal (CRIM)

6	Analyse des formes d’expression de biais	*Aleksandrova D*
7	Expériences de classification	
—	fastText	*Aleksandrova D*
—	Descente stochastique de gradient	*Ménard PA*
—	Régression logistique	*Ménard PA*
8	Résultats et discussion	
—	Analyse des résultats	*Ménard PA, Aleksandrova D*
8.1	Analyse des erreurs	*Aleksandrova D*

Tableau 1. Résumé des contributions des auteurs par étape

Chapitre 2 Biais et partialité dans Wikipédia

La mise en place d'une politique de neutralité dans Wikipédia ne garantit pas automatiquement son respect, car la responsabilité d'y adhérer incombe aux contributeurs pour qui la détection de certains types de partialité reste difficile. Par exemple, Yano *et al*. (2010) démontrent que la partisanerie politique dans les textes d'opinion passe inaperçue lorsqu'elle est conforme aux idées et aux croyances du lecteur. De façon similaire, les auteurs sont influencées par leurs préférences personnelles. S'ils estiment une personne, ils auront tendance à la tenir responsable de ses succès, tout en attribuant ses échecs à des facteurs externes (Hutto et al., 2015). De plus, le niveau d'implication des contributeurs semble être corrélé avec la force de leurs sentiments. Zhou *et al*. (2015) ont trouvé que le degré d'engagement personnel des gens dans un sujet lié à la guerre influence le nombre de mots et le nombre de concepts subjectifs qu'ils exprimeront.

L'existence de ces « angles morts » d'une part et l'étendue de l'encyclopédie de l'autre ont motivé plusieurs à élaborer des outils automatiques de détection de partialité (cf. §3). Avant de procéder au résumé des travaux antérieurs et à l'exposé du nôtre, nous allons décrire dans ce chapitre les grandes lignes de la politique NPV de Wikipédia qui sert à la fois d'encadrement et de définition de la notion de biais.

2.1 Neutralité de point de vue (NPV)

Le principe fondateur de neutralité de point de vue (NPV) dicte que « les articles doivent être écrits de façon à ne pas prendre parti pour un point de vue plutôt qu'un autre »[4]. Les lignes directrices du principe de neutralité sont les suivantes :

[4] *ibid.*

2.1.1 Éviter de présenter des opinions comme des faits

Lorsqu'un article contient des opinions pertinentes exprimées sur son sujet, ces derniers ne doivent pas être exprimées dans la voix de Wikipédia. Ils devraient plutôt être attribués dans le texte à des sources particulières ou, lorsque cela se justifie, décrits comme des opinions répandues. Par exemple, un article ne devrait pas dire que « le génocide est une action malfaisante », mais il peut dire que « le génocide a été décrit par Jean X comme l'incarnation du mal humain ».

En cas d'affirmations contradictoires sur une question (apportées par des sources fiables), il faut traiter ces affirmations comme des opinions plutôt que comme des faits.

2.1.2 Éviter de présenter des faits comme des opinions

Les affirmations factuelles non contestées et non controversées faites par des sources fiables devraient normalement être directement énoncées dans la voix de Wikipédia (avec un lien de référence à l'appui). Il ne faut pas rédiger le passage de façon à ce qu'il paraisse contesté.

2.1.3 Écrire sans porter de jugement

Un point de vue neutre ne se montre ni en accord ni en désaccord avec le sujet (ou avec ce que des sources fiables disent par rapport à ce sujet).

2.1.4 Accorder aux points de vue différents une place proportionnelle à leur importance dans les études sur le sujet

« La neutralité de point de vue ne signifie pas qu'il faille présenter nécessairement tous les points de vue existants sur un sujet. Ne doivent l'être que ceux dont la présence dans une encyclopédie est pertinente. »[5]

[5] ibid.

2.2 Neutralité de ton

En plus de la définition du phénomène proscrit de partialité, Wikipédia explicite un ensemble d'expressions à éviter à cause de leur potentiel à véhiculer un biais. Il existe des listes de mots à utiliser avec précaution pour 36 langues, y compris les trois langues concernées par notre étude : l'anglais, le français et le bulgare. Il importe de mentionner ici qu'il n'y a pas de mots censurés sur Wikipédia, mais « [...] certains termes [sic] doivent être utilisés avec précaution car ils peuvent introduire un biais. Ils risquent alors d'appuyer un point de vue, être péjoratifs ou encore amener un style non souhaitable. »[6] La répartition en catégories de ce vocabulaire varie légèrement d'une langue à l'autre, mais de façon générale, il est question de six classes de mots à éviter :

2.2.1 Les mots évasifs ou non spécifiques (tergiversations)

Les tergiversations (*weasel words* en anglais) représentent un certain type de formulation qui tente de masquer un manque de neutralité tout en prenant implicitement parti, par manque de précision, sur l'origine des thèses avancées. Le contenu évasif n'est pas suffisamment précis pour être vérifiable, ce qui prive le lecteur de la possibilité d'évaluer la source du point de vue. Une forme courante de contenu évasif est l'attribution vague, où une déclaration est revêtue d'autorité, mais n'a pas de fondement substantiel, car l'identité des agents est camouflée : *Certains affirment…; La plupart des gens pensent que…; Une minorité soutient que…; Tout le monde sait que…; Les dernières recherches montrent que…; Il est à noter que…; On…*

2.2.2 Les mots rédactionnels

Les soi-disant mots rédactionnels sont les mots susceptibles de mettre en avant un point de vue que l'on emploie librement dans les articles d'opinion. Dans Wikipédia par contre, utiliser certains adverbes pour souligner la certitude ou l'absence de doute devrait généralement être évité afin de maintenir un ton impartial. *Bien sûr, naturellement, bien entendu, selon toute évidence, comme de raison, essentiellement, principalement, en gros, dans le fond,*

[6] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Termes_%C3%A0_utiliser_avec_pr%C3%A9caution

manifestement, etc., sont souvent utilisés pour éviter de fournir des explications, des arguments ou de citer des sources. Par ce fait, ils peuvent, s'ils sont mal employés, trahir un point de vue. Par contre, parmi les exemples donnés ci-haut, il existe des adverbes polysémiques dont l'utilisation dans des contextes appropriés n'est guère proscrite. C'est le cas, par exemple, de l'adverbe *naturellement* dans la phrase « Le plutonium est produit naturellement dans les réacteurs nucléaires naturels du Gabon »[7].

Encore plus subtil est le choix de conjonction de coordination tel que *mais, malgré, pourtant, cependant*, ou *même si*, qui peut également produire un effet de partialité en modifiant, souvent de façon injuste ou incorrecte, le poids des propositions ou même en créant un lien là où il n'en existe pas.

2.2.3 Les mots controversés (*loaded words*)

Cette classe regroupe différents types de mots chargés de connotations et d'émotions dont l'utilisation sous-entend un jugement de valeur. Traiter quelqu'un d'*extrémiste* ou d'*activiste* au même titre qu'appeler une croyance *culte, secte* ou *fondamentalisme* a comme effet de communiquer le point de vue partial soit de l'endogroupe soit de l'exogroupe. Même s'il existent des contextes appropriés pour l'usage de ces mots, Wikipédia suggère de les éviter lorsqu'un mot neutre peut être employer à la place afin d'éviter qu'ils soit qu'ils soient perçus comme un point de vue par certains groupes. Par exemple, l'adjectif *États-Unien* est parfois utilisé comme synonyme d'*américain* même quand ce dernier est le seul terme indiqué dans le code de rédaction interinstitutionnel de l'Union européenne[8]. Cependant, cet emploi est à proscrire, car *États-Unien* apparaît souvent dans des textes critiques à l'égard des États-Unis d'où il comporte une connotation négative. Même lorsque ces « étiquettes » ne sont pas péjoratives, elles reflètent un point de vue du narrateur ou une idée reçue qui peut être en désaccord avec les idéologies des personnes ou groupes impliqués. Alors pour ne pas imposer

[7] La phrase fait référence aux dépôts d'uranium du Gabon considérés comme les « seuls réacteurs nucléaires naturels connus au monde à ce jour », https://fr.wikipedia.org/wiki/R%C3%A9acteur_nucl%C3%A9aire_naturel_d%27Oklo

[8] http://publications.europa.eu/code/fr/fr-5000500.htm

son point de vue dans la description de l'homéopathie par exemple, il est mieux d'éviter l'épithète *pseudoscientifique* en faveur d'une description sans mots controversés :
« L'homéopathie gagne en popularité, mais ni son empirisme ni son fondement hypothétique ne répondent aux normes scientifiques »[9].

2.2.4 Les expressions de doute

L'introduction de propos rapportés par *prétendre*, *assurer, avancer, soutenir*, etc., laissent entendre que Wikipédia soutient un point de vue ou au contraire doute de celui-ci. Ils sont donc généralement à éviter en l'absence d'éléments assurant le caractère douteux ou inexact de l'affirmation contestée. L'adjectif *soi-disant*, qui signifie 'qui prétend être tel' comporte une connotation négative en français, contrairement à l'emploi courant en anglais de *so-called*, qui a le sens de 'que l'on appelle'. Il est donc à utiliser avec parcimonie.

2.2.5 Les mots excessivement positifs ou flatteurs

Certains mots posent une affirmation très forte ou promeuvent le sujet d'un article, sans pour autant communiquer ni résumer de façon claire des informations vérifiables. Il est question le plus souvent d'intensificateurs subjectifs agissant en tant que modificateurs nominaux ou verbaux ayant la forme d'adjectif (*meilleur, légendaire, extraordinaire, célébré, notable*, etc.) ou d'adverbe (*extrêmement, infiniment, véritablement*, etc.), mais aussi de superlatifs ou de quantifieurs. Ces expressions possèdent une forte connotation laudative qui brise la neutralité de ton.

2.2.6 Certains synonymes de dire

Rapporter des propos de façon neutre et exacte se fait au moyen des verbes dire, déclarer, décrire, écrire, commenter ou de la préposition selon. Toute autre formulation risque de transmettre davantage d'information non vérifiable ou carrément fausse. Par exemple, écrire

[9] Exemple tiré de:
https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Termes_%C3%A0_utiliser_avec_pr%C3%A9caution#Termes_qui_sous-entendent_un_jugement_de_valeur

qu'une personne a *clarifié, expliqué, exposé, trouvé, signalé* ou *révélé* quelque chose présuppose la véracité des propos. Écrire qu'une personne a *insisté*, *noté, observé, spéculé* ou *présumé* suggère un degré de prudence ou de certitude de la personne à qui on attribue la citation, même lorsque ces aspects sont invérifiables.

2.3 Les articles non neutres

Lorsqu'un article est considéré comme biaisé, un wikipédien[10] peut le signaler en ajoutant une balise[11] telle que `{{NPV}}` à sa source[12], qui fait s'afficher un avertissement (Figure 1) au-dessus du bandeau de la page, de la section ou du passage concerné.

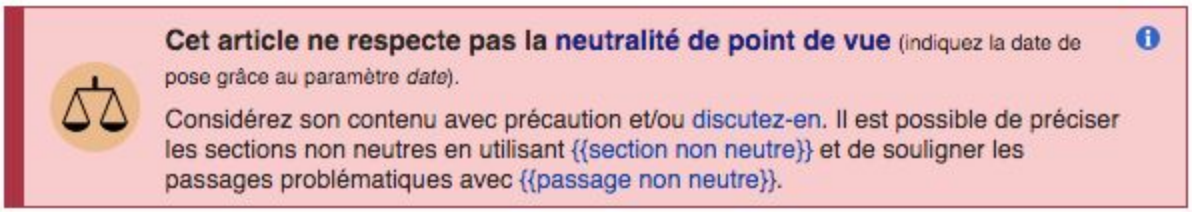


Figure 1. Bannière d'avertissement

Ces grandes lignes directrices de la politique de neutralité (et les rédacteurs qui les appliquent) aident graduellement à réduire les préjugés dans Wikipédia grâce à un processus continu de révision collaborative du contenu (Pavalanathan et al., 2018). Cependant, de nouveaux cas de préjugés sont introduits aussi souvent que des anciens sont négligés en raison de la difficulté inhérente des humains à détecter les biais face aux expressions subtiles des points de vue (cf. §6). En effet, Recasens *et al.* (2013) ont montré que lorsqu'on leur présente une phrase biaisée de Wikipédia, les annotateurs parviennent à identifier correctement le mot porteur de biais dans seulement 37% des cas.

[10] Les wikipédiens sont les personnes qui écrivent et éditent les articles de Wikipédia.

[11] Les balises (appelée aussi «modèles ») servent à reproduire sur plusieurs pages le même message ou le même formattage. En wikicode, un modèle est délimité par des doubles accolades, « {{ » et « }} ». https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Mod%C3%A8le_de_non-neutralit%C3%A9

[12] La source d'une page Wikipédia (appelée aussi «wikicode») est composée d'une syntaxe et d'un système de code interprétables par le logiciel de MediaWiki.

Chapitre 3 Travaux antérieurs

Les approches de détection de biais existantes varient principalement en fonction de leur choix de corpus, de méthode de vectorisation et d'algorithme de classification (Tableau 2). Nous résumons, dans ce chapitre, les travaux pertinents à cette tâche selon ces trois axes.

CORPUS	**VECTORISATION**	**CLASSIFICATION**
Emploi de corpus existants	Vecteurs-mots creux ou denses	Listes de mots (dictionnaires)
Annotation manuelle	Vecteurs de caractéristiques faites sur mesure	Méthodes d'apprentissage automatique (*Naive Bayes, Random Forest*)
Présélection automatique et annotation manuelle	Un mélange des deux approches	Méthodes d'apprentissage profond (*Convolutional Neural Network*, *Recurrent neural network*)
Extraction et annotation automatiques		

Tableau 2. Résumé des approches de détection de biais selon les trois axes principaux

3.1 Corpus

La tâche de détection de biais par apprentissage automatique nécessite, dans un premier temps, la composition d'un corpus d'exemples positifs et négatifs sur lequel on entraîne, par la suite, un ou des algorithmes d'apprentissage automatique.

Parmi les travaux connexes à la détection de biais dans Wikipédia, on retrouve ceux qui s'appuient sur des corpus existants (Ganter & Strube, 2009; Kuang & Davison, 2016; Vincze, 2013), d'autres qui annotent manuellement (Al Khatib et al., 2012; Ganter & Strube, 2009; Herzig et al., 2011; Hube & Fetahu, 2018; Hutto et al., 2015), d'autres encore qui annotent

manuellement ce qu'ils avaient extrait automatiquement (Yano et al., 2010; Yiwei Zhou, 2016), et enfin, ceux qui tentent d'extraire automatiquement des exemples étiquetés (Ganter & Strube, 2009; Hube & Fetahu, 2018; Recasens et al., 2013). Notre approche s'inscrit parmi ces derniers.

3.1.1 Corpus existants

L'utilisation de corpus existants, tout en étant la méthode la moins chère en termes de temps et de main-d'oeuvre, impose des limites quant aux types de biais à explorer et aux langues à l'étude. Vincze (2013) utilise WikiWeasel, un des corpus de la tâche partagée CoNLL-2010 (Farkas et al., 2010) pour étudier le phénomène d'incertitude contextuelle. Disponible en anglais seulement, WikiWeasel comprend 2 484 phrases contenant des mots évasifs ou non spécifiques (*weasel words*), ce qui ne représente qu'une des six classes de mots à éviter (cf. §2.2) ainsi qu'une des nombreuses formes d'expression de biais que l'on trouve dans Wikipédia (cf. §6). De façon similaire, Ganter et Strube (2009) étudient la détection de mots vagues (*hedges*) dans un corpus de 1 000 phrases contenant la balise `{{weasel}}` tirées de Wikipédia. Enfin, Kuang et Davison (2016) rapportent un gain de précision sur le corpus (entièrement en anglais) de Recasens *et al.* (2013) grâce à des représentations vectorielles contextuellement riches produites par un modèle de langue neuronal entraîné sur Wikipédia.

3.1.2 Annotation manuelle

L'annotation manuelle produit normalement des corpus de très bonne qualité, mais nécessite souvent un investissement de temps et d'argent non négligeable, surtout pour les corpus multilingues. La qualité des annotations peut aussi varier selon l'aptitude des annotateurs. Si certains phénomènes linguistiques (p.ex. la grammaticalité d'une phrase) se prêtent tout aussi bien à l'étiquetage par des linguistes qu'à l'externalisation ouverte (*crowdsourcing*), d'autres demandent des connaissances plus fines ou spécialisées du phénomène à l'étude.

Dans son étude exploratoire des marqueurs d'incertitude contextuelle, Vincze (2013) développe une nouvelle ressource lexicale. Des 4 530 articles du corpus WikiWeasel comptant

20 756 phrases, deux linguistes ont identifié et annoté 10 794 marqueurs d'incertitude contenus dans 7 336 phrases (ou dans 35% du corpus). Par exemple, dans la phrase *While the Skyraider is not as iconic as* ***some other*** *aircraft, it has been featured in some Vietnam-era films such as The Green Berets (1968) and Flight of the Intruder (1991)*, la paire de déterminant et adjectif (en gras) sous-spécifie un des arguments de la phrase et se classe alors parmi les mots évasifs ou non spécifiques (cf. §2.2.1).

Hutto *et al.* (2015) ont fait annoter les phrases de cinq nouvelles fictives sur les actions des présidents George W. Bush et Barack Obama par 91 participants afin d'obtenir leur jugement sur le degré de partialité (légèrement, moyennement ou extrêmement partiale).

Hube et Fetahu (2018), quant à eux, ont appris à détecter le biais dans Wikipédia à la base d'un corpus de 685 phrases annotées manuellement par des utilisateurs de la plateforme CrowdFlower[13] où environ la moitié (323) étaient considérées partiales. Il est intéressant de noter que les auteurs ont tiré leur corpus de Conservapedia[14], un wiki façonné selon des idées conservatrices de droite sans prétention d'objectivité et avec de vives critiques de la politique libérale et des membres du Parti démocrate des États-Unis. Bien que similaires en termes de sujets traités, les deux wikis se distinguent par leur respect de la neutralité de point de vue. Dans le cas de Conservapedia, ce principe fait place à une volonté de présenter les faits d'un point de vue conservateur, ce qui assure un bon nombre d'exemples partiaux. Les auteurs motivent cet emploi de source alternative par une volonté de réduire les coûts liés à l'annotation manuelle. De plus, ils évoquent la rareté du phénomène à l'étude[15] (ici, la partialité) dans la source textuelle (en l'occurrence, Wikipédia) comme une difficulté importante pour la présélection d'exemples. Autrement dit, lorsque l'expression ou la construction que l'on cherche à étudier se fait rare dans les documents à notre disposition, la

[13] https://crowdflower.com/
[14] http://www.conservapedia.com/
[15] Hube et Fetahu (2018) estiment qu'environ 4% des phrases dans Wikipédia sont biaisées. Leur approximation résulte de l'application de leur classifieur sur un échantillion de phrases en anglais (du domaine de la politique américaine). Pour une estimation basée sur l'entièreté de Wikipédia, cf. §4.3 où nous présentons les résultats de notre extraction de révisions biaisées pour trois langues.

méthode d'échantillonnage aléatoire ne fournit plus suffisamment d'exemples positifs. Il s'impose alors des méthodes hybrides composées de présélection automatique et de validation/annotation manuelle pour la création de corpus d'entraînement et d'évaluation.

3.1.3 Présélection automatique et annotation manuelle

Pour leur étude exploratoire de la partialité, Yano *et al*. (2010) ont extrait les phrases les plus partiales (pour un total de 1 041 phrases) d'un ensemble de billets de blogues politiques (équilibré entre blogues libéraux et blogues conservateurs) qu'ils ont ensuite fait annoter par cinq personnes sur Mechanical Turk[16]. Pour identifier ces phrases de façon automatique, les auteurs ont recouru à trois critères :

- la présence de *sticky partisan bigrams* (les mille bigrammes libéraux et conservateurs les plus fréquents, moins les doublons et le bruit)
- la présence de mots indicateurs de sentiments (selon les dictionnaires spécialisés de LIWC[17])
- la présence de *kill verbs* (liste de verbes employés par Greene et Resnik (2009) dans leur étude du lien entre la structure syntaxique et la perception des sentiments).

La rareté des exemples a aussi amené Zhou *et al*. (2016) à employer une approche hybride pour la construction de leur corpus de phrases influentes en provenance de Wikipédia. Ces phrases non neutres (à charge positive ou négative) représentent un type particulier de partialité et se caractérisent par leur capacité d'affecter la réputation d'une entité nommée. Par exemple, dans Wikipédia en allemand, Angela Merkel est décrite comme "*a well-known student with excellent performance in any event at the University of Leipzig"*. La détection de telles expressions nécessite une classification binaire en deux étapes, où la première filtre les phrases influentes des autres phrases et la deuxième distingue les positives des négatives. Pour extraire des phrases potentiellement influentes, les auteurs ont d'abord recueilli toute mention de 219

[16] https://www.mturk.com/
[17] *The Linguistic Inquiry and Word Count*, un logiciel d'analyse textuelle conçu pour l'étude des aspects émotifs, cognitifs et structuraux des corpus textuels http://liwc.wpengine.com/

entités nommées (corporations, politiciens, vedettes et athlètes) ainsi que des variations de leurs formes de surface (grâce à l'ontologie DBpedia[18]), pour un total de 1 196 403 phrases. Ils ont ensuite classé les phrases selon leur score d'objectivité calculé comme la somme de la polarité des mots (dans SentiWordNet[19]) normalisée par le nombre de mots dans la phrase. Les 2 500 phrases les plus subjectives et environ autant de phrases choisies aléatoirement ont formé le sous-corpus de phrases que des membres de CrowdFlower[20] ont classées comme influentes ou neutres.

La création du corpus WikiWeasel (Farkas et al., 2010) résulte d'une extraction automatique de tous les paragraphes contenant la balise `{{weasel}}`[21] dans l'ensemble du Wikipédia en anglais (5 874 paragraphes), suivie par l'annotation d'un échantillon de 438 paragraphes choisis au hasard. Pour chacun de ces paragraphes, les annotateurs ont identifié les phrases partiales et les constructions évasives les plus courantes. Ensuite, deux autres ensembles de paragraphes de Wikipédia ont été recueillis selon qu'ils contenaient ou non de telles constructions. Les paragraphes initialement extraits ont fourni les exemples de la classe positive, tandis que le reste de Wikipédia a fourni des exemples de la classe négative, où les mêmes constructions se trouvaient dans des contextes neutres.

3.1.4 Extraction et annotation automatiques

L'historique des éditions de Wikipédia donne la possibilité d'automatiser l'extraction de contenu et de créer des corpus étiquetés en plusieurs langues relativement facilement. En explorant les révisions de chaque article dans le temps, on peut extraire des tuples de fautes lexicales ou grammaticales avec leurs corrections (Cahill et al., 2013; Nelken & Yamangil, 2008; Wisniewski et al., 2010), des exemples de paraphrases (Max & Wisniewski, 2010), de simplification (Yamangil & Nelken, 2008) ou de résumé (Nelken & Yamangil, 2008). De façon générale, les approches automatiques ou hybrides assurent la quantité au détriment de la

[18] http://dbpedia.org/
[19] http://sentiwordnet.isti.cnr.it/
[20] *Crowd Flower, loc. cit.*
[21] Un marqueur de contenu évasif (cf. §2.2.1)

qualité, mais elles permettent (lorsqu'elles sont bien conçues) d'étendre rapidement la portée du corpus à plusieurs langues.

Une approche complètement automatique est celle de Bhosale *et al.* (2013), qui ont compilé un corpus de contenu promotionnel à partir de 13 000 articles en anglais porteurs de l'étiquette `{{advert}}`. Quant à la classe négative, ils ont d'abord créé un ensemble bruité de 26 000 articles sans étiquettes qui contenaient potentiellement du contenu promotionnel ignoré par les rédacteurs de Wikipédia. Ils ont extrait par la suite 11 000 articles de qualité[22] (*featured articles*) en tant qu'exemples d'articles neutres par excellence.

Contrairement à Bhosale *et al.* (2013), qui ont constitué leur corpus d'articles complets (promotionnels ou neutres), Recasens *et al.* (2013) ont proposé une heuristique d'extraction automatique de phrases partiales où chacun des mots est identifié comme neutre ou biaisé. Leur tâche était alors d'apprendre à identifier le mot problématique étant donné une phrase non neutre. Pour identifier et extraire des phrases non neutres, ils ont compté sur les commentaires laissés par les rédacteurs lors de leurs révisions. Si un commentaire faisait mention de la politique de neutralité (*p. ex. Attempts at presenting some claims in more NPOV*[23] *way*; ou *Merging in a passage from the researchers article after basic NPOV-ing*), ils supposaient que la rédaction associée avait corrigé un problème de biais contenu dans la version antérieure de l'article, comme dans ces exemples tirés de paires de révisions successives :

(1) a. Usually, smaller cottage-style houses have been demolished to make way for these **McMansions**.
b. Usually, smaller cottage-style houses have been demolished to make way for these **homes**.

(Recasens et al., 2013)

(2) a. Kuypers **claimed** that the mainstream press in America tends to favor liberal viewpoints.

[22] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Articles_de_qualit%C3%A9
[23] *NPOV* étant l'équivalent de NPV en anglais.

b. Kuypers **stated** that the mainstream press in America tends to favor liberal viewpoints.

(Recasens et al., 2013)

Ils procédaient alors à l'extraction de ces deux états du même article qu'ils comparaient par la suite afin d'en isoler les différences. Des 16 819 éditions ainsi identifiées dans le Wikipédia en anglais, Recasens *et al*. (2013) n'ont conservé que les phrases de la version « avant » où un mot avait été remplacé par un à cinq mots dans la version « après » (en ignorant les changements mineurs définis par une distance Levenshtein de moins de 4 caractères). En tout, le corpus final comptait 2 235 phrases partiales où les mots-cibles des segments sources portaient l'étiquette de la classe positive et les autres mots, celle de la classe négative. Bien que fiable, cette méthode produit très peu d'exemples pour l'anglais et aucun pour les Wikipédias plus petits, en raison de sa forte dépendance aux commentaires de révisions (qui sont facultatifs).

3.2 Vectorisation

La vectorisation d'un corpus de documents textuels (mots, phrases, paragraphes, articles ou autre) consiste à représenter les propriétés de chacun des documents par un vecteur de nombres réels (appelés caractéristiques) et la totalité des documents par une matrice de caractéristiques. Les vecteurs-mots ainsi obtenus peuvent encoder (à un degré variable selon la méthode de vectorisation) de l'information sur leur sens lexical et grammatical, sur leurs cooccurrents ainsi que sur leur contexte. Les approches de vectorisation les plus simples sont basées sur la fréquence d'occurrence des mots du corpus (tf, tf-idf, n-grammes) ou bien sur des plongements de mots extraits de modèles de langue neuronaux déjà entraînés. Les approches plus complexes impliquent l'extraction de caractéristiques (*feature extraction*) où on rajoute les valeurs de variables (potentiellement) discriminantes que l'on a définies *ad hoc*. C'est la technique la plus répandue dans les travaux sur la partialité dans Wikipédia, suivie par l'emploi de représentations issues de modèles de langue préentraînés. Si l'efficacité de la première approche peut varier largement selon la créativité des auteurs et la richesse des ressources lexicales utilisées, l'étude de Bellows (2018) n'a démontré aucune différence significative entre

la performance de classifieurs entraînés sur des représentations de *word2vec*, *GloVe* ou *fastText*. L'efficacité des divers plongements de mots pour cette tâche est ainsi comparable.

Plusieurs chercheurs (Bhosale et al., 2013; Ganter & Strube, 2009; Hube & Fetahu, 2018; Recasens et al., 2013) choisissent d'incorporer des caractéristiques stylométriques, lexicales, contextuelles et linguistiques dans les représentations vectorielles de leurs corpus. Quoique utiles, ce type de caractéristiques dépendent souvent de listes de mots construites sur mesure, de ressources lexicales spécialisées telles que SentiWordNet (Baccianella et al., 2010), subjClue (Gitari et al., 2015), LIWC (Pennebaker et al., 2001), etc., ainsi que de logiciels d'analyse grammaticale ou morphologique souvent disponibles en anglais seulement.

Recasens *et al*. (2013) ont représenté les tokens de leur corpus en tant que vecteurs comprenant les dimensions suivantes :

- la présence ou l'absence de chaque mot dans huit dictionnaires de mots partiaux tirés d'autres travaux;
- sa partie du discours;
- son lemme;
- des caractéristiques contextuelles du mot dans une fenêtre de cinq mots;
- le ratio entre le nombre de fois qu'un mot a fait l'objet d'une révision et sa fréquence d'occurrence. L'objectif de la dernière caractéristique est d'identifier les mots ambigus qui ne sont partiaux que dans certains contextes.

Kuang *et al*. (2016) adressent plusieurs critiques à la méthode de détection de partialité textuelle par listes de mots (Recasens et al., 2013). D'une part, elle n'est adéquate que lorsque la partialité concerne des unités lexicales monosémiques. Or, toute unité lexicale polysémique, dont certains sens ne posent pas problème pour la partialité de la phrase, produirait de faux positifs selon l'approche par tables de consultation. D'autre part, cette méthode n'est pas suffisamment flexible, car elle détecte uniquement les mots contenus dans ces listes tout en

ignorant leur contexte, assurant ainsi un rappel faible. Leur solution consistait à rajouter aux vecteurs de mots la représentation vectorielle de chaque document en tant que caractéristique contextuelle supplémentaire dans le but de distinguer entre les homographes. Appliquée au corpus de phrases partiales de Recasens *et al*. (2013), cette approche a démontré que les caractéristiques contextuelles ne sont utiles que dans le cas des mots ambigus. Étant donné la rareté de tels mots dans le corpus en question (706 sur 3 249 mots au total), ces caractéristiques augmentent inutilement le nombre de dimensions à considérer.

Hutto *et al*. (2015) se sont penchés sur l'identification des caractéristiques les plus pertinentes pour la perception de la partialité. Pour chaque phrase de leur corpus, ils ont extrait 26 caractéristiques qu'ils ont utilisées en guise de représentation. Parmi les caractéristiques à l'étude, les auteurs distinguent celles d'ordre syntaxique, pragmatique et structural (sentiments, subjectivité, certitude, mode grammatical, difficulté du texte) des caractéristiques lexicales (présence ou absence dans un vocabulaire prédéfini). Ces dernières comprennent les caractéristiques de partialité épistémique et de subjectivité de Recasens *et al*. (2013), une collection de vocabulaires du LIWC et quelques caractéristiques supplémentaires sous forme de listes de mots (des mots d'opinion, des modificateurs de degré, des marqueurs de cohérence).

En amont de la classification des 685 phrases de Conservapedia composant leur corpus, Hube et Fetahu (2018) ont construit un vocabulaire de presque 10 000 mots partiaux sur lequel ils ont basé deux de leurs caractéristiques discriminantes. Afin de dresser cette liste, ils ont d'abord identifié quelques sujets controversés et polarisants (*p. ex. médias, immigrants, avortement*). Ils ont ensuite parcouru à la main les mots les plus proches de la représentation vectorielle de chacun de ces sujets dans un modèle *word2vec* (Mikolov et al., 2013) entraîné sur Conservapedia afin d'en sortir d'autres expressions sémantiquement proches qu'ils ont rajouté à la liste initiale. Par exemple, dans cet espace multidimensionnel conservateur, les voisins de *media* sont *arrogance, whining, despises* et *blatant*. La liste de sujets ainsi enrichie comptait 100 mots polarisants autour desquels graviteraient, selon l'hypothèse des auteurs, encore plus de mots non neutres. Pour les trouver, Hube et Fetahu (2018) utilisent un modèle de langue

entrainé sur l'ensemble d'articles Wikipédia en anglais. Plutôt que de simplement en extraire les mots les plus proches de chacun des 100 mots polarisants, ils calculent la moyenne des vecteurs de dix mots polarisants choisis aléatoirement de l'ensemble et retiennent les 1 000 mots entourant le vecteur combiné de ce noyau de mots partiaux. Le tableau 3 donne comme exemple les 20 mots les plus proches du mot-clé *indoctrinate* (à gauche) et d'un noyau de dix mots-clés dont *indoctrinate* fait partie (à droite).

Un mot-clé		**Dix mots-clés**	
cajole	outmaneuver	hypocritical	scorn
emigrates	helmswoman	indifference	downplaying
ingratiate	outflank	ardently	discrediting
endear	renditioned	professing	demeaning
abscond	redeploy	homophobic	prejudices
americanize	seregil	mocking	humiliate
reenlist	unnerve	complacent	determinedly
overawe	titzikan	recant	frustration
disobey	unbeknown	hatred	ridicule
reconnoiter	terrorise	vilify	disrespect

Tableau 3. Les mots dont les vecteurs se trouvent le plus proche de celui d'*indoctrinate* (à gauche) et les mots entourant un noyaux de mots partiaux (*indoctrinate, resentment, defying, irreligious, renounce, slurs, ridiculing, disgust, annoyance, misguided*)

De cette façon, ils arrivent à identifier plusieurs mots partiaux tout en limitant le nombre de faux positifs, ce qui leur permet de construire un dictionnaire de 9 742 mots (après la suppression des doublons).

Zhou *et al*. (2016) explorent différentes combinaisons de trois groupes de caractéristiques :

- stylométriques (nombre de mots dans la phrase, n-grammes, ponctuation, partie du discours, dépendances);

- lexicaux (des dictionnaires de mots d'opinion, de mots partiaux, subjectifs, etc.);
- non supervisés (une modélisation du sujet de chaque document à l'aide d'un partitionnement en K sujets, et un modèle de plongements de mots normalisé avec le tf-idf de chaque mot).

Les résultats démontrent qu'une matrice trop grande nuit à la classification de phrases issues de domaines largement différents. De plus, les caractéristiques trop spécifiques à un domaine particulier provoquent le surapprentissage. Des trois types de caractéristiques testés, les plongements de mots semblent le plus performants pour la tâche de classification de phrases.

3.3 Classification

Recasens *et al*. (2013) sont parmi les premiers à publier des résultats sur la tâche de détection de biais dans Wikipédia. La précision de leur classifieur est souvent citée en tant que point de référence dans les expériences qui ont suivi. Cependant, leur tâche ne visait pas à classer une phrase comme partiale ou impartiale, mais à trouver le mot partial dans chaque phrase donnée. Pour ce faire, ils ont dû assigner à chaque mot de la phrase une probabilité d'appartenance à la classe positive (mots partiaux). Ensuite, ils ont sélectionné comme candidats les un, deux ou trois mots dont la probabilité était la plus élevée. Selon le nombre de candidats retournés, la précision de leur modèle de régression logistique variait entre 0,34 et 0,58. Ces résultats étaient comparables à la précision avec laquelle des humains avaient effectué la même tâche, soit entre 0,37 et 0,59.

Hutto *et al*. (2015) ont évalué la pertinence de chacune des 26 caractéristiques utilisées comme représentation vectorielle par Recasens *et al*. (2013) lors de leurs expériences sur le même corpus. À l'aide d'un modèle de régression linéaire, ils ont éliminé les 12 moins performantes. Cet espace vectoriel réduit, auquel ils ont rajouté de nouvelles caractéristiques, a procuré une précision de 0,95 et un rappel de 0,85 dans une tâche d'identification du degré de partialité de la phrase.

Vincze (2013) a identifié les marqueurs d'incertitude dans un corpus de Wikipédia manuellement annoté selon une méthode déterministe à base de dictionnaires. Dans son expérimentation de détection du token partial dans une phrase partiale donnée, si un candidat était annoté comme marqueur d'incertitude dans au moins 50% de ses occurrences dans le corpus, alors il était mis dans la classe positive. Les résultats de la détection des mots flatteurs (P=0,40; R=0,53; F_1=0,45) indiquent qu'avec cette approche, ces marqueurs sont les plus difficiles à identifier, à cause de leur ambiguïté élevée. Par exemple, le mot-forme ***most*** serait évasif dans la phrase ***Most*** *agree that…*, flatteur dans *...the* ***most*** *touristic beach.*, marqueur d'incertitude dans *...**most** of the time* et neutre dans *He did the **most** he could*. En contrepartie, il est possible de relever les marqueurs d'incertitude (*hedges*) de façon assez précise (P=0,91; R=0,71; F_1=0,80) avec rien qu'un dictionnaire, ce qui suggère qu'ils sont moins ambigus.

Bellows (2018) a obtenu une précision de 0,68 sur un corpus équilibré de 2 143 phrases (partiales et neutres) tirées d'articles de journaux, vectorisées selon la méthode tf-idf et classifiées dans deux classes par un modèle de classification naïve bayésienne. Elle rapporte également des précisions de 0,77 et 0,78 obtenues par un réseau de neurones convolutif et par un réseau récurrent, respectivement.

En utilisant un algorithme de forêts d'arbres décisionnels sur un corpus de 686 phrases de *Conservapedia* manuellement annotées, Hube *et al*. (2018) obtiennent une précision de 0,74 et un rappel de 0,66 lors de la classification de phrases partiales. Les caractéristiques les plus discriminantes dans leur cas ont été le pourcentage de mots-clés par phrase, le contexte (dans une fenêtre de trois mots) et les caractéristiques extraites à l'aide des dictionnaires de LIWC.

La détection de partialité dans Wikipédia et ailleurs constitue un champ de recherche actif depuis plus que dix ans. L'étude du phénomène a produit plusieurs corpus d'exemples annotés, des résultats d'enquêtes ou bien, des analyses linguistiques des formes diverses d'expression de biais. Cependant, la majorité des approches proposées jusqu'à maintenant ont été développées pour l'anglais, sans la possibilité d'être validées dans une autre langue. D'où notre

volonté de développer une heuristique d'extraction d'exemples annotés indépendante de la langue d'entrée. Nous avons opté pour la construction automatique de corpus au détriment du contrôle de la qualité que l'annotation manuelle peut offrir, dans le but de limiter les frais de production, mais aussi pour maximiser la taille des corpus résultant pour qu'ils soient exploitables par les nouveaux algorithmes d'apprentissage automatique.

Chapitre 4 Corpus

Pour les besoins de notre étude de la partialité, nous avons développé une procédure d'extraction automatique d'exemples étiquetés d'un Wikipédia en n'importe quelle langue, que nous avons ensuite appliquée sur les sauvegardes des Wikis en bulgare, français et anglais du mois d'avril 2019[24]. Le code (en Python) permettant de refaire nos expérimentations ou de traiter d'autres langues est public[25]. La chaîne de traitement détaillée dans ce chapitre comprend les étapes principales suivantes (dont la première se fait manuellement) :

1. Dresser une liste de balises relatives à la politique de neutralité dans la langue cible (manuellement).
2. Télécharger le ou les fichier(s) d'une sauvegarde contenant tous les articles et l'historique complet de leurs révisions.
3. Extraire des paires de révisions contenant une des balises.
4. Nettoyer le texte du balisage et le segmenter en phrases et en tokens.
5. Filtrer les révisions.
6. Comparer les révisions de chaque couple pour n'en extraire que les différences.
7. Filtrer les phrases.
8. Équilibrer les deux classes, étiqueter les exemples, segmenter le corpus en trois parties (entraînement, développement et validation) et sauvegarder.

Il existe plusieurs informations que le format structuré de Wikipédia permet d'extraire relativement facilement : des dictionnaires bilingues (Yu & Tsujii, 2009), des relations sémantiques (Al-Rajebah & Al-Khalifa, 2010; Nakayama et al., 2008; Wang et al., 2007), des relations géographiques (Blessing & Schütze, 2010), des données structurées (Morsey, 2012), des locutions (Bekavac & Tadic, 2008) et même des faits divers (Tsurel et al., 2016). Dans tous ces cas, le contenu ciblé est plus ou moins identifiable grâce à sa position dans la structure XML

[24] https://dumps.wikimedia.org
[25] https://github.com/crim-ca/wiki-bias

d'un article (qui distingue la date de révision, le titre d'un article, etc.) ou à la syntaxe du HTML qui organise le contenu de l'article et qui permet de reconnaître facilement les adresses URL externes, les liens internes, les images, les en-têtes, etc. Contrairement à ces informations, la partialité n'a pas de forme fixe et facilement identifiable. On retrouve des biais à l'intérieur de phrases isolées, dans des paragraphes complets ou bien dans des phrases dispersées à travers un article. Ainsi, le seul indice de la présence de ton non neutre est la balise « NPV » insérée par des contributeurs dans l'article. Son rôle de mise en garde, par contre, ne laisse pas savoir où exactement se trouve le problème. Ce n'est qu'à la suite d'une révision visant à réinstaurer la neutralité que l'on peut réellement retracer les passages concernés en comparant les deux états (c'est-à-dire avant et après la révision) de l'article. Ainsi, la liste d'articles à neutralité contestée (à l'heure actuelle) que l'on retrouve sur Wikipédia[26] ne peut pas servir à l'extraction d'exemples, car les passages non-neutres n'ont pas encore été identifiés (Herzig et al., 2011). Pour extraire des révisions pertinentes, Recasens *et al*. (2013) ont compté sur les commentaires que les contributeurs avaient laissés lors de leur intervention (cf. §3.1.4). Si un commentaire faisait mention de la politique de neutralité, les auteurs conservaient la révision en question et la version précédente du même article. Mais comme la possibilité de laisser un commentaire est optionnelle, cette approche ne retourne qu'une partie des révisions pertinentes pour l'étude de la neutralité. Nous proposons alors une méthode qui s'appuie sur les balises de neutralité contestée incluses dans le texte de l'article.

4.1 Balises de neutralité contestée

La première étape consistait à dresser la liste de balises relatives à la politique de neutralité de point de vue (*p. ex*. `{{POV}}`, `{{NPOV}}`, `{{neutral point of view}}`, `{{peacock}}`, etc.) pour chacune des langues à l'étude (anglais, français et bulgare). En anglais et en français, il existe une page de maintenance[27] qui donne la liste de (certaines)

[26] En anglais: https://en.wikipedia.org/wiki/Category:NPOV_disputes
en français:
https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Article_soup%C3%A7onn%C3%A9_de_partialit%C3%A9

[27] En anglais: https://en.wikipedia.org/wiki/Category:Neutrality_templates
en français: https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Mod%C3%A8le_de_non-neutralit%C3%A9

balises NPV. Par contre, elle n'existe que pour 13 autres langues, excluant le bulgare. Alors, nous avons dû dresser à la main la liste des balises en bulgare, en parcourant les articles à neutralité contestée[28]. Cette lacune dans les pages de maintenance en langue bulgare n'était qu'un des obstacles à l'automatisation de cette étape de la chaîne de traitement. Nous avons constaté également de la variabilité dans l'orthographe de certaines balises, décrite dans leurs pages détaillées. En anglais, par exemple `{{POV}}` signale un problème de neutralité au même titre que les balises `{{Npov}}`, `{{Pov}}`, `{{NPOV}}`, `{{Neutrality}}`, `{{Point Of View}}`, `{{PoV}}`, `{{Neutral}}`. En prenant en compte les variantes décrites dans la documentation de Wikipédia, nous avons initialement recueilli 45 balises de neutralité contestée en anglais, 24 en français et 19 en bulgare (cf. Annexe 1).

Mais comme Wikipédia est une plateforme collaborative, il y a plus de variation dans l'usage réel des balises que dans leur documentation. Par exemple, parmi les 16 variantes les plus fréquentes de la balise `{{weasel}}` données dans le Tableau 4 , il n'y en a que cinq (en gras) qui sont documentées. Les autres représentent 35% de l'usage de cette balise qui signale la présence de contenu évasif et non neutre.

Balise	Occurrences	Fréquence relative
weasel	201 092	0,5748
weasel-inline	89 352	0,2554
weasel words	21 755	0,0622
weasel word	16 991	0,0486
weasel section	3 954	0,0113
weasel-section	3 743	0,0107
weasel inline	2 631	0,0075

28 https://bg.wikipedia.org/wiki/%D0%9A%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D0%B8%D1%8F:%D0%A1%D0%BF%D0%BE%D1%80%D0%BD%D0%B0_%D0%BD%D0%B5%D1%83%D1%82%D1%80%D0%B0%D0%BB%D0%BD%D0%BE%D1%81%D1%82

weaselinline	2 213	0,0063
weasel-words	2 176	0,0062
weasel-word	2 102	0,0060
weaselword	1 967	0,0056
weasel-name	956	0,0027
weaselwords	503	0,0014
weasel_section	225	0,0007
weasel_words	124	0,0004
weasel_word	80	0,0002

Tableau 4. Les principales variantes de la balise `{{weasel}}` en anglais

Bien qu'il soit facile pour les humains d'interpréter la signification de ces variantes, il n'est pas trivial d'identifier automatiquement toutes celles qui sont liées à la politique de neutralité. Par exemple, l'extraction naïve de toute balise commençant par *weasel* procure des résultats sans rapport, comme `{{weasel, back-striped}}` (un animal) ou `{{weasel, ben}}` (un chanteur punk). Pour cette raison, nous avons d'abord compilé des listes exhaustives de balises (avec leur fréquence relative), puis nous avons sélectionné manuellement les variations pertinentes de chacune. Cette procédure nous a permis d'enrichir considérablement les listes de balises que nous avions au début, comme le montre le Tableau 5.

	Liste initiale	**Liste enrichie**
BG	11	19
FR	18	74
EN	45	201

Tableau 5. La taille des listes de balises par langue

4.2 Extraction de révisions

Un des moyens d'obtenir les articles de Wikipédia dans une langue donnée, y compris l'historique exhaustif de leurs révisions, est de télécharger une des sauvegardes (disponibles en 285 langues) offertes sans frais par Wikimédia[29]. Le nombre et la taille des fichiers peuvent varier largement en fonction de la langue. La sauvegarde la plus complète en bulgare, par exemple, comptait un seul fichier (compressé en format BZIP2) d'environ 4 Go qui contenait un fichier XML de 25 fois sa taille (~100 Go). Le contenu en français, par contre, était séparé en plus de 100 fichiers de taille variable, tandis que le contenu en anglais (qui est aussi l'archive Wikipédia la plus grande) comptait plus de 600 fichiers. La taille considérable des fichiers décompressés a nécessité la conception d'un traitement à la volée permettant d'éviter l'écriture des fichiers sources.

Dans la sauvegarde en format XML, on retrouve au niveau le plus haut toutes les pages de l'encyclopédie, y compris les articles, les pages de discussion, les pages de maintenance, la documentation, etc., où chaque type de page porte un identificateur différent. Nous avons donc procédé à un premier filtrage pour ne conserver que les pages de type `ns01` (*name space 01*), qui correspondent aux articles. Enchâssées sous chaque article, on trouve ses révisions ($r_1, r_2, \ldots, r_N$) en ordre chronologique, de la plus récente à la plus ancienne. Nous avons procédé à un filtrage des révisions de chaque article, deux par deux (r_i, r_{i+1}) afin de trouver des paires où la version plus ancienne (r_{i+1}) contient une balise NPV, mais la version plus récente (r_i) n'en contient plus (Figure 2).

[29] https://dumps.wikimedia.org/

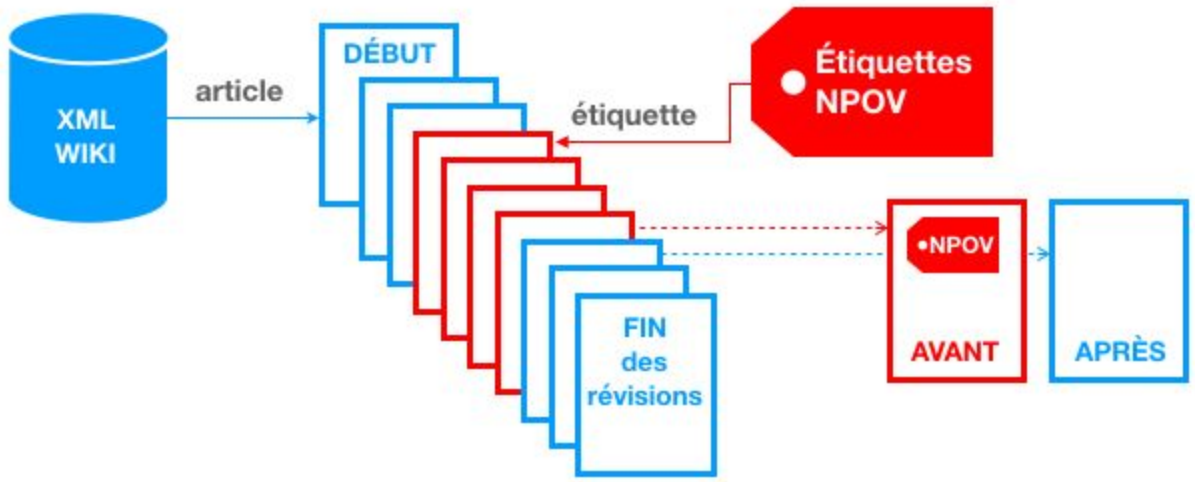


Figure 2. Illustration de la procédure de sélection des paires de révisions contenant des exemples de neutralité contestée

Nous supposons que cette disparition de la balise correspond à des changements qui ont visé les problèmes de neutralité, soit en réécrivant les passages concernés, soit en enlevant ou en rajoutant de l'information. Nous avons alors extrait ces paires de révisions (cf. Annexe 2 pour un exemple) accompagnées des balises qui y étaient incluses. Le Tableau 6 donne le nombre de révisions parcourues et le nombre de paires de révisions extraites par langue. On peut aussi voir que dans le cas de l'anglais et du français, la proportion de paires de révisions qui remplissent notre condition se situe autour de 0,03 à 0,04%. En bulgare par contre, il semble y avoir environ trois fois moins d'exemples.

	Révisions parcourues	Paires de rév. extraites	%
EN	577 164 752	197 953	0,03
FR	105 695 459	46 331	0,04
BG	7 125 897	1 021	0,01

Tableau 6. Nombre de révisions parcourues et nombre et proportion des paires de révisions extraites par langue

4.3 Traitement et filtrage des paires de révisions

Comme notre algorithme cherche à extraire des révisions successives où seulement la première contient une balise NPV, il extrait également des paires de révisions provenant d'un article qui a fait l'objet d'une redirection ou qui a subi un acte de vandalisme (le plus souvent, l'effacement complet du texte ou son remplacement par un message non pertinent), car dans ces cas, la balise a été effacée. Nous avons donc exclu toute paire où la deuxième révision était trop courte (< 400 caractères).

Ensuite, nous avons procédé au nettoyage des paires de révisions à l'aide d'expressions régulières afin d'extraire que le texte (y compris la partie textuelle des liens) et la ponctuation essentielle. Nous avons supprimé les éléments suivants (dans cet ordre) :

- les titres de section
- les références : `<ref ...>...</ref>`
- les guillemets doubles de formatage : `''`
- les étiquettes : `<tag ...>...</tag>`
- les sauts de ligne : `<br/>`
- les commentaires HTML : `<!--...-->`
- les balises HTML vides : `<.../>`

- les liens (externes) HTML, tout en préservant le texte visible
- les liens (internes) Wiki, tout en préservant le texte visible
- les liens vers le même article écrit en d'autres langues : `[[langue: titre]]`
- les liens vers les catégories auxquelles appartient l'article
- les adresses URL : `http(s)://...`
- les icônes et commentaires imbriqués : `{{...({{...}})...}}`
- les tableaux : `{...}`
- les titres d'images, de figures et autres : `[...|...]`
- les accolades orphelines : `{}`
- les crochets orphelins : `[]`
- les balises HTML orphelines
- les caractères unicode HTML : `&...;`
- les puces des listes non ordonnées et les sauts de lignes : `*`
- les chiffres (remplacés par un token `NUMTKN`)
- les points multiples
- les espaces multiples
- les points d'exclamation multiples

Une fois le texte nettoyé, nous l'avons segmenté en tokens et en phrases à l'aide de méthodes par règles de la librairie spaCy (Honnibal & Montani, 2017). Nous avons d'abord envisagé les méthodes de segmentation probabilistes de spaCy dépendantes de modèle de langue préentraîné. Cependant, lorsque nous avons comparé la performance des deux types de méthodes sur le segment test des corpus français FTB, GSD et Sequoia de la collection UD[30] pour les tâches de segmentation en tokens et en phrases (Tableau 7), nous avons observé une précision comparable, alors que la méthode fondée sur des règles était 30% plus rapide lors de la segmentation en phrases (un gain non négligeable étant donné la taille des corpus à traiter).

[30] https://universaldependencies.org/

		FTB		GSD		SEQ	
		régles	modèle	règles	modèle	règles	modèle
Segmentation en tokens	Précision	0.747	0.747	0.875	0.873	0.850	0.847
	Rappel	0.867	0.869	0.929	0.930	0.927	0.927
	F-mesure	0.803	0.803	0.901	0.900	0.886	0.885
	Temps (*min*)	0.24	0.23	0.14	0.15	0.14	0.16
Segmentation en phrases	Précision	0.896	0.859	0.977	0.905	0.890	0.887
	Rappel	0.831	0.964	0.932	0.973	0.695	0.938
	F-mesure	0.862	0.908	0.954	0.938	0.781	0.912
	Temps (*min*)	0.23	0.48	0.14	0.21	0.14	0.21

Tableau 7. Résultats de la segmentation de trois corpus français en tokens et en phrases avec spaCy

De plus, la version actuelle de spaCy (2.1.3) ne propose des modèles linguistiques préentraînés que pour une poignée de langues, à l'exclusion du bulgare. En revanche, la même version offre des méthodes déterministes de segmentation adaptées à 51 langues[31], ce qui la rend très utile à notre approche multilingue. Nous avons dû rajouter une liste d'exceptions à la fonction de segmentation en phrases pour la langue bulgare, car les points suivant plusieurs contractions courantes dans cette langue (*p. ex.* с т р., т. н а р., г., etc.) étaient pris à tort pour des fins de phrase. Comme la segmentation de phrases à base de règles n'est pas parfaite, nous avons exclu les plus courtes (de moins de cinq tokens) ainsi que les plus longues (de plus de 300 tokens, incluant la ponctuation). Un examen qualitatif d'une portion de ces phrases nous a permis de confirmer leur caractère erroné, à cause d'imperfections de la segmentation et/ou

[31] https://spacy.io/usage/models

du texte source. À la suite de la segmentation en phrases, nous avons supprimé la ponctuation restante et nous avons tout converti en minuscules. Un autre filtrage par longueur (entre cinq et 300 tokens) a été appliqué, cette fois en absence de ponctuation.

Nous avons ensuite comparé les deux révisions r_i et r_{i+1} de chaque paire afin d'en extraire les phrases différentes, c'est-à-dire les phrases que les deux révisions n'avaient pas en commun (Figure 3). Ainsi, la révision r_i nous a fourni les phrases enlevées ou corrigées qui, selon notre hypothèse, étaient partiales, et la révision r_{i+1} nous a donné des phrases rajoutées que l'on considérait (probablement, mais pas assurément) neutres.

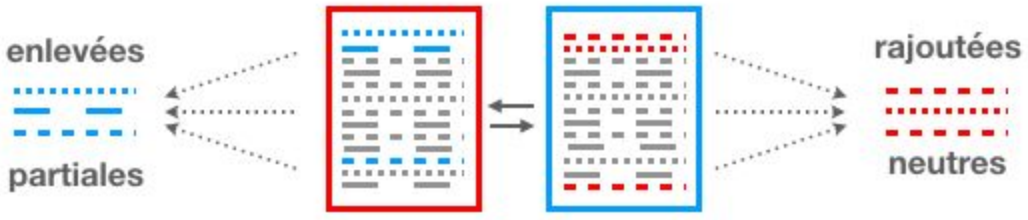


Figure 3. Comparaison des phrases d'une paire de révisions et extraction des phrases partiales et neutres.

Il importe de noter que certaines révisions n'introduisaient que des insertions ou des effacements. Suite à une analyse des exemples, nous avons décidé d'exclure ces paires de révisions du corpus pour deux raisons :

- la qualité des exemples que les simples effacements fournissaient était inférieure à celle des exemples sortis de révisions plus complexes
- les simples insertions ne fournissaient pas d'exemples à la classe positive de phrases partiales (à laquelle nous étions particulièrement intéressés)

Nous avons ensuite examiné les changements entre chaque paire de révisions successives en fonction du nombre de phrases modifiées. Dans environ 20% des cas pour une langue donnée,

la seule différence entre deux états d'un article résidait dans l'absence de la balise NPV dans la version plus récente. Autrement dit, le problème (potentiel) de neutralité a été éliminé par l'élimination de la signalisation. Selon nous, ces cas témoignent de guerres d'opinion (Sumi et al., 2011; Yasseri et al., 2012) qui, tout comme les transgressions de la politique de neutralité, affectent de façon disproportionnée les articles des sujets controversés. Dans un autre 20% des révisions initialement extraites, nous n'avons constaté que des différences de ponctuation et/ou de capitalisation de lettres que nous avions déjà masquées lors de l'étape de nettoyage. Ensuite, nous avons constaté que dans un petit nombre de révisions (que nous avons alors écartées), plus de 400 phrases avaient été modifiées suite à des actes de vandalisme (Dutrey et al., 2011; Yang et al., 2017). Enfin, nous avons éliminé toutes les paires de phrases à différences mineures (dont la distance Levenshtein était égale à 1), qui contenaient des corrections de ponctuation ou d'orthographe et non de partialité. Le Tableau 8 donne la diminution du nombre initial de paires de révisions par langue selon les différents critères de filtrage.

Paires de révisions	**BG**	**FR**	**EN**
nombre initial	1 021	46 331	197 953
guerre éditoriale	−257	−10 255	−61 397
ponctuation/capitalisation	−194	−5 967	−44 345
redirection/vandalisme	−56	−1 524	−17 154
effacements	−33	−2 740	−11 331
insertions	−28	−2 819	−2 938
orthographe	−3	−136	−400
vandalisme	−2	−153	−609
Nombre final	448	22 737	59 779

Tableau 8. Filtrage des paires de révisions par langue

Une mise à jour des proportions de paires de révisions conservées après le filtrage (Tableau 9) nous permet d'approximer l'étendue du contenu partial dans Wikipédia. Soulignons qu'il ne s'agit que du biais remarqué et signalé et cela, dans une perspective diachronique. Notre estimation du nombre de pages biaisées diffère largement de celle de Hube et Fetahu (2018) par rapport aux phrases biaisées même si les deux ne sont pas directement comparables. Leur approximation est probablement exagérée, car leur classifieur a été entrainé sur des exemples de partialité plus explicites que ceux rencontrés dans une encyclopédie. Quant à la notre, elle sousestime la quantité réelle de partialité en se limitant aux révisions signalées. Or, pour correctement quantifier le phénomène en incluant tout biais ignoré ou passé inaperçu, il faut recourir à un classifieur comme celui que l'on est en train de construire (ou bien, approximer la proportion à partir d'un échantillon manuellement annoté).

	Révisions parcourues	**Paires de rév. extraites**	**Paires de rév. conservées**	**%**
EN	577 164 752	197 953	59 779	0,01
FR	105 695 459	46 331	22 737	0,02
BG	7 125 897	1 021	448	0,006

Tableau 9. Nombre de révisions parcourues, nombre de paires de révisions extraites et nombre et proportion des paires de révisions conservées par langue

4.4 Extraction de phrases

Pour construire le corpus final, nous avons placé les phrases enlevées (=partiales) et celles rajoutées (=neutres) dans les classes positive et négative, respectivement. Un filtrage entre et à l'interieur des classes a éliminé les doublons. Nous avons choisi au hasard des phrases inchangées que nous avons rajoutées à la classe négative afin d'équilibrer le jeu de données (voir Figure 4) dans lequel nous avions davantage de phrases partiales.

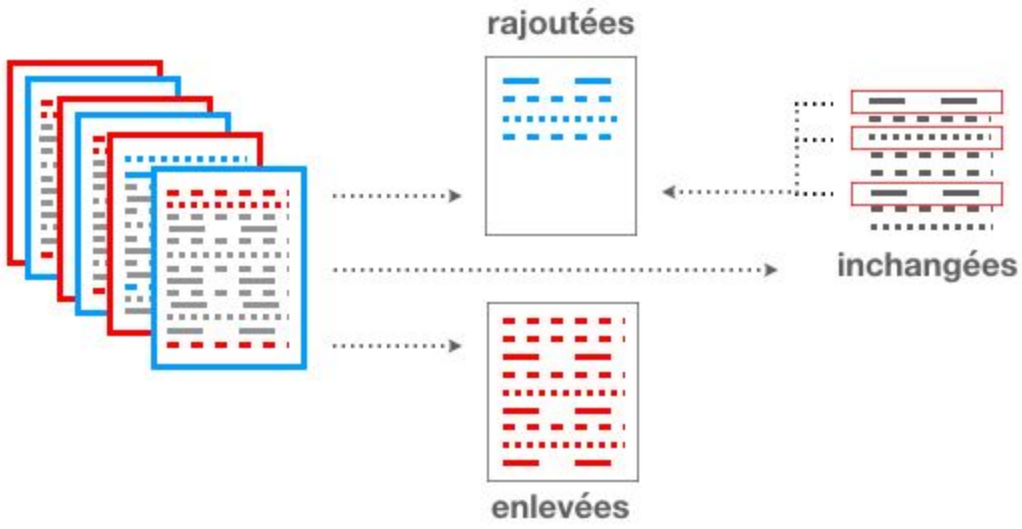


Figure 4. Construction des classes équilibrées du corpus de phrases

Le Tableau 10 ci-dessous donne le nombre de phrases des corpus finaux par type et par langue. Soulignons que notre corpus bulgare est 2 à 4 fois plus grand que ce qui a été créé précédemment pour l'anglais.

Phrases	**BG**	**FR**	**EN**
Enlevées (=partiales)	4 756	105 939	800 191
Rajoutées (=neutres)	3 288	72 183	494 993
Inchangées (=neutres)	1 468	33 756	305 198
Nombre final	9 512	211 878	1 600 382

Tableau 10. La taille des corpus finaux (en phrases)

Chapitre 5 Évaluation des corpus

Étant donné la création automatisée de nos corpus, il était important d'évaluer manuellement un échantillon des phrases que notre algorithme avait classées partiales, afin de mesurer sa robustesse et par extension, la qualité des corpus. Selon notre hypothèse, les phrases effacées par une révision éliminant en même temps la balise de neutralité contestée auraient de fortes chances d'être des phrases non neutres. Ce sont donc ces phrases que nous avons visées par l'évaluation ci-dessous.

5.1 Analyse préliminaire

Nous avons d'abord examiné la distribution des paires de révisions en fonction du nombre de phrases ciblées par une révision donnée. Comme nous avions déjà limité le nombre de différences entre deux révisions consécutives à 400 phrases ou moins (cf. §4.3), le corpus contenait alors des paires de révisions où entre 1 et 400 phrases avaient été changées d'une révision à l'autre. La figure 4 illustre la distribution zipfienne[32] de ces paires (en français) sur une échelle logarithmique où il devient évident que l'écrasante majorité des révisions n'affecte qu'une à deux phrases. Autrement dit, une grande partie des paires de révisions extraites ne contribuent qu'une à deux phrases à notre corpus, tandis que certaines paires situées dans la queue de la distribution peuvent apporter des dizaines, voir des centaines d'exemples de phrases modifiées entre deux version d'un article. (cf. Annexe 3 pour les distributions des corpus bulgare et anglais.)

[32] À l'exception d'un peu de variation dans la queue de la distribution, attribuable à des actes de vandalisme successifs sur le même article que notre approche diachronique extrait de l'historique des révisions.

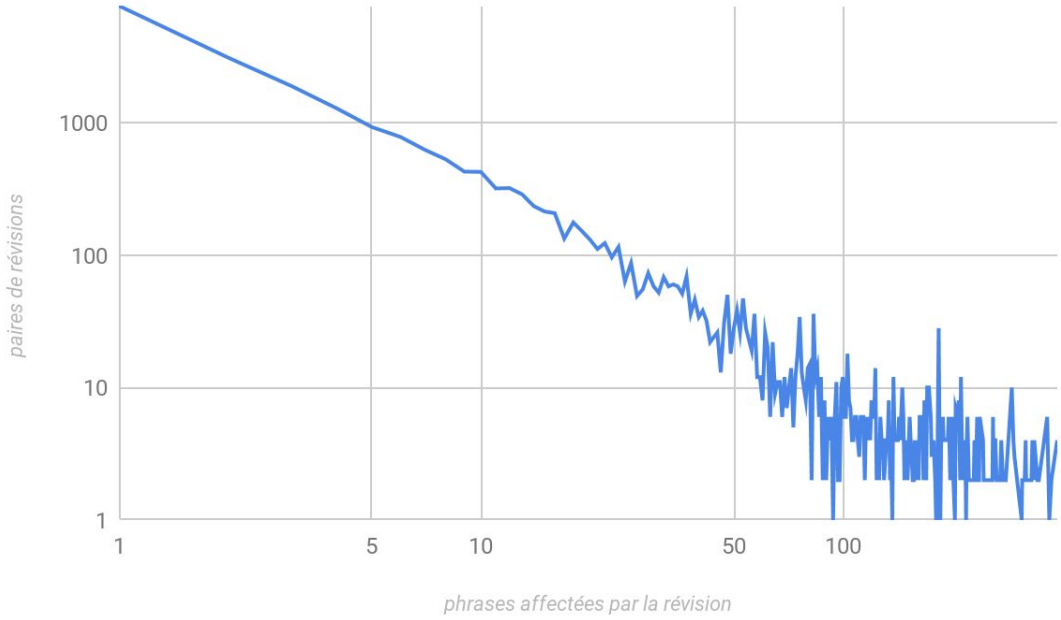


Figure 5. Distribution des paires de révisions (en français) en fonction du nombre de phrases affectées par la révision

Selon notre intuition, la probabilité qu'une phrase extraite par notre algorithme soit véritablement partiale devrait varier en fonction du nombre de phrases affectées par la révision. Moins il y a eu de changements entre une révision étiquetée partiale (r_i) et celle qui fait disparaitre l'étiquette (r_{i+1}), plus il y a des chances que ces changements corrigent un biais et non pas le style, l'orthographe ou tout autre aspect de l'article. Nous avons donc regroupé les paires de révisions dans quatre classes en fonction du nombre de phrases touchées par une révision, ce qui nous a permis également d'échantillonner chaque section de la distribution. Les bornes de ces classes ont été définies empiriquement afin d'obtenir des classes de taille comparable en termes de paires de révisions.

- classe 1 : une ou deux phrases
- classe 2 : entre trois et six phrases
- classe 3 : entre sept et quinze phrases

- classe 4 : plus de seize phrases

La figure 6 montre la segmentation (agrégée pour les trois langues) en classes plus au moins égales en nombre de révisions, mais très déséquilibrées quant à leur contribution de phrases au corpus. Par exemple, on retrouve dans la première classe 35% des paires de révisions et seulement 2% des phrases du corpus, tandis que la quatrième classe contient 26% des paires de révisions mais fournit 84% de toutes les phrases. (cf. Annexe 4 pour les données par langue.)

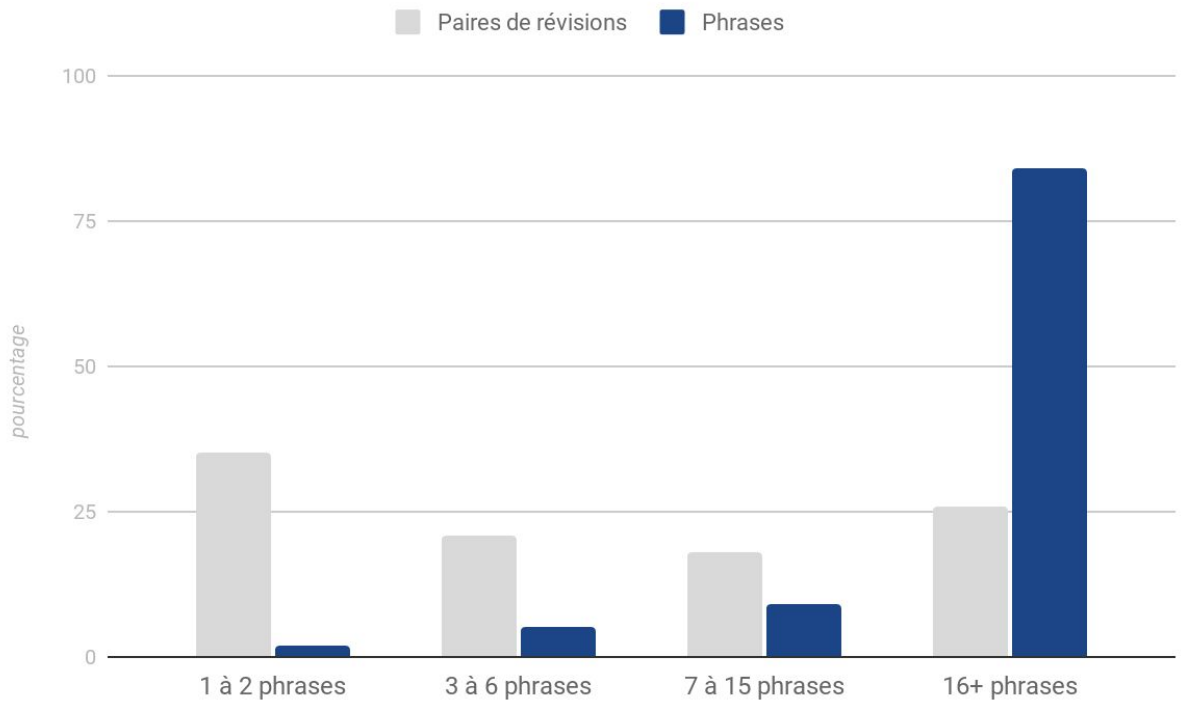


Figure 6. Proportions de révisions et de phrases par classe (agrégation des trois langues)

De chaque classe, nous avons choisi aléatoirement 74 phrases jugées partiales par notre algorithme d'extraction (pour un total de 296 phrases) que nous avons fait annoter par 2 ou 3 annotateurs par langue. 24% (72) des phrases dans une langue donnée étaient identiques pour tous les annotateurs de la langue, alors que les autres étaient uniques. Cette configuration (Tableau 11) nous a permis de calculer l'accord inter-annotateur (AIA) sans pour autant limiter

la couverture de l'évaluation (en termes de phrases uniques annotées). Les échantillons des corpus en bulgare et en français ont été annotés chacun par deux locuteurs natifs. L'échantillon du corpus anglais a été traité par trois personnes ayant une compétence quasi-native. Étant donné la taille très variable des trois corpus, la couverture des échantillons que nous avons pu faire évaluer est, elle aussi, très variable; de 7% en bulgare à seulement 0,06% en anglais.

Langue	Tous	Ann1	Ann2	Ann3	Total	%
BG	72	224	224	—	520	7,23
FR	72	224	224	—	520	0,31
EN	72	224	224	224	744	0,06

Tableau 11. Nombre de phrases assignées par annotateur et proportion des phrases évaluées de l'ensemble de phrases automatiquement étiquetées comme partiales

5.2 Protocole

Les annotateurs ont reçu des instructions identiques en anglais (cf. Annexe 5) détaillant les aspects de la politique de neutralité de point de vue de Wikipédia (cf. §2), ainsi que des exemples de mots à éviter (en anglais). Pour chaque phrase de leur échantillon, ils devaient dire si elle violait la politique NPV ou pas. Les annotateurs recevaient une par une les phrases à annoter en contexte, accompagnées par les deux états de la révision (avant et après), juxtaposés. Ils étaient donc libres de consulter la réécriture de la phrase jugée partiale ainsi que d'examiner son contexte. (cf. Annexe 6 pour un exemple abrégé de fiche d'annotation.)

5.3 Résultats de l'évaluation

Comme nous avions trois annotateurs pour l'anglais, nous avons utilisé le kappa de Fleiss (κ) pour mesurer l'AIA. Le tableau 12 affiche le taux de phrases évaluées comme partiales (à gauche) et l'AIA (à droite) par langue et par classe. Les deux dernières lignes donnent

respectivement la moyenne (*moy.*) et l'écart type (*s*) des annotations par langue, tandis que les deux dernières colonnes de chaque tableau résument les résultats par classe. Ainsi, on peut y lire que la proportion des phrases véritablement partiales est la plus élevée en bulgare (0,56), mais c'est aussi la langue pour laquelle le désaccord entre les annotateurs semble être le plus grand – avec un écart type relativement élevé de 0,13 et un accord inter-annotateur de 0,16. Des trois langues, les résultats obtenus de l'annotation de phrases en anglais se situent le plus proche de la tendance centrale pour les deux mesures. Les moyennes des quatre classes varient entre 0,44 et 0,52 et se caractérisent par une faible dispersion (*s*=0,03). À l'intérieur des classes, la variabilité semble le plus élevée pour la quatrième classe avec un écart type de 0,12 entre les résultats des trois langues.

En moyenne, pour l'ensemble des langues et des classes, les annotateurs ont confirmé le caractère partial pour 48% des phrases dans leur échantillon, avec un AIA global de 0,41. Si l'on ne tient pas compte de la classe 4 en bulgare (la seule avec un kappa de Fleiss négatif), on obtient un taux moyen de phrases partiales de 47% (*s*=0,08) et une valeur moyenne de κ de 0,46 (*s*=0,14).

Classe	**BG**	**EN**	**FR**	***moy.***	***s***	**BG**	**EN**	**FR**	***moy.***	***s***
1	0,34	0,51	0,47	0,44	0,07	0,32	0,55	0,67	0,51	0,15
2	0,64	0,45	0,45	0,52	0,09	0,22	0,58	0,44	0,41	0,15
3	0,63	0,45	0,38	0,48	0,11	0,32	0,31	0,61	0,41	0,14
4	0,63	0,52	0,34	0,50	0,12	−0,23	0,39	0,68	0,28	0,38
moy.	0,56	0,48	0,41	0,48	0,06	0,16	0,46	0,60	0,41	0,18
s	0,13	0,03	0,05	0,03	0,10	0,23	0,11	0,10	0,08	0,21

Tableau 12. La proportion des phrases évaluées comme partiales par classe (à gauche) et l'accord inter-annotateur (kappa de Fleiss) (à droite)

Les coefficients d'accord inter-annotateur que nous avons obtenus sont comparables à ceux mesurés par Vincze (2013) qui avait fait annoter 200 articles de Wikipédia en anglais par deux linguistes pour des mots évasifs (κ=0,48), des mots flatteurs (κ=0,45) et pour des expressions de doute (κ=0,46). Ils sont également supérieurs à l'AIA de 0,35 rapporté par Hube et Fetahu (2018), qui avaient externalisé l'annotation de phrases de Conservapedia en deux catégories (partiales ou pas). Enfin, ils sont légèrement inférieurs aux résultats de Yano *et al*. (2010), qui ont mesuré un accord de 0,55 entre les annotateurs de phrases extraites de blogues politiques américains qu'il fallait classer comme impartiales, légèrement partiales ou très partiales. Les résultats soulignent la complexité de la tâche pour les humains, dont les biais inhérents obscurcissent souvent la perception. Comme l'ont observé Yano *et al*. (2010), les personnes sont enclines à considérer une phrase comme partiale lorsqu'elles sont en désaccord avec son propos. Inversement, elles se montrent clémentes envers la partialité lorsque cette dernière est conforme à leurs convictions. Toujours selon eux, le nombre élevé de phrases annotées (dans leur expérience) avec l'étiquette *I think it is biased, but am not sure which side* indique que les participants sont souvent capables de détecter un parti pris, sans pour autant en comprendre la motivation en l'absence de tout contexte.

Environ la moitié des phrases annotées s'avèrent neutres à cause de différents facteurs et sources de bruit dont nous discutons brièvement ci-dessous. (cf. Annexe 7 pour les résultats de l'annotation par annotateur.) Nous n'avons pas constaté de corrélation entre le nombre de phrases visées par une rédaction et la probabilité que ces phrases soient véritablement partiales. À l'encontre de notre intuition première, une révision qui ne change qu'une ou deux phrases n'est pas plus susceptible de fournir de bons exemples de phrases partiales qu'une autre qui efface des paragraphes au complet. Ainsi, la classe première contient le plus grand nombre de phrases réellement partiales (selon l'évaluation manuelle) seulement en français. En anglais, la distribution entre les classes est presque uniforme et en bulgare, l'association est inverse de celle en français (Figure 7). Cependant, il faut souligner comme limites importantes de la présente évaluation le petit nombre d'annotateurs par langue ainsi que la couverture restreinte de deux des trois langues.

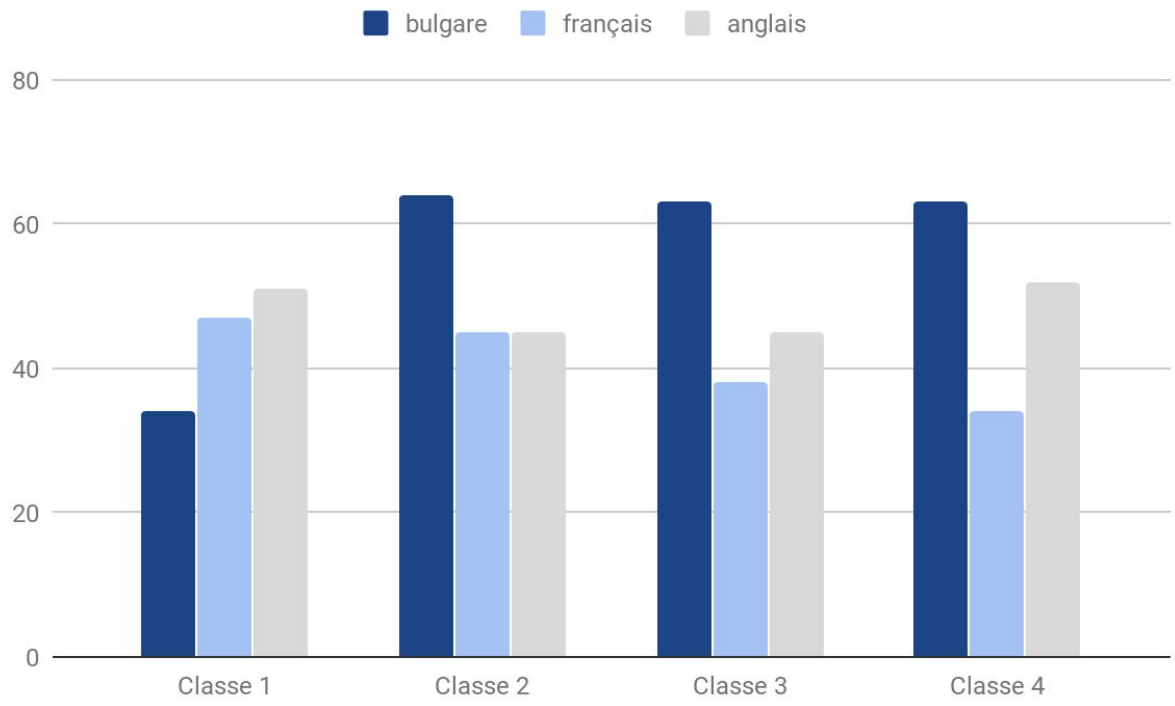


Figure 7. Proportion des phrases partiales par classe et par langue

5.4 Sources de bruit

L'étude du sous-ensemble de phrases choisies pour l'annotation nous a permis d'identifier deux sources de bruit : la procédure d'extraction et de nettoyage, et les rédacteurs.

Le bruit lié à la chaîne de traitement est soit un bruit introduit lors de la phase de prétraitement (par exemple, en raison d'une segmentation incohérente comme dans les phrases (3, 4)[33]), soit un bruit qui subsiste malgré nos efforts de filtrage et de nettoyage, comme en (4). De tels artéfacts sont, par exemple, les modifications (plus longues qu'un caractère) non-liées directement à la question de partialité, mais qui corrigent plutôt des fautes d'orthographe (5) ou de style (6).

(3) a. *(avant)* **anta diop** né le à diourbel mort le à dakar est un historien et anthropologue sénégalais
b. *(après)* **cheikh anta diop** né le à diourbel **<saut de ligne>** mort le à dakar est un historien et anthropologue sénégalais

(4) a. *(avant)* durant la saison des migrations du numtkn septembre au numtkn octobre la tour est en partie éteinte et l intensité de l éclairage est réduit afin de ne pas perturber les oiseaux et la tour **cn image**
b. *(après)* durant la saison des migrations du numtkn septembre au numtkn octobre la tour est en partie éteinte et l intensité de l éclairage est réduit afin de ne pas perturber les et la tour **cn image cntowernastyfall jpg vue plongeante image cn tower numtkn**

(5) a. *(avant)* thérèse avait comme parrain et marraine et d hautefeuille berte et léonide dieudonné marcelle denis emery et marcelle hume
b. *(après)* thérèse avait comme parrain et marraine et d hautefeuille berte et léonide dieudonné **et** marcelle denis emery et marcelle hume

(6) a. *(avant)* danielle arbid est une réalisatrice de films **franco libanaise**
b. *(après)* danielle arbid est une réalisatrice de films **française d origine libanaise**

D'autres exemples de ce type de bruit constituent les restes d'infobox[34], qui sont particulièrement épineux pour le nettoyage par expressions régulières. Enfin, on a observé des

[33] Les paires d'exemples accompagnées d'indication *avant/après* sont tirées de nos corpus.
[34] cf. https://fr.wikipedia.org/wiki/Aide:Infobox

paires de révisions où la correction apportée à une des phrases ne faisait que réécrire un nombre en lettres.

L'autre type de bruit, dû au comportement des rédacteurs de Wikipédia, est plus difficile à éviter. Dans ce groupe, on retrouve l'ajout (intentionnel ou non) de propos partiaux (7, 8); les actes de vandalisme; les corrections d'erreurs factuelles non liées à la question de neutralité; ainsi que le remplacement d'un propos partial par un autre (9). Enfin, on retrouve des exemples de « modifications collatérales ». Nous désignons par ce terme tout changement de phrases neutres voisinnant des phrases problématiques lors de la réécriture d'un passage ou d'un article complet (10).

(7) a. *(avant)* cardinal health inc is a holding company
b. *(après)* cardinal health is a healthcare company **dedicated to making healthcare safer and more productive**

(8) a. *(avant)* kandahar ak est une épreuve de l histoire du ski alpin
b. *(après)* kandahar ak est une épreuve **très importante** de l histoire du ski alpin

(9) a. *(avant)* its support is low only in the cholla province which has for nearly numtkn years supported kim dae jung a well known **leftist** politician born in that province who also served as president of south korea numtkn numtkn
b. *(après)* its support is low only in the jeolla province which has for nearly numtkn years supported kim dae jung a well known **progressive** politician born in that province who also served as president of south korea numtkn numtkn

(10) a. *(avant)* from the numtkn th century confucianism was losing its influence on vietnamese society monetary economy began to develop but unfortunately in negative ways
b. *(après)* from the numtkn th century confucianism was losing its influence on vietnamese society and a monetary economy began to develop

Chapitre 6 Formes d'expression de biais

Dans le contexte de Wikipédia et de sa politique de neutralité de point de vue, le biais des contributeurs peut prendre différentes formes plus ou moins explicites. Recasens *et al*. (2013) ont proposé une classification binaire en biais épistémiques (*epistemological bias*) et biais de cadrage (*framing bias*), où la première classe englobe des types plus implicites que la deuxième.

Le cadrage, un terme employé dans les domaines de la science politique et des communications, se définit comme le processus d'élimination de quelques éléments de la réalité perçue et d'assemblage d'un récit alternatif pour promouvoir une interprétation particulière (Entman, 2007). Au niveau de la phrase, le cadrage consiste à introduire ou à renforcer la saillance de certaines idées par l'ajout d'intensificateurs subjectifs, de vocabulaire partial (*one-sided terms*) ou spécialisé. Les différentes manifestations linguistiques du cadrage, pour la plupart explicites, ont été étudiées dans les travaux sur la détection de subjectivité, d'opinions ou de prise de position (*stance detection*) (Baccianella et al., 2010; C. Lin et al., 2011; Liu et al., 2005; Murray & Carenini, 2009; Riloff & Wiebe, 2003; Wiebe & Riloff, 2005; Wilson & Raaijmakers, 2008; Yano et al., 2010; Yiwei Zhou, 2016). La paire de phrases ci-dessous donne un exemple de propos partisan (apparemment libéral) en (11-a) mis en contraste avec sa forme neutre en (11-b).

(11) a. Without Sestak's challenge, we would have Specter, **comfortably ensconced as** a Democrat **in name only**.
b. Without Sestak's challenge, Specter would have no incentive to side more frequently with Democrats.

Yano et al., 2010

Quant à la partialité épistémique, il s'agit d'une classe introduite par Recasens *et al*. (2013) pour regrouper certaines formes implicites de biais responsables de l'accréditation d'opinions (cf. §2.1.1) ou, à l'inverse, de la discréditation de faits (cf. §2.1.2) moyennant des présuppositions et des implications (*entailments*). Dans le cas des présuppositions, un lexème particulier, un grammème ou une construction servira de déclencheur de sens supplémentaire

que l'auteur laisse entendre sans asserter et qu'il présuppose pour que son énoncé soit significatif dans le contexte en question (Frege, 2003; C. Potts, 2015). Quoique implicites, les présuppositions conventionnelles (par opposition aux conversationnelles) relèvent de certains aspects sémantiques des déclencheurs et peuvent alors être éliminées par une réécriture, lorsque identifiées comme un véhicule de partialité. Par exemple, le guide de style de Wikipédia ne recommande que quelques verbes pour rapporter des propos de façon neutre et exacte (cf. §2.2.7), car il s'agit de verbes exempts de présuppositions. De façon similaire, les implications dépendent du sens des lexèmes. Elles émanent du contenu sémantique de l'énoncé ainsi que de certaines relations sémantiques unilatérales que ses lexèmes entretiennent à l'intérieur du lexique (Bach, 1994, 2006; Vanderveken, 1990). Par exemple, le sens 'assassiner' implique les sens 'tuer' et 'prémédité', sans que cette relation soit réciproque.

À ces deux classes de biais, nous rajoutons celle des implicatures afin d'y classer trois autres formes implicites d'expression de partialité que nous avons recensées lors de la campagne d'évaluation (cf. §5), notamment :

- la narration descriptive
- les omissions et les imprécisions
- l'emploi de la voix active

Grice (1975) fut le premier à définir l'implicature comme l'intention communicative du locuteur, indépendante du contenu propositionnel de l'énoncé et sujette au principe de coopération (Bach, 2006). De cette façon, l'acte de réaliser un énoncé peut, selon la situation, encoder différentes implicatures conversationnelles. Par exemple, le sens de l'énoncé (12) comprend l'implicature implicite en (12-b) ainsi que la proposition explicitement communiquée (12-a).

(12) *Mother (to child crying over a cut on his knee):*
You're not going to die.
a. You (Billy) are not going to die from that cut.
b. You (Billy) should stop making such a fuss about it.

(Bach 1994)

Outre les implicatures conversationnelles, certains auteurs distinguent une forme conventionnelle d'implicatures qui, à l'instar des présuppositions conventionnelles et des implications, relève du sens conventionnel de certaines expressions linguistiques comme *mais* et *donc*. L'intuition derrière les implicatures conventionnelles est qu'elles sont liées à des significations dans la phrase, mais distinctes de l'interprétation compositionnelle de celle-ci (Bach, 1999; Frege, 2003; Grice, 1975; Neale, 1999; C. Potts, 2007, 2015). Potts (2015) en fournit la définition suivante :

> Le sens *p* est une implicature conventionnelle de la phrase *S* ssi :
>
> a. *p* est une propriété conventionnelle (sémantiquement encodée) d'une unité lexicale ou d'une construction dans *S*;
> b. *p* est impliqué par *S*; et
> c. la valeur de vérité de *p* n'a aucun effet sur le contenu de *S.*

Dans la phrase *Jean est énorme, mais agile* le contenu sémantique serait 'Jean est énorme et Jean et agile', tandis que l'implicature conventionnelle correspondrait à 'être énorme empêche l'agilité' où la présence et le contenu de l'implicature sont déterminés (au moins en partie) par le sens conventionnel du lexème *mais*. D'autres exemples d'implicatures conventionnelles, en anglais, sont les adverbes *almost*, *already*, *barely*, *even*, *only*; les épithètes anaphoriques (*the jerk*); les conjonctions *but*, *nevertheless*, *so*, *therefore*; les diminutifs; les jurons; les exclamatives; les conjonctions de subordination *although*, *despite*, *even though*, etc. (voir Potts (2015, p. 188) pour la liste complète accompagnée de références).

En plus de ces trois classes de bias (cadrage, présuppositions, implicatures) sur l'axe pragmatique/sémantique on peut imaginer une autre organisation des formes d'expression de partialité sur l'axe implicite/explicite[35]. Au sujet de cette distinction Carston (2009) avance que

[35] Carston (2009) souligne que les deux distinctions se recoupent: non seulement les processus pragmatiques contribuent au contenu explicite de l'énoncé, mais le sens de certains mots ne contribue pas au contenu explicite, mais fonctionne plutôt comme une contrainte sur les processus inférentiels que doit effectuer le locuteur pour dériver les implicatures.

les locuteurs dérivent pragmatiquement les deux types de propositions, mais que seulement dans le cas des propositions explicites, les inférences sont linguistiquement ancrées. Dans le contexte de détection automatique de biais, cette organisation a l'avantage de nous informer sur le degré de difficulté d'identification que pose chacune des formes d'expression de partialité. Si un biais est identifié comme explicite, il est plus probable de l'associer à un élément dans la phrase que s'il était considéré implicite. Par conséquent, les approches de détection dépendantes de lexiques ne seraient efficaces que dans la détection des formes explicites de partialité.

Nous présentons à la figure 8 une proposition d'organisation des formes d'expression de biais en fonction des deux axes discutées. Nous allons les discuter un par un dans cet ordre dans le reste de ce chapitre.

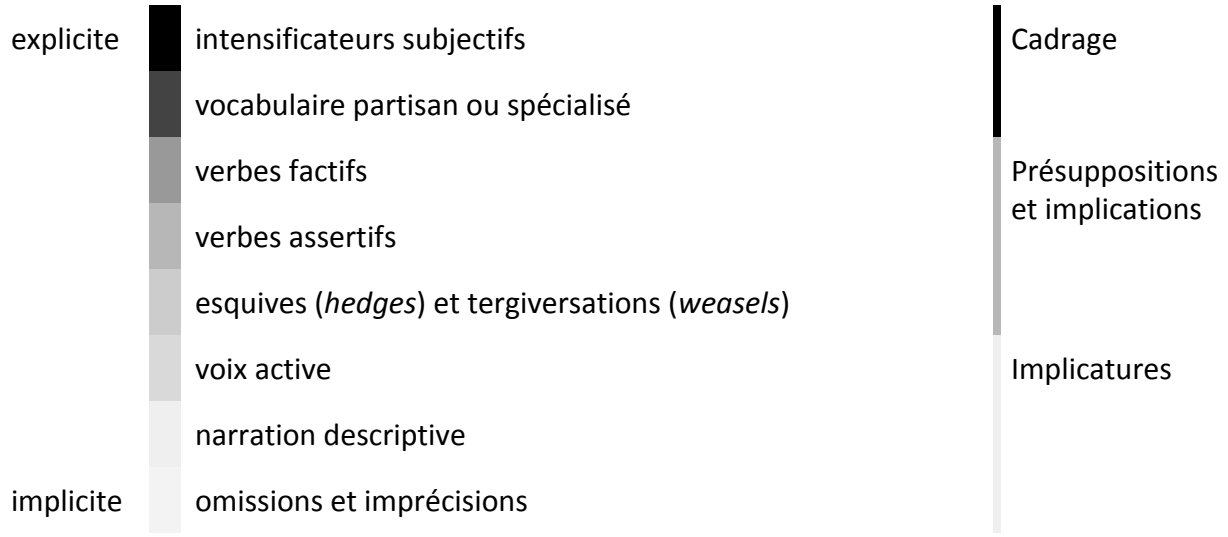


Figure 8. Formes d'expression de biais

6.1 Intensificateurs subjectifs

La classe des intensificateurs subjectifs en tant que porteurs de biais ne se limite pas à un type de modificateur (verbal, nominal, adjectival) ni à une catégorie grammaticale (adverbes, adjectifs) (Athanasiadou, 2007; Bolinger, 2013; Paradis, 2001), mais englobe toute expression dont la fonction est celle d'intensificateur, y compris les superlatifs, les quantifieurs et les mots excessivement positifs ou flatteurs proscrits par Wikipédia (cf. §2.2.5). Les membres de la classe correspondent généralement aux valeurs de la fonction lexicale d'intensification **Magn** (Mel'čuk, 1982; Wanner, 1996).

La manifestation linguistique du point de vue du rédacteur se réalise le plus souvent par l'entremise d'un adverbe modifiant un verbe (13 – 15) ou un adjectif (17) ou d'un adjectif modifiant un nom (15, 16). Mais comme les exemples 14 et 17 le démontrent, la lexicalisation du prédicat verbal peut aussi traduire un parti pris.

(13) a. *(avant)* le succès de la société de tabacologie et de ses journées de tabacologie le souci d indépendance qu elle revendique **ne plaisent guère** aux firmes qui commercialisent les aides pharmaceutiques à la cessation du tabagisme
b. *(après)* le succès de la société de tabacologie et de ses journées de tabacologie le souci d indépendance et l esdprit critique qu elle revendique **suscitent l opposition** de firmes qui commercialisent les aides pharmaceutiques à la cessation du tabagisme

(14) a. *(avant)* some prominent liberals including scott reid **were strongly critical** of volpe s response criticized volpe s response
b. *(après)* some prominent liberals including scott reid **criticized** volpe s response

(15) *(avant)* he is **truly** one of the **greatest** americans

(16) a. *(avant)* this is an **absurd** statement because the cavalry of any age is designed first and foremost to run over the enemy and separate them as to make them far more vulnerable to being overwhelmed and overrun
b. *(après)* this is **wrong** because the cavalry of any age is designed first and foremost to run over the enemy and separate them as to make them far more vulner- able to being overwhelmed and overrun

(17) a. *(avant)* la ville **est bien pourvue** en commerces et services de proximité établis principalement en bordure de la route de rochefort et sur la place des vieilles forges
b. *(après)* la ville **dispose** de commerces et services de proximité établis principalement en bordure de la route de rochefort et sur la place des vieilles forges

(18) a. *(avant)* on y trouve enfin des serres maraîchères **très modernes** occupées par la société arc at plants qui exporte des plants maraîchers à destination des professionnels dans l ensemble de la france et qui emploie jusque numtkn salariés
b. *(après)* on y trouve enfin des serres maraîchères occupées par la société arc at plants qui exporte des plants maraîchers à destination des professionnels dans l ensemble de la france et qui emploie jusque numtkn salariés

6.2 Vocabulaire partisan ou spécialisé

Le vocabulaire partisan (*one-sided terms*) englobe des termes unilatéraux qui ne reflètent qu'une partie d'une question litigieuse ou ceux qui sous-entendent un jugement de valeur. Ils apparaissent dans le contexte de sujets controversé (la religion, le terrorisme, l'avortement, certains conflits localisés, etc.) où le même événement peut être vu sous deux ou plusieurs

angles opposés (19, 20) (W.-H. Lin et al., 2006; Pryzant et al., 2019). Le vocabulaire spécialisé, quant à lui, constitue une sous-classe d'expressions proscrites par la politique de rédaction de Wikipédia qui inclut les clichés, le jargon ainsi que certaines locutions et métaphores. En dehors des considérations de clarté et de compréhension, on proscrit ces expressions au nom de la neutralité à cause de leur caractère subjectif et de leur forte connotation (21).

(19) a. *(avant)* Israeli forces **liberated** the eastern half of Jerusalem.
b. *(après)* Israeli forces **captured** the eastern half of Jerusalem.

(20) a. *(avant)* Colombian **terrorist** groups.
b. (après) Colombian **paramilitary** groups.

(21) *(avant)* x force was **concocted** by illustrator rob liefeld who started **penciling** the new mutants comic book in numtkn

De façon générale, l'emploi du vocabulaire partisan ou spécialisé est discuté dans la littérature sur la détection d'opinion (*stance detection*) puisqu'il trahit la position du locuteur par rapport au sujet concerné (Conrad et al., 2012; Park et al., 2011; Somasundaran & Wiebe, 2010; Yano et al., 2010). Il s'agit d'une forme explicite de prise de position où un choix particulier de vocabulaire transgresse la politique de neutralité de Wikipédia.

6.3 Verbes factifs

En tant que prédicats à deux arguments, les verbes factifs (*découvrir, réaliser, constater, se rendre compte, regretter*, pour n'en nommer que quelques-uns) présupposent la vérité de la proposition qu'ils prennent comme sujet ou comme complément (Kiparsky & Kiparsky, 1968). Par exemple, la proposition en (22-a) présuppose celle en (22-b) :

(22) a. Paul regrette d'avoir mangé un deuxième dessert.
b. Paul a mangé un deuxième dessert.

(exemples inventés)

L'utilisation d'un verbe factif peut alors introduire de la partialité en affirmant la proposition subordonnée (23-a) plutôt que de l'annoncer sans présupposer sa vérité (23-b) (Recasens et al., 2013) :

(23) a. *(avant)* **He realized** that the oppression of black people was more of a result of economic exploitation than anything innately racist.
b. *(après)* **His stand was** that the oppression of black people was more of a result of economic exploitation than anything innately racist.

6.4 Verbes assertifs

Les verbes assertifs (*declare*, *claim*, *acknowledge*, *reply*, *say*, *agree*, etc.), tout comme les verbes factifs, sélectionnent une proposition subordonnée infinitive comme sujet ou complément (Hooper, 1975; Kiparsky & Kiparsky, 1968). Par contre, les verbes assertifs non factifs (pour les distinguer des verbes à la fois assertifs et factifs) ne font qu'affirmer la proposition, sans présupposer sa valeur de vérité. Selon la force de l'affirmation, Hooper (1975) distingue deux sous-classes d'assertifs : les assertifs faibles (*think*, *believe*, *suppose*, *imagine*, *guess*, *appear*, *seem*, *figure*) et les assertifs forts (*admit*, *affirm*, *argue*, *remark*, *reply*, *report, presume*, etc.), dont certains (*say*, *state, suggest*, *answer*, etc.) sont considérés comme neutres par Wikipédia. C'est notamment cette force affirmative qui, selon le contexte, peut introduire un biais de deux façons :

- lorsqu'un verbe assertif faible introduit un fait largement connu ou une proposition dont l'interprétation fait largement consensus (24)

(24) a. *(avant)* Kuypers **claimed** that the mainstream press in America tends to favor liberal viewpoints.
b. *(après)* Kuypers **stated** that the mainstream press in America tends to favor liberal viewpoints.

- lorsqu'un verbe assertif fort introduit une opinion minoritaire (25)

(25) a. *(avant)* The "no Boeing" theory is a controversial issue, even among conspiracists, many of whom have **pointed out** that it is disproved by ...
b. *(après)* The "no Boeing" theory is a controversial issue, even among conspiracists, many of whom have **said** that it is disproved by...

Dans les deux cas, le biais est le résultat d'une discordance de force entre le prédicat affirmatif et la proposition qu'il introduit. Un prédicat trop faible jette un doute sur la vérité de son complément et présente ainsi un fait comme une opinion. À l'inverse, un prédicat trop fort atteste son complément de façon démesurée, ce qui permet de faire passer une opinion pour un fait. L'emploi d'un verbe assertif fort est neutre lorsqu'il introduit une proposition dont la vérité ne prête pas à controverse (26) :

(26) a. *(avant)* Cooper says that slavery was worse in South America and the US than Canada, but **clearly states** that it was a horrible and cruel practice.
b. *(après)* Cooper says that slavery was worse in South America and the US than Canada, but **points out** that it was a horrible and cruel practice.

6.5 Esquives (*hedges*) et tergiversations (*weasels*)

Lakoff (1975) fut le premier à définir les esquives (*hedges* en anglais) comme une classe de mots dont la présence à l'intérieur de la phrase la rend floue (*fuzzy*). Plus précisément, le rôle des esquives consiste à qualifier la confiance de l'auteur dans la vérité de son propos (Bach, 2008; Farkas et al., 2010; Hyland, 1998; Markkanen & Schröder, 1989; Vincze, 2013; Zuck & Zuck, 1986). Le plus souvent, il s'agit de l'implication d'incertitude (27-a) par laquelle l'auteur évite de s'engager dans des affirmations catégoriques (27-b) (Bach, 2008). L'insertion d'incertitude peut alors nuire à la neutralité lorsque l'esquive est utilisée pour discréditer un fait (28, 29) (Recasens et al., 2013).

(27) a. Sheffield, **I believe**, is north of Nottingham.
b. Sheffield is north of Nottingham.

(Bach, 2008)

(28) a. *(avant)* The lower cost of living in more rural areas means a **possibly** higher standard of living.
b. *(après)* The lower cost of living in more rural areas means a higher standard of living.

(29) a. *(avant)* in the first invasion operation litani in numtkn the israeli military and south lebanon army sla occupied a narrow strip of land **ostensibly** as a security zone
b. *(après)* in the first operation litani in numtkn the israel defense forces and south lebanon army occupied a narrow strip of land **described** as the security zone

Une autre façon de déguiser une opinion en fait est de la présenter sans l'attribuer véritablement à une source crédible (Ganter & Strube, 2009; Vincze, 2013). Dans le contexte de Wikipédia, les événements dont la source est manquante ou n'est précisée que de manière vague ou trop générale (30) sont appelés tergiversations (*weasels*).

(30) a. **On** recense davantage de pays pauvres que de pays riches
b. **Certains** affirment…
c. **Les dernières recherches** montrent que…

(exemples inventés)

Camoufler l'identité de l'agent auquel on fait allusion se fait aussi en utilisant la voix passive (31), la nominalisation (32) ou bien le « il » explétif (33).

(31) Cette opinion **est** largement **considérée** comme…

(32) La **coordination des moyens** a permis **la réduction des désagréments**.

(33) **Il** ne fait aucun doute que…

(exemples inventés)

Selon Vincze (2013), la sous-spécification de n'importe quel argument dans la phrase, non seulement le sujet, constitue potentiellement une tergiversation. Dans l'exemple en (33), le déterminant *some* et l'adjectif *other* comblent un manque de précision et introduisent par conséquent de l'incertitude.

(33) While the Skyraider is not as iconic as **some other** aircraft, it has been featured in some Vietnam-era films such as *The Green Berets* (1968) and *Flight of the Intruder* (1991).

(Vincze, 2013)

Il importe de mentionner que la deuxième occurrence du déterminant *some* dans la phrase en (33) n'est pas problématique, car son argument ne manque pas de précision. D'ailleurs, Vincze (2013) souligne l'ambiguïté de certaines formes d'expression de biais et la difficulté qu'elle pose aux méthodes de classification appuyées sur des listes de mots (Ganter & Strube, 2009; Recasens et al., 2013). Par exemple, *most* peut introduire un biais en étant une tergiversation (34), une esquive (35), un intensificateur (36) ou peut être utilisé de façon neutre (37).

(34) **Most** agree that…

(35) … **most** of the time.

(36) … the **most** touristic beach…

(37) He did the **most** he could.

(Vincze, 2013)

6.6 Voix active

Si en général l'emploi de la voix passive est à proscrire afin d'alléger le style et d'améliorer la compréhension (38), certains contextes l'exigent pour éliminer le biais que l'emploi de la voix

active peut apporter. Dans sa forme avant la rédaction, la phrase (39-a) assigne à un des participants le rôle d'agent (plutôt que celui de patient) par l'entremise de la voix active et d'un verbe support positivement connoté, soulignant ainsi sa contribution présumée.

(38) a. *(avant)* dariush a publié l album **appelé** donyaye en man roozaye qui **a été publié** le numtkn juin numtkn et récemment la nouvelle chanson « divar » le numtkn août numtkn
b. *(après)* dariush a publié l album donyaye en man roozaye le numtkn juin numtkn et la nouvelle chanson divar le numtkn août numtkn

(39) a. *(avant)* the united states department of justice indicted the company but **amway secured an acquittal**
b. *(après)* the united states department of justice indicted the company but **amway were acquitted**

L'implication dans la phrase à la voix active (39-a) peut être décrite comme ceci : la compagnie a joué un rôle actif dans l'obtention de son acquittement et c'est grâce à ses efforts qu'elle l'a mérité. La formulation au passif (39-b) n'implique pas les sens de 'agence' et de 'mérite'. Nous avançons qu'il s'agit ici d'implicature plutôt que de supposition ou de présupposition, car les sens rajoutés par l'utilisation de la voix active ne découlent pas du sens du prédicat et ne déterminent pas sa vérité.

6.7 Narration descriptive

Raconter en détail ou analyser plutôt que d'annoncer l'événement est une forme de partialité difficile à détecter, car elle ne relève pas du vocabulaire ni de structures particulières (40,41).

(40) *(avant)* he was a former club rugby and an opening batsman in club cricket but did not have the ability to make it all the way to the top level these two sports have become his particular area of expertise however he is very knowledgable on all sports that are played

(41) *(avant)* however the most important consequence of the battle was that president lincoln was able to sieze upon the victory claim it as a strategic victory for the north and release his emancipation proclamation

Suite à une analyse des erreurs de leur détecteur de contenu promotionnel dans Wikipédia, Bhosale *et al*. (2013) signalent que la majorité des pages non neutres omises par leur algorithme (les faux négatifs) contenaient cette forme implicite de partialité. Parmi les exemples qu'ils avaient relevés figuraient l'histoire détaillée d'un personnage fictif ainsi qu'un compte-rendu exhaustif du développement d'une entreprise. Nos observations confirment leurs constats en ce qui concerne les articles biaisés composés de phrases neutres. Durant l'évaluation des corpus, nous avons recensé plusieurs occurrences d'articles sur des personnalités composés de phrases à priori neutres, mais qui dans l'ensemble trahissaient un certain point de vue par la quantité importante de détails.

Dans cette forme particulière d'expression de biais, l'implicature contenant le point de vue de l'auteur émane non pas de l'interprétation de l'énoncé, mais de sa taille.

6.8 Omissions

Selon plusieurs auteurs, l'omission n'est pas une absence d'action, mais plutôt une forme négative d'action (volontaire ou involontaire) (Bach, 2010; Gheorghe, 2010; Payton, 2016). En tant qu'action, l'omission est alors dotée d'agence et capable d'entraîner des conséquences. C'est la raison pour laquelle omettre de l'information d'une phrase influence son interprétation au même titre qu'y insérer des faussetés. Que ce soit volontairement, par oubli ou par ignorance, passer un fait sous silence, quelle que soit son importance, introduit de la partialité par le fait de déformer l'interprétation des propos. Peindre un portrait incomplet du sujet traité nuit à la neutralité de sa représentation. Par exemple, l'omission peut servir à faire diminuer l'importance d'un sujet (42), à cacher un fait désagréable (43) ou bien à déguiser une opinion en fait en omettant des précisions importantes (44).

(42) a. *(avant)* as of numtkn it is the ethnic minority party in romania with representation in the romanian parliament
b. *(après)* as of numtkn it is the ethnic minority party in romania with representation in the romanian parliament **and is part of the governing coalition along with the justice and truth alliance and the conservatives**

(43) a. *(avant)* les arts martiaux qui existent à l époque comme le taekkyeon et le soo bak sont

très éloignés du taekwondo actuel mais incitent néanmoins déjà sur les techniques de jambes

b. *(après)* cependant les arts martiaux qui existent à l époque comme le taekkyeon et le soo bak sont très éloignés du taekwondo actuel même s ils incitent néanmoins déjà sur les techniques de jambes **et le lien entre le taekwondo et ces arts anciens est qualifié de mensonger considéré comme un argument de propagande par certains historiens de la discipline**

(44) a. *(avant)* in numtkn the journal won the praise of fascist leaders

b. *(après)* **there are some authors who retain** that the journal won the praise of fascist leaders

L'omission d'informations factuelles est sans doute le type d'expression de biais le plus difficile à détecter et à classifier, non seulement pour les machines, qui sont censées reconnaître le manque comme une caractéristique informative, mais aussi pour les humains, puisque combler des lacunes factuelles nécessite des connaissances spécifiques au domaine.

Chapitre 7 Expériences de classification

L'objectif des expériences était d'évaluer l'utilité des corpus dans une tâche de classification binaire de phrases. Les trois corpus que nous avons construits (cf. § 4.4) comprenaient deux classes de phrases en proportion égale:

- classe positive: des phrases «partiales» que nous avons définies comme la divergence entre deux révisions consécutives
- classe négative: des phrases «neutres» en provenance de la concordance entre deux révisions consécutives

Chaque corpus a été divisé en un ensemble d'entraînement (80%), un ensemble de développement (10%) sur lequel nous avons optimisé les paramètres, et un ensemble de test (10%) sur lequel nous avons effectué une seule évaluation de la meilleure configuration.

7.1 Classification avec fastText

Nous avons utilisé la librairie fastText (Joulin et al., 2017) qui met en œuvre un algorithme de régression logistique multinomiale en sus de plongements de mots appris sur le corpus fourni ou préentraînés. L'algorithme de fastText construit la représentation d'une phrase (d'un document) comme un sac de vecteurs de mots. Quant à la représentation de chaque mot du vocabulaire, elle constitue la somme des représentations vectorielles de ses n-grammes. Nous avons identifié et choisi les hyperparamètres optimaux (en gras) à partir des valeurs suivantes :

- époques : **5**, 10, 25;
- taux d'apprentissage : **0,1**, 0,01 ou 0,05;
- taille des n-grammes de mots : **1** à 5;
- taille minimale des n-grammes de caractères : 2 ou **3**;
- taille maximale des n-grammes de caractères : **6**;
- nombre minimal d'occurrences d'un mot : 1 à **5**;

- nombre de dimensions des plongements de mots : **100** ou 300;
- fonction de perte : **softmax**, ns[36], hs[37];
- utilisation des plongements de mots préentraînés : vrai ou **faux**;
- taux de la mise à jour du taux d'apprentissage : 50 ou **100**;

En appliquant les vecteurs préentraînés de fastText[38], nous avons obtenu des résultats comparables pour l'anglais et le français sans aucun gain significatif, ainsi que des performances inférieures pour le bulgare. Par conséquent, nous avons choisi le modèle final dont les performances globales (dans les trois langues) sont optimales et dont les représentations des mots sont apprises à partir de nos corpus. Nous l'avons ensuite évalué sur l'ensemble de test.

7.2 Classification avec sklearn

Nous avons également testé deux algorithmes de classification de la librairie sklearn[39]: la régression logistique (RL) (Hosmer et al., 2013) et la machine à vecteurs de support (SVM) (Fan et al. 2008) avec une vectorisation des données par sac de mots. Chacun des algorithmes a été exécuté en utilisant les mêmes hyperparamètres sur chacun des corpus afin d'obtenir les meilleures performances globales moyennes en termes de précision, rappel et mesure F_1. Une recherche exhaustive (*grid search*) nous a permis d'identifier les meilleurs hyperparamètres de l'ensemble de leurs valeurs possibles. Le seul prétraitement que nous avons appliqué aux corpus en amont de leur vectorisation a été d'enlever les mots vides (*stop words*) qui se caractérisent par une distribution uniforme à travers les documents du corpus.

L'optimisation du SVM a nécessité 72 permutations des hyperparamètres suivants :

- taille des n-grammes de mots : unigrammes, unigrammes et bigrammes, unigrammes à trigrammes;

[36] *Skipgram negative sampling*
[37] *Skipgram hierarchical softmax*
[38]Offerts en 157 langues, préentraînés sur Common Crawl et Wikipédia (Grave et al., 2018) https://fasttext.cc/docs/en/crawl-vectors.html
[39] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

- taille du sac de mots : 100, 150, 300, 500, 1,000 et 3,000;
- utilisation d'une pondération FID (fréquence inverse de document) : vrai ou faux;
- valeur de α[40] : 0.01, 0.001.

Les autres hyperparamètres ont été laissés à leur valeur par défaut[41].

Quant à RL, nous avons testé 504 permutations avec les hyperparamètres suivants :

- taille des n-grammes de mots et du sac de mot identique à la configuration de SVM;
- C[42] : 1.0e–3, 1.0e–2, 1.0e–1, 1.0e0, 1.0e+1, 1.0e+2 et 1.0e+3.
- Solver[43] : sag, saga.

Nous avons sélectionné le meilleur modèle en fonction de sa performance sur les trois ensembles de développement donnée par la moyenne des trois mesures F_1. La même configuration d'hyperparamètres optimale a été ensuite utilisée pour entraîner trois modèles (un par langue) dont nous avons testé la performance sur les ensembles de test respectifs.

[40] Un hyperparamètre qui contrôle la force de la régularisation.
[41] Dans la version 0.21.2 de la librairie sklearn.
[42] Un hyperparamètre qui contrôle la force de la régularisation.
[43] L'algorithme de résolution du problème d'optimisation.

7.3 Résultats et discussion

		SVM		fastText		RL	
Mesure	**Lng**	dev	test	dev	test	dev	test
Précision	BG	0,5387	**0,5886**	0,5324	0,5330	0,5182	0,5032
	FR	0,5059	0,5087	0,5533	**0,5520**	0,5151	0,5161
	EN	0,5112	0,5083	0,5656	**0,5634**	0,5230	0,5224
Rappel	BG	0,4318	0,5049	0,4752	0,4937	0,6219	**0,6303**
	FR	0,8877	**0,8363**	0,5724	0,5721	0,6751	0,6739
	EN	0,8357	**0,8277**	0,5686	0,5718	0,5344	0,5354
F_1	BG	0,4794	0,5435	0,5022	0,5126	0,5653	**0,5596**
	FR	0,6444	**0,6146**	0,5627	0,5619	0,5844	0,5845
	EN	0,6334	**0,6291**	0,5671	0,5676	0,5286	0,5288

Tableau 13. Résultats pour la classe de phrases partiales par langue, ensemble et méthode de classification

Le tableau 13 résume les résultats des expériences détaillées dans le chapitre précédent (cf. §7) pour les algorithmes SVM, fastText et régression logistique. Pour chaque mesure de performance, ensemble de données, algorithme et langue, nous fournissons des résultats concernant la classe de biais. La performance la plus élevée obtenue sur l'ensemble de test de chaque langue est en caractères gras.

Pour l'algorithme RL, nous avons obtenu les meilleures performances en utilisant une valeur C de 0,001, la méthode de descente de gradient avec un terme de régularisation L1, un modèle d'unigrammes de 100 dimensions sans repondération de la fréquence inverse de documents (fid). Les meilleurs paramètres pour le SVM utilisaient un modèle mettant ensemble

unigrammes, bigrammes et trigrammes avec un α de 0,001 et une repondération fid. Quant à fastText, l'ensemble de paramètres le plus performant utilisait les valeurs par défaut[44], unigrammes et un minimum de 5 occurrences par token.

Paramètres	lr	0,1	0,1	0,1	0,05	0,05
	minCount	5	1	5	1	5
	dim	100	300	300	300	300
Précision	EN	**0,565**	**0,565**	0,564	0,564	0,564
	FR	0,558	0,555	0,559	0,560	**0,561**
	BG	0,521	**0,522**	0,517	0,507	0,492
	Globale	**0,548**	0,547	0,547	0,544	0,539

Tableau 14. La précision de quelques configurations d'hyperparamètres de fastText sur les trois corpus

Les résultats d'une sélection d'expérimentations (Tableau 14) illustrent la relation entre certains hyperparamètres et la taille des corpus d'entraînement. Nous avons fait varier (entre autres) le taux d'apprentissage (lr), le nombre minimal d'occurrences d'un token (minCount) et le nombre de dimensions d'un vecteur (dim). Si la performance sur l'anglais semble stable à travers toutes les configurations, celle du français bénéficie d'un nombre accru de dimensions, d'un taux d'apprentissage plus petit et d'un vocabulaire plus restreint (minCount=5). Au contraire, un apprentissage (relativement) optimal pour le bulgare nécessite un vocabulaire sans restrictions et ne bénéficie guère d'un taux d'apprentissage inférieur à la valeur par défaut, étant donné la petite taille du corpus. Ces variations de performance, quoique faibles, suggèrent l'avantage d'adapter l'algorithme d'apprentissage à la taille du corpus et potentiellement à la langue. Par exemple, restreindre le vocabulaire d'une langue à

[44] Nous avons utilisé un *wrapper* en Python pour fastText (version 0.8.3) https://pypi.org/project/fasttext/0.8.3/ qui, lui, utilisait la version 0.1.0 de fastText https://github.com/facebookresearch/fastText/releases

morphologie flexionnelle offre généralement une augmentation de la performance par le biais d'une similarité distributionnelle améliorée (Avraham & Goldberg, 2017; Levy et al., 2015).

Dans l'ensemble, le fait d'avoir obtenu des résultats similaires sur les corpus de développement et de test pour chaque algorithme écarte la possibilité de surajustement. En outre, les trois mesures (précision, rappel et F_1) présentent une variance relativement faible entre les langues, à l'exception du rappel en bulgare avec SVM, considérablement inférieur aux deux autres langues.

Nous observons que les méthodes de vectorisation et de classification de fastText offrent une meilleure précision sur des corpus plus grands, mais SVM et RL assurent un rappel beaucoup plus élevé, quel que soit le nombre d'exemples.

Bien que relativement meilleure, la performance de SVM peut être largement améliorée. Selon nous, la présence de nombreuses phrases neutres parmi les phrases biaisées serait responsable de la faible performance observée (cf. §5.4). L'ambiguïté lexicale et contextuelle de certaines formes d'expression de biais (cf. §6) serait un autre facteur affectant les résultats. Toutefois, nous observons des performances comparables entre des corpus de taille variable et de langues appartenant à des familles différentes. Sur l'ensemble de test, nos meilleures mesures F_1 se situent entre 0,56 et 0,63. Ce résultat est inférieur à celui de Vincze (2013) (0,71) ou à celui de Hube et Fetahu (2018) (0,70), mais notre approche déploie un vaste corpus automatiquement dérivé de Wikipédia en trois langues, en utilisant un minimum de ressources linguistiques spécifiques, contrairement à l'approche monolingue de Vincze, ou bien celle de Hube et Fetahu qui dépend de ressources lexicales monolingues.

Il est également possible de nous comparer aux résultats de la compétition de détection d'esquives dans un corpus de 11 000 phrases en anglais tirées de Wikipédia, dont 2 346 contenaient des marqueurs d'incertitude (cf. §6.6) (Farkas et al., 2010). Des dix-sept

soumissions à la tâche 1W, le meilleur système a obtenu un F_1 de 60,2 grâce à un modèle de sac de mots sans caractéristiques syntaxiques supplémentaires.

Nos résultats établissent une référence pour la détection multilingue de biais en trois langues. Pour améliorer la performance par rapport à une seule langue, on peut envisager une optimisation des hyperparamètres en fonction d'un modèle (plutôt que de trois), l'introduction de caractéristiques supplémentaires ou un nettoyage approfondi des deux classes du corpus.

7.4 Analyse des erreurs

Nous avons procédé à une analyse des résultats de la classification avec fastText de l'ensemble test du corpus bulgare avec l'objectif de vérifier l'étiquetage des faux positifs et des faux négatifs. Sachant qu'une portion des phrases avaient été mal classées lors de la génération du corpus (environ la moitié, selon l'évaluation manuelle des corpus; cf. §5), nous voulions estimer, à l'instar de Bhosale *et al*. (2013), la proportion de phrases biaisées mal étiquetées, mais correctement détectées par l'algorithme de classification. La taille restreinte de ce corpus (952 phrases) nous a permis de lire chacun des 196 faux positifs et 253 faux négatifs afin d'approximer sa classe (neutre ou biaisé).

Nous estimons que 49% des faux positifs (cf. Tableau 15) sont bien des phrases neutres que l'algorithme a reconnues à tort comme biaisées. Cependant, environ la moitié contiennent un élément qui semble biaisé et qui aurait pu influencer leur classification. Par exemple, 18 phrases neutres traitent d'un sujet controversé (religion, politique, histoire, sport ou biographie) qui a déjà alimenté le corpus de nombreux exemples biaisés. Dans neuf cas, on retrouve une terminologie neutre, mais peu fréquente, tandis que cinq phrases contiennent du vocabulaire partial, mais dans un contexte de citation ou de titre d'ouvrage. Durant l'évaluation du corpus, nous avons constaté la présence de plusieurs phrases d'en-tête dans la classe des phrases biaisées. L'en-tête est le début de l'article commençant par une définition du sujet, suivie par un résumé succinct de l'article. Comme la structure de ces phrases est très rigide, les quatre faux positifs de ce type sont probablement dus à une surgénéralisation. Quant aux

énumérations, elles sont généralement proscrites dans le cadre des narrations descriptives et le corpus en contient des exemples étiquetés comme biaisés. Enfin, certaines phrases semblent neutres, mais il est impossible de le confirmer en isolation. Il peut s'agir soit d'omissions dont l'identification exige des connaissances du domaine, soit d'ambiguïté contextuelle.

Une autre portion de 16% des faux positifs est composée de bruit sous forme de phrases incomplètes, d'actes de vandalisme ou de propos mal écrits, sans respect du style encyclopédique. La dernière tranche de 35% consiste en des phrases, selon nous, biaisées, qui ont été mal étiquetées lors de la création du corpus. Il s'agit alors de véritables instances de partialité, correctement identifiées par le classifieur, mais cachés parmi les faux positifs.

TYPE	**PHRASES**	**PROPORTION**
neutre	49	25%
neutre (controversé)	18	9%
neutre (terminologie)	9	5%
neutre (entête)	4	2%
neutre (vocabulaire biaisé)	5	3%
neutre (énumération)	5	3%
neutre (incertain)	4	2%
biaisé	68	35%
incomplet	26	13%
vandalisme	1	1%
style problématique	5	2%
TOTAL	**194**	**100%**

Tableau 15. Répartition des faux positifs par type

L’analyse des faux négatifs (cf. Tableau 16) démontre un véritable échec de détection de biais dans seulement 40% des cas. Une autre tranche de 40% des phrases étiquetées comme partiales s’avèrent en réalité neutres et ont été donc correctement identifiées par le classifieur comme telles. Les 20% restant sont du bruit.

TYPE	**PHRASES**	**PROPORTION**
neutre	101	40%
biaisé	100	40%
incomplet	49	19%
style problématique	2	1%
TOTAL	**252**	**100%**

Tableau 16. Répartition des faux négatifs par type

Les vrais faux négatifs, c.-à-d. les phrases correctement identifiées comme partiales, mais incorrectement classifiées comme neutres, sont souvent écrits dans un style narratif ou trop détaillé comportant de longues énumérations. Certaines de ces phrases ne paraissent partiales qu’en contexte, car quoique neutres en isolation, elles font partie de relations détaillés de réalisations (dans les pages biographiques), d’avantages ou de caractéristiques (dans les pages de produits), de pratiques religieuses (dans les pages de cultes ou de croyances), etc.

Chapitre 8 Conclusion

Nous avons présenté[45] une méthode semi-automatique pour l'extraction de phrases biaisées de Wikipédia en bulgare, en français et en anglais (cf. §4). Comme cette méthode ne repose pas sur des ressources linguistiques spécifiques, à l'exception d'une liste de balises NPV et d'une liste de mots vides, elle est facilement applicable aux archives de Wikipédia dans d'autres langues. Elle exploite les balises ajoutées par les rédacteurs humains dans les articles qu'ils considèrent comme biaisés. Nous récupérons la dernière révision balisée, que nous présumons biaisée, et celle d'après (non balisée), que nous présumons neutre, afin de les comparer et d'en extraire les phrases supprimées et ajoutées.

Nous avons annoté manuellement 1 784 des phrases supprimées, pour les trois langues combinées, et nous avons constaté que seule la moitié de cet échantillon était réellement biaisée. Une valeur moyenne de κ de 0,41 (0,46 si l'on ignore une valeur aberrante) indique que la tâche n'est pas triviale, même pour les humains (cf. §5).

Nous avons apporté une contribution modeste à la classification des formes linguistiques d'expression de biais, par l'ajout de nouveaux exemples (tirés de nos corpus) aux classes déjà existantes et par la description de quelques nouvelles formes (cf. §6). De plus, nous avons organisé les expressions de partialité dans un continuum selon leur caractère plus ou moins explicite. À l'un des extêmes de ce continuum on trouve les formes de cadrage qui renforcent la saillance de certaines idées par l'utilisation d'intensificateurs subjectifs, de vocabulaire partisan ou spécialisé. Les présuppositions et les implications se situent au milieu, car ces formes linguistiques explicites (verbes de supposition et verbes d'implication; verbes factifs; verbes assertifs; esquives et tergiversations) déclenchent des sens supplémentaires implicites. À l'autre extrême se situent les formes d'implicature capables d'introduire un point de vue par des moyens si subtils que la voix active, la narration descriptive ou l'omission.

[45] Une version concise de ce travail (Aleksandrova et al., 2019) a d'abord été présentée à RANLP 2019.

En utilisant nos corpus, nous avons testé trois algorithmes de classification : la vectorisation par sac de mots avec SVM, fastText et la régression logistique (cf. §7). La discussion des résultats (§7.3) est accompagnée par une analyse des erreurs de classification de l'algorithme fastText sur le bulgare (cf. §7.4).

L'algorithme de génération de corpus de phrases partiales pourrait servir de base à de futurs travaux visant à minimiser le bruit et par conséquent, à améliorer la qualité des ensembles de données automatiquement étiquetés. Les aspects qui bénéficieraient de développement supplémentaire sont : la segmentation en phrases (pour minimiser le nombre de phrases incomplètes); l'extraction et le formatage des listes non numérotées; les modifications d'orthographe mineures qui accompagnent des fois la réécriture d'articles biaisés. Une autre optimisation envisageable consiste à segmenter les corpus en deux ou plusieurs sous-corpus selon les principales formes d'expression de biais (par exemple, explicite contre implicite) et d'investiguer les réssultats de la classification en fonction de ce classement. Il serait alors possible d'explorer et d'évaluer séparément les différentes formes de biais, ce qui pourrait à son tour motiver des techniques de classification différentes. Enfin, tirer les exemples de phrases neutres des articles vedettes de Wikipédia (à l'instar de Bhosale *et al*. (2013)) pourrait contribuer à réduire la ressemblance des classes en renforçant le contraste entre le ton encyclopédique neutre et les expressions de biais.

Références bibliographiques

Aleksandrova, D., Lareau, F., & Ménard, P. A. (2019). Multilingual sentence-level bias detection in Wikipedia. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 42‑51.

Al Khatib, K., Schütze, H., & Kantner, C. (2012). Automatic Detection of Point of View Differences in Wikipedia. *Proceedings of COLING 2012*, 33‑50.

Al-Rajebah, N. I., & Al-Khalifa, H. S. (2010). Semantic relationship extraction and ontology building using wikipedia: A comprehensive survey. *International Journal of Computer Applications in Technology*, *12*(3), 6‑12.

Athanasiadou, A. (2007). On the subjectivity of intensifiers. *Language sciences* , *29*(4), 554‑565.

Avraham, O., & Goldberg, Y. (2017). The Interplay of Semantics and Morphology in Word Embeddings. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 422‑426.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Lrec*, *10*, 2200‑2204.

Bach, K. (1994). Conversational impliciture. *Mind and language*, *9*(2), 124‑162.

Bach, K. (1999). The Myth of Conventional Implicature. *Linguistics and philosophy*, *22*(4), 327‑366.

Bach, K. (2006). The top 10 misconceptions about implicature. *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, *21*, 30.

Bach, K. (2008). Applying Pragmatics to Epistemology. *Philosophical Issues. A Supplement to Nous*, *18*, 68‑88.

Bach, K. (2010). Refraining, omitting, and negative acts. *A Companion to the Philosophy of Action*, 50‑57.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54‑61.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting Factuality of Reporting and Bias of News Media Sources. Dans *arXiv [cs.IR]*. arXiv. http://arxiv.org/abs/1810.01765

Bekavac, B., & Tadic, M. (2008). A Generic Method for Multi Word Extraction from Wikipedia. *ITI 2008 - 30th International Conference on Information Technology Interfaces*, 663‑668.

Bellows, M. (2018). *EXPLORATION OF CLASSIFYING SENTENCE BIAS IN NEWS ARTICLES WITH MACHINE LEARNING MODELS* [University of Rhode Island]. https://digitalcommons.uri.edu/theses/1309/

Bhosale, S., Vinicombe, H., & Mooney, R. (2013). Detecting promotional content in Wikipedia. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1851‑1857.

Blessing, A., & Schütze, H. (2010). Fine-Grained Geographical Relation Extraction from Wikipedia. *LREC*. https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/519_Paper.pdf

Bolinger, D. (2013). *Degree Words*. Walter de Gruyter.

Bonart, M., Samokhina, A., Heisenberg, G., & Schaer, P. (2019). An Investigation of Biases in Web Search Engine Query Suggestions. Dans *arXiv [cs.IR]*. arXiv. http://arxiv.org/abs/1912.00651

Cahill, A., Madnani, N., Tetreault, J., & Napolitano, D. (2013). Robust systems for preposition error correction using wikipedia revisions. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 507‑517.

Carston, Robyn. 2009. “The Explicit/Implicit Distinction in Pragmatics and the Limits of Explicit Communication.” International Review of Pragmatics 1 (1): 35–62.

Conrad, A., Wiebe, J., & Hwa, R. (2012). Recognizing arguing subjectivity and argument tags. *Proceedings of the workshop on extra-propositional aspects of meaning in computational linguistics*, 80‑88.

Dutrey, C., Bouamor, H., Bernhard, D., & Max, A. (2011). *Typologie des modifications dans les révisions de Wikipédia* (Vol. 1, p. 139). wicopaco.limsi.fr.

https://wicopaco.limsi.fr/pub/typologie-modifications-wikipedia.pdf

Entman, R. M. (2007). Framing Bias: Media in the Distribution of Power. *The Journal of communication*, *57*(1), 163‑173.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. “LIBLINEAR: A Library for Large Linear Classification.” Journal of Machine Learning Research: JMLR 9 (Aug): 1871–74.

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010). The CoNLL-2010 Shared Task: Learning to Detect Hedges and Their Scope in Natural Language Text. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning --- Shared Task*, 1‑12.

Frege, G. (2003). On sense and reference. *oversatt av Max Black, i J. Guitérrez-Rexach (red.): Semantics: Crictical concepts in linguistics*, *1*, 7‑25.

Ganter, V., & Strube, M. (2009). Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 173‑176.

Gheorghe, S. (2010). *Negative Acts*. https://philarchive.org/rec/GHENA

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215‑230.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. Dans *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1802.06893

Greene, S., & Resnik, P. (2009). More Than Words: Syntactic Packaging and Implicit Sentiment. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 503‑511.

Grice, H. P. (1975). Logic and Conversation. Dans *Speech Acts* (p. 41‑58). Brill.

Herzig, L., Nunes, A., & Snir, B. (2011). An Annotation Scheme for Automated Bias Detection in

Wikipedia. *Proceedings of the 5th Linguistic Annotation Workshop*, 47‑55.

Hirning, N. P., Chen, A., & Shankar, S. (2017). *Detecting and Identifying Bias-Heavy Sentences in News Articles*. Stanford University.

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Hooper, J. B. (1975). On Assertive Predicates. Dans *Syntax and Semantics volume 4* (p. 91‑124). Brill.

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

Hube, C., & Fetahu, B. (2018). Detecting Biased Statements in Wikipedia. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 1779‑1786.

Hutto, C. J., Folds, D., & Appling, D. (2015). Computationally detecting and quantifying the degree of bias in sentence-level text of news stories. *Proceedings of Second International Conference on Human and Social Analytics*. International Conference on Human and Social Analytics.

Hyland, K. (1998). *Hedging in Scientific Research Articles*. John Benjamins Publishing.

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014). Political ideology detection using recursive neural networks. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1113‑1122.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427‑431.

Kiparsky, P., & Kiparsky, C. (1968). *Fact*. Linguistics Club, Indiana University.

Kuang, S., & Davison, B. D. (2016). Semantic and context-aware linguistic model for bias detection. *Proc. of the Natural Language Processing meets Journalism IJCAI-16 Workshop*, 57‑62.

Lakoff, G. (1975). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. Dans D. Hockney,

W. Harper, & B. Freed (éds.), *Contemporary Research in Philosophical Logic and Linguistic Semantics: Proceedings of a Conference Held at the University of Western Ontario, London, Canada* (p. 221‑271). Springer Netherlands.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for*. https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00134

Lin, C., He, Y., & Everson, R. (2011). Sentence subjectivity detection with weakly-supervised learning. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 1153‑1161.

Lin, W.-H., Wilson, T., Wiebe, J., & Hauptmann, A. G. (2006). Which side are you on? Identifying perspectives at the document and sentence levels. *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, 109‑116.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the Web. *Proceedings of the 14th international conference on World Wide Web*, 342‑351.

Markkanen, R., & Schröder, H. (1989). Hedging as a translation problem in scientific texts. *Special languages: From human thinking to thinking machines*, 171‑179.

Massa, P., & Scrinzi, F. (2012). Manypedia: comparing language points of view of Wikipedia communities. *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 1‑9.

Max, A., & Wisniewski, G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. *LREC*. ftp://192.44.78.170/public/827_Paper.pdf

Mel'čuk, I. A. (1982). Lexical functions in lexicographic description. *Annual Meeting of the Berkeley Linguistics Society*, *8*, 427‑444.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Dans *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1301.3781

Morsey, M. (2012). DBpedia and the live extraction of structured data from Wikipedia (C. Stadler & S. Hellmann, trad.). *Rossiiskaya Akademiya Nauk. Programmirovanie*, *46*(2), 157‑181.

Murray, G., & Carenini, G. (2009). Detecting subjectivity in multiparty speech. *Tenth Annual Conference of the International Speech Communication Association*. https://www.isca-speech.org/archive/interspeech_2009/i09_2007.html

Nakayama, K., Hara, T., & Nishio, S. (2008). Wikipedia Link Structure and Text Mining for Semantic Relation Extraction. *SemSearch*, 59‑73.

Neale, S. (1999). Coloring and composition. *Philosophy and linguistics*, 35‑82.

Nelken, R., & Yamangil, E. (2008). Mining Wikipedia's article revision history for training computational linguistics algorithms. *Proceedings of the AAAI Workshop on Wikipedia*. https://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-006.pdf

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Paradis, C. (2001). Adjectives and boundedness. *Cognitive Linguistics*, *12*(1), 47‑65.

Park, S., Lee, K.-S., & Song, J. (2011). Contrasting opposing views of news articles on contentious issues. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 340‑349.

Pavalanathan, U., Han, X., & Eisenstein, J. (2018). Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm. *Proc. ACM Hum. -Comput. Interact.*, *2*(CSCW), 137:1‑137:23.

Payton, J. D. (2016). The logical form of negative action sentences. *Canadian journal of philosophy*, *46*(6), 855‑876.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Perez, C. C. (2019). *Invisible Women: Data Bias in a World Designed for Men*. Abrams.

Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S., & Weikum, G. (2018). On Measuring Bias in Online Information. *SIGMOD Rec.*, *46*(4), 16‑21.

Potts, C. (2007). Conventional implicatures, a distinguished class of meanings. *The Oxford handbook of linguistic interfaces*, *475501*. https://web.stanford.edu/~cgpotts/papers/potts-interfaces.pdf

Potts, C. (2015). Presupposition and implicature. *The handbook of contemporary semantic theory*. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118882139#page=180

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2019). Automatically Neutralizing Subjective Bias in Text. Dans *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1911.09709

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 1650‑1659.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 105‑112.

Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of personality and social psychology*, *35*(7), 485.

Shadowen, N. (2019). Ethics and Bias in Machine Learning: A Technical Study of What Makes Us « Good ». Dans N. Lee (éd.), *The Transhumanism Handbook* (p. 247‑261). Springer International Publishing.

Somasundaran, S., & Wiebe, J. (2010). Recognizing stances in ideological on-line debates. *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 116‑124.

Sumi, R., Yasseri, T., Rung, A., Kornai, A., & Kertesz, J. (2011). Edit Wars in Wikipedia. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 724‑727.

Tsurel, D., Pelleg, D., Guy, I., & Shahaf, D. (2016). Fun Facts: Automatic Trivia Fact Extraction from

Wikipedia. Dans *arXiv [cs.SI]*. arXiv. http://arxiv.org/abs/1612.03896

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124‑1131.

Vanderveken, D. (1990). *Meaning and Speech Acts: Volume 1, Principles of Language Use*. Cambridge University Press.

Vincze, V. (2013). Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 383‑391.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. *Ninth international AAAI conference on web and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10585

Wang, G., Yu, Y., & Zhu, H. (2007). PORE: Positive-Only Relation Extraction from Wikipedia Text. *The Semantic Web*, 580‑594.

Wanner, L. (éd.). (1996). *Lexical functions in lexicography and natural language processing* (Vol. 31). John Benjamins.

Wiebe, J., & Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *Computational Linguistics and Intelligent Text Processing*, 486‑497.

Wilson, T., & Raaijmakers, S. (2008). Comparing word, character, and phoneme n-grams for subjective utterance recognition. *Ninth Annual Conference of the International Speech Communication Association*. https://www.isca-speech.org/archive/archive_papers/interspeech_2008/i08_1614.pdf

Wisniewski, G., Max, A., & Yvon, F. (2010). Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipédia. *Actes de TALN*.

Yamangil, E., & Nelken, R. (2008). Mining Wikipedia revision histories for improving sentence compression. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 137‑140.

Yang, D., Halfaker, A., Kraut, R., & Hovy, E. (2017). Identifying semantic edit intentions from revisions in wikipedia. *Proceedings of the 2017*. https://www.aclweb.org/anthology/papers/D/D17/D17-1213/

Yano, T., Resnik, P., & Smith, N. A. (2010). Shedding (a Thousand Points of) Light on Biased Language. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 152‑158.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS One*, *7*(6), e38869.

Yiwei Zhou, A. I. C. (2016). Towards detection of influential sentences affecting reputation in wikipedia. *Proceedings of the 8th ACM Conference on Web Science*, 244‑248.

Yu, K., & Tsujii, J. (2009). Bilingual dictionary extraction from wikipedia. *Proceedings of machine translation summit xii*, 379‑386.

Zhou, Y., Cristea, A., & Roberts, Z. (2015). Is wikipedia really neutral? A sentiment perspective study of war-related wikipedia articles since 1945. *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, 160‑168.

Zhou, Y., Demidova, E., & Cristea, A. I. (2016). Who Likes Me More?: Analysing Entity-centric Language-specific Bias in Multilingual Wikipedia. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, 750‑757.

Zuck, J. G., & Zuck, L. V. (1986). Hedging in newswriting. *Beads or bracelets*, 172‑180.

Annexes

Annexe 1. Balises de neutralité contestée

ANGLAIS	FRANÇAIS	BULGARE
Cleanup-resume	Article mal proportionné	NPOV
Close connection	Article non neutre	POV
COI	Contenu évasif	есе
Colloquial	Controverses	КНИ
Editorial	Curriculum vitae	НГТ
Editorializing	Dithyrambe	неутралност
Essay-like	Désaccord de neutralité	Паун
FixPOV	Hagiographique	Паун-раздел
GenderPOV	name dropping	паун-раздел
Like resume	NdPV	прессъобщение
Like-resume	non neutre	реклама
Likeresume	NPOV	Спорна неутралност
LikeResume	passage lyrique	Спорна неутралност-раздел
Lopsided	Passage non neutre	шаблон:pov
Lopsided	passage promotionnel	шаблон:нгт
Neutral	PdV	шаблон:паун
Neutrality	PdVN	шаблон:прессъобщение
Npov	POV	шаблон:спорна неутралност
NPOV	POV fork	шаблон:спорна неутралност-раздел
NPOV language	POV pushing	
Peacock	promotionnel	
Peacock term	Section non neutre	
Point Of View	Ton journalistique	
Pov	Tribune	
PoV		
POV Article		
POV check		
POV check inline		
POV lead		
POV map		
POV section		
POV statement		
POV title		
POV-statement		
Puffery		
resume		

Systemic bias
Systemic bias
Undisclosed paid
Undue weight
Weasel
Weasel inline
Weasel words
Weasel-words
Weaselwords

Annexe 2. Paire de révisions d'un article

Nous avons ajouté des sauts de ligne et un alignement afin de faciliter la lisibilité et la comparaison des versions.

	AVANT	APRÈS
1	{{Désaccord de neutralité|Politique et société}}	
2	{{Désaccord de neutralité|Politique et société}}	{{article incomplet}}
3	{{À sourcer}}	{{À sourcer}}
4	'''L'ethnocentrisme''' est un concept [[ethnologie|ethnologique]] qui est apparu, en même temps que le mot, au milieu du {{XXe siècle}}. Il signifie la {{citation|tendance, '''plus ou moins consciente''', à privilégier les valeurs et les formes culturelles du groupe ethnique auquel on appartient}}<ref>[http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos//''Dictionnaire de l'Académie française, 8{{e}} édition, 1932-1935.'']</ref>. Une autre définition restreint l'ethnocentrisme à un {{citation|Comportement social et [une] attitude	'''L'ethnocentrisme''' est un concept [[ethnologie|ethnologique]] qui est apparu, en même temps que le mot, au milieu du {{XXe siècle}}. Il signifie la {{citation|tendance, plus ou moins consciente, à privilégier les valeurs et les formes culturelles du groupe ethnique auquel on appartient}}<ref>[http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos//''Dictionnaire de l'Académie française, 8{{e}} édition, 1932-1935.'']</ref>. Une autre définition restreint l'ethnocentrisme à un {{citation|Comportement social et [une] attitude inconsciemment motivée}}<ref

	'''inconsciemment''' motivée}}<ref name="atilf">[http://atilf.atilf.fr/tlf.htm ''Le trésor de la langue française informatisé.'']</ref> qui amènent en particulier à {{citation|'''surestimer''' le groupe racial, géographique ou national auquel on appartient, aboutissant parfois à des '''préjugés''' en ce qui concerne les autres peuples}}<ref name="atilf"/>.	name="atilf">[http://atilf.atilf.fr/tlf.htm ''Le trésor de la langue française informatisé.'']</ref> qui amènent en particulier à {{citation|surestimer le groupe racial, géographique ou national auquel on appartient, aboutissant parfois à des préjugés en ce qui concerne les autres peuples}}<ref name="atilf"/>.
5	==Dans le domaine des sciences humaines ==	==Dans le domaine des sciences humaines ==
6	=== L'ethnocentrisme : trait universel de l'humanité ?<ref name="clv">L'ethnocentrisme, chap. 3 in ''Race et histoire'' par C. Lévi-Strauss, Paris, UNESCO, 1952</ref> ===	=== L'ethnocentrisme : trait universel de l'humanité ? ===
7	L'anthropologie a constaté à maintes reprises dans les sociétés et civilisations ''premières'' que la notion d'humanité est presque toujours restreinte au groupe d'êtres humains auquel l'individu appartient. Le plus souvent, le mot qui définit le concept d'être humain (ou ''hommes'') dans la langue du groupe considéré ne concerne que les membres dudit groupe <ref>{{citation|un grand nombre de populations dites primitives se désignent d'un nom qui signifie les ''hommes'' (…) impliquant ainsi que les autres (…) ne participent pas des vertus – ou même de la nature – humaines (...).}} (in ''Race et histoire'', réf. ci-dessus)</ref>. [[Claude Lévi-Strauss|C. Lévi-Strauss]] estime même que {{citation|la notion d'humanité, englobant, sans distinction de race ou de civilisation, toutes les formes de l'espèce humaine, est d'apparition fort tardive}}<ref name="clv"/> ; d'une part, et que le rejet hors de l'humanité de tous ceux trop différents pour en faire	L'anthropologie a constaté à maintes reprises dans les sociétés et civilisations ''premières'' que la notion d'humanité est presque toujours restreinte au groupe d'êtres humains auquel l'individu appartient. Le plus souvent, le mot qui définit le concept d'être humain (ou ''hommes'') dans la langue du groupe considéré ne concerne que les membres dudit groupe <ref>{{citation|un grand nombre de populations dites primitives se désignent d'un nom qui signifie les ''hommes'' (…) impliquant ainsi que les autres (…) ne participent pas des vertus – ou même de la nature – humaines (...).}}, ''Race et histoire''.</ref>. [[Claude Lévi-Strauss]] estime même que {{citation|la notion d'humanité, englobant, sans distinction de race ou de civilisation, toutes les formes de l'espèce humaine, est d'apparition fort tardive}}<ref>Claude Lévi-Strauss, ''Race et histoire''.</ref> ; d'une part, et que le rejet hors de l'humanité de tous ceux trop

	partie<ref>{{citation|l'attitude la plus ancienne, et qui repose sans doute sur des fondements psychologiques solides puisqu'elle tend à réapparaître chez chacun de nous quand nous sommes placés dans une situation inattendue, consiste à répudier purement et simplement les formes culturelles (...) qui sont les plus éloignées de celles auxquelles nous nous identifions.}} (in ''Race et histoire'', réf. ci-dessus)</ref>est, paradoxalement, un trait de comportement universel<ref>{{citation|Cette attitude de pensée, au nom de laquelle on rejette les ''sauvages'' (ou tous ceux qu'on choisit de considérer comme tels) hors de l'humanité, est justement l'attitude la plus marquante et la plus distinctive de ces sauvages mêmes. (...) En refusant l'humanité à ceux qui apparaissent comme les plus "sauvages" ou "barbares" de ses représentants, on ne fait que leur emprunter une de leurs attitudes typiques.}} (in ''Race et histoire'', réf. ci-dessus)</ref>, d'autre part.	différents pour en faire partie<ref>{{citation|l'attitude la plus ancienne, et qui repose sans doute sur des fondements psychologiques solides puisqu'elle tend à réapparaître chez chacun de nous quand nous sommes placés dans une situation inattendue, consiste à répudier purement et simplement les formes culturelles (...) qui sont les plus éloignées de celles auxquelles nous nous identifions.}}, ''Race et histoire''.</ref>est, paradoxalement, un trait de comportement universel<ref>{{citation|Cette attitude de pensée, au nom de laquelle on rejette les ''sauvages'' (ou tous ceux qu'on choisit de considérer comme tels) hors de l'humanité, est justement l'attitude la plus marquante et la plus distinctive de ces sauvages mêmes. (...) En refusant l'humanité à ceux qui apparaissent comme les plus "sauvages" ou "barbares" de ses représentants, on ne fait que leur emprunter une de leurs attitudes typiques.}} , ''Race et histoire''.</ref>, d'autre part.
8	=== L'ethnocentrisme : le propre de l'ethnologue ?===	=== L'ethnocentrisme : le propre de l'ethnologue ?===
9	En ce qui concerne les sciences humaines, en général, et l'anthropologie, en particulier, un auteur comme [[Clifford Geertz|C. Geertz]] considère que, n'étant justement pas des sciences expérimentales à la recherche de lois, mais des sciences interprétatives à la recherche de sens, toute description implique un ethnocentrisme relatif mais inévitable. Pour Geertz, l'observateur (l'ethnographe) ne peut qu'essayer {{citation|de lire par-dessus l'épaule}}{{Citation nécessaire}}de la population	En ce qui concerne les sciences humaines, en général, et l'anthropologie, en particulier, un auteur comme [[Clifford Geertz|C. Geertz]] considère que, n'étant justement pas des sciences expérimentales à la recherche de lois, mais des sciences interprétatives à la recherche de sens, toute description implique un ethnocentrisme relatif mais inévitable. Pour Geertz, l'observateur (l'ethnographe) ne peut qu'essayer {{citation|de lire par-dessus l'épaule}}{{Citation nécessaire}}de la population

	étudiée. Les linguistes{{Citation nécessaire}}, quant à eux, ont pu démontrer que la langue même, en ce qu'elle est un construit culturel, participe à cette tendance<ref>{{citation|[pour] un grand nombre de populations dites primitives (…) les autres(…) sont tout au plus composées de ''mauvais'', de ''méchants'', de ''singes de terre'' ou d'''oeufs de pou''. On va jusqu'à priver l'étranger de ce dernier degré de réalité en en faisant un ''fantôme'' ou une ''apparition''.}} (in ''Race et histoire'', réf. ci-dessus)</ref>.	étudiée. Les linguistes{{Citation nécessaire}}, quant à eux, ont pu démontrer que la langue même, en ce qu'elle est un construit culturel, participe à cette tendance<ref>{{citation|[pour] un grand nombre de populations dites primitives (…) les autres(…) sont tout au plus composées de ''mauvais'', de ''méchants'', de ''singes de terre'' ou d'''oeufs de pou''. On va jusqu'à priver l'étranger de ce dernier degré de réalité en en faisant un ''fantôme'' ou une ''apparition''.}}, ''Race et histoire''.</ref>.
10	==Débats, controverses et polémiques==	
11	===Glissements de l'ethnocentrisme===	
12	{{passage non neutre|L'ethnocentrisme peut se réaliser dans le [[communautarisme identitaire]], le « [[chauvinisme]] », le [[nationalisme]], le [[colonialisme]], l'[[esclavage]] la [[xénophobie]] et le [[racisme]].}}	
13	{{passage non neutre|Par exemple, dans le cas de l'ethnocentrisme français qui dévalorise et méprise, la [[diversité culturelle]] est promue sur le plan international (revendication de son [[exception culturelle]]) alors qu'elle menace gravement sur le plan intérieur après le nivellement linguistique imposé par la IIIe république }}	
14	(cf [[politique linguistique de la France]]) : la plupart des langues régionales de France sont classées comme étant en danger par l'[[UNESCO]]. Ainsi la France exige t-elle des pays entrant dans l'[[Union européenne]] qu'ils	

	signent la [[charte des langues régionales]] alors qu'elle refuse elle-même de la ratifier.	
15	{{passage non neutre|On considère ainsi qu'elle est à l'origine de la ségrégation raciale, de conflits « inter-ethniques » et de guerres.}}	
16	{{passage non neutre|L'[[anthropologie]], dont la prétention est de tenir un discours objectif sur les groupes humains, est elle-même accusée d'avoir soutenu des thèses ethnocentriques]]. Même si cette attitude est particulièrement critiquée chez les fondateurs de la discipline, qui s'inscrivaient majoritairement dans le courant de pensée [[Évolutionnisme (anthropologie)|évolutionniste]], {{passage non neutre|la vigilance est toujours de vigueur dans les définitions de concept pouvant être soupçonnées d'[[eurocentrisme|européo-centrisme]].}}	
17	Ainsi [[Pierre Clastres]], dans ''La société contre l'État'', critique l'idée selon laquelle il existerait des peuples sans pouvoir, sous prétexte qu'on n'y observerait aucune coercition, alors qu'il faudrait plutôt remettre en question notre notion de [[Pouvoir (sociologie)|pouvoir]] comme héritière d'un certain contexte culturel. {{passage non neutre}} {{passage non neutre|La critique de l'ethnocentrisme devient ainsi une attitude essentielle dans l'exercice de l'anthropologie.}} Ce concept a aussi fait l'objet d'études par [[Claude Lévi-Strauss]] dans ''[[Race et Histoire]]''. Celui-ci est toujours d'actualité. À l'inverse on parle d'interculturalisme: effort pour se décentrer, se	

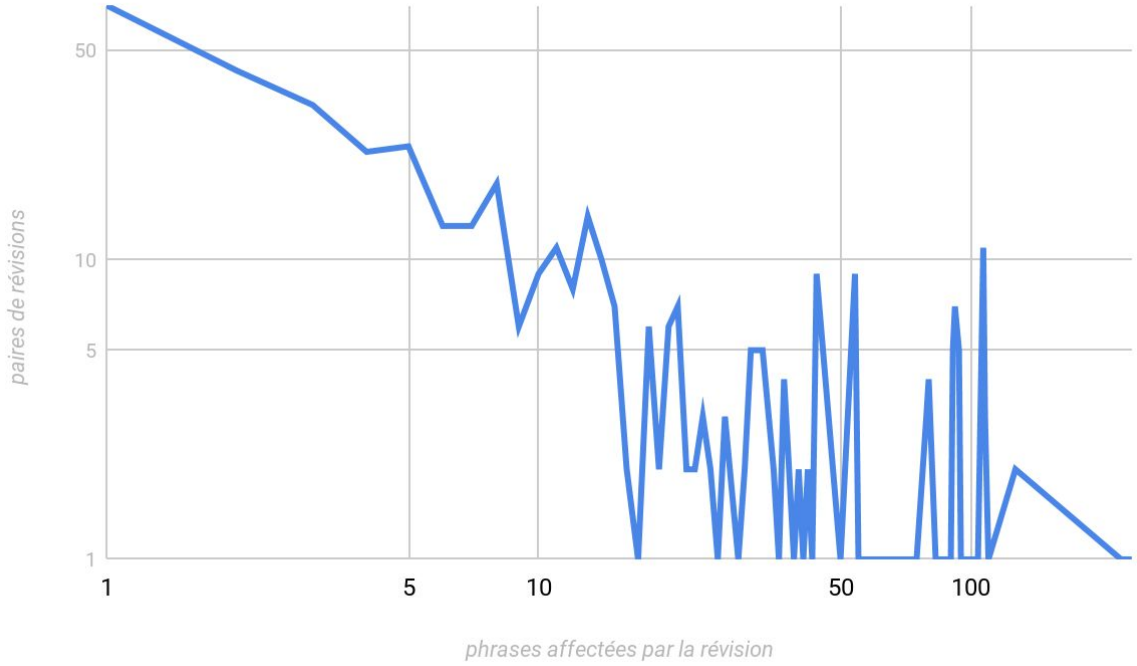
	mettre à la place de l'autre, coopérer, comprendre comment l'autre nous perçoit. L'ethnocentrisme signifie qu'on considère sa propre ethnie comme le centre du monde. :Exemples de formes prises : #France : [[francocentrisme]] #Europe : [[eurocentrisme]] #Afrique : [[afrocentrisme]]	
18	== Voir aussi ==	== Voir aussi ==
19	==Notes et références==	
20	<references/>	
21	===Liens internes===	===Articles connexes===
22	Théories pouvant se référer à l'ethnocentrisme : [[Évolutionnisme (anthropologie)]] - [[Nationalisme]] - [[Régionalisme (politique)|Régionalisme]] - [[Racisme]] - [[Racialisme]] - [[Communautarisme identitaire]] - [[Colonialisme]] - [[Patriotisme]] - [[Religions]] - [[ethnodifférencialisme]] - [[polycentrisme]] - [[Médiation interculturelle]] - [[Habitus (sociologie)]]	* [[Francocentrisme]] * [[Eurocentrisme]] * [[Afrocentrisme]] Théories pouvant se référer à l'ethnocentrisme : [[Évolutionnisme (anthropologie)]] - [[Nationalisme]] - [[Régionalisme (politique)|Régionalisme]] - [[Racisme]] - [[Racialisme]] - [[Communautarisme identitaire]] - [[Colonialisme]] - [[Patriotisme]] - [[Religions]] - [[ethnodifférencialisme]] - [[polycentrisme]] - [[Médiation interculturelle]] - [[Habitus (sociologie)]]
23	=== Liens externes ===	=== Liens externes ===
24	*[http://www.ethnociel.qc.ca/ethnocentrisme.html Ethnocentrisme et relativisme culturel sur ''Ethnociel''] *Ethnocide et ethnocentrisme: [http://vadeker.club.fr/corpus/ethnocide.html] * Sow, Adama: [http://www.aspr.ac.at/epu/research/Sow.pdf Ethnozentrismus als Katalysator bestehender Konflikte in Afrika südlich der Sahara, am Beispiel der Unruhen in Côte d`Ivoire] au: [[European University Center for Peace Studies]] (EPU), Stadtschleining	*[http://www.ethnociel.qc.ca/ethnocentrisme.html Ethnocentrisme et relativisme culturel sur ''Ethnociel''] *[http://vadeker.club.fr/corpus/ethnocide.html Ethnocide et ethnocentrisme] ==Notes et références== {{Références|colonnes=2}} [[Catégorie:Ethnocentrisme|*]] [[ca:Etnocentrisme]] [[cs:Etnocentrismus]] [[da:Etnocentrisme]] [[de:Ethnozentrismus]] [[el:Εθνοκεντρισμός]]

```
2005 {{de_icon}}
[[Catégorie:Ethnocentrisme|*]]
[[ca:Etnocentrisme]]
[[cs:Etnocentrismus]]
[[da:Etnocentrisme]]
[[de:Ethnozentrismus]]
[[el:Εθνοκεντρισμός]]
[[en:Ethnocentrism]]
[[eo:Etnocentrismo]]
[[es:Etnocentrismo]]
[[fi:Etnosentrismi]] [[he:
אתנוצנטריות]] [[it:Etnocentrismo]]
[[ja:エスノセントリズム]]
[[ka:ეთნოცენტრიზმი]]
[[nl:Etnocentrisme]]
[[pl:Etnocentryzm]]
[[pt:Etnocentrismo]]
[[ru:Этноцентризм]]
[[sk:Etnocentrizmus]]
[[sl:Etnocentrizem]]
[[sr:Етноцентризам]]
[[sv:Etnocentrism]]
[[uk:Етноцентризм]]
```

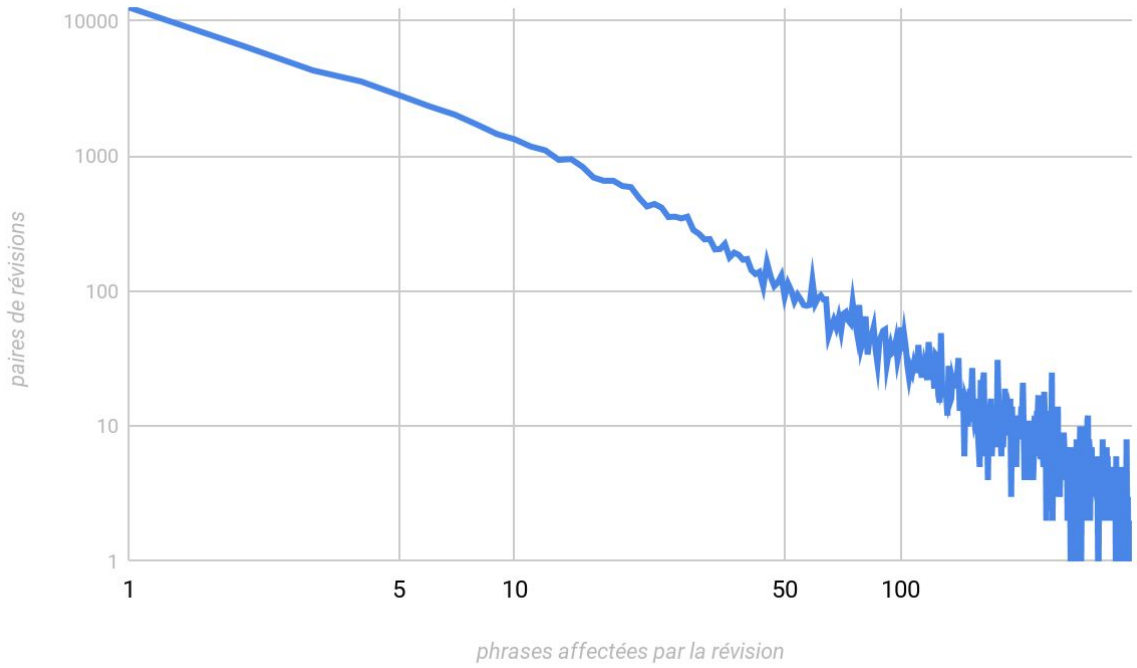
```
[[en:Ethnocentrism]]
[[eo:Etnocentrismo]]
[[es:Etnocentrismo]]
[[fi:Etnosentrismi]] [[he:
אתנוצנטריות]] [[it:Etnocentrismo]]
[[ja:エスノセントリズム]]
[[ka:ეთნოცენტრიზმი]]
[[nl:Etnocentrisme]]
[[pl:Etnocentryzm]]
[[pt:Etnocentrismo]]
[[ru:Этноцентризм]]
[[sk:Etnocentrizmus]]
[[sl:Etnocentrizem]]
[[sr:Етноцентризам]]
[[sv:Etnocentrism]]
[[uk:Етноцентризм]]
```

Annexe 3. Distributions des paires de révisions

- Bulgare



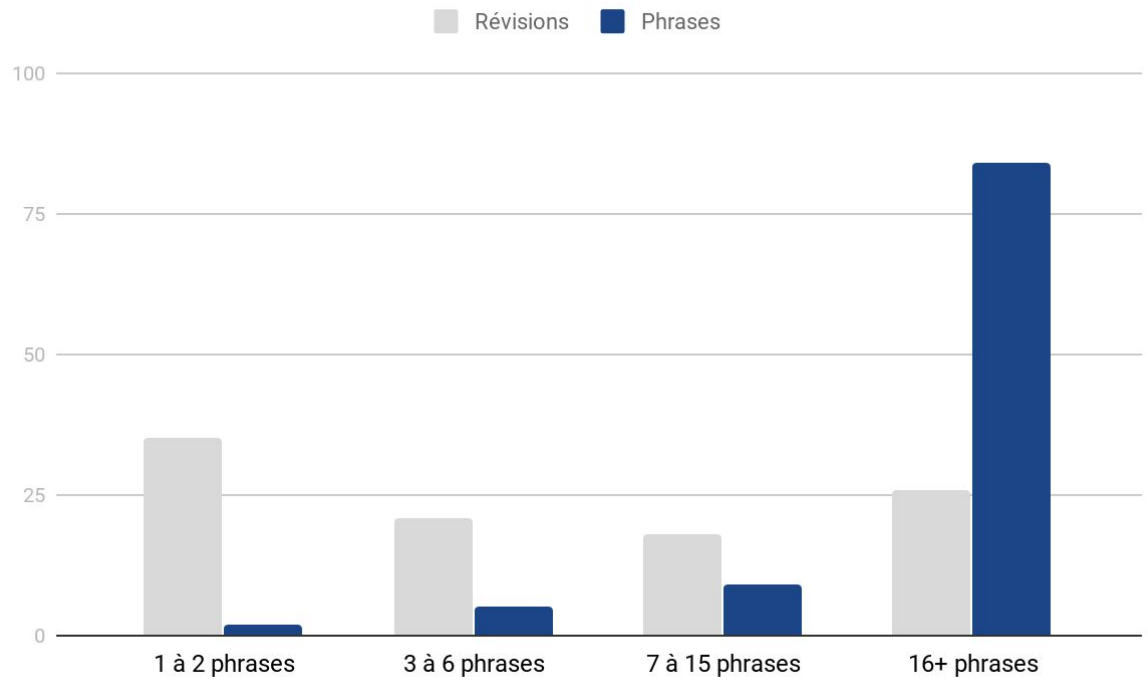
– Anglais


10000
1000
100
10
1
paires de révisions
1
5
10
50
100
phrases affectées par la révision

Annexe 4. Proportion des paires de révisions et de phrases par classe et par langue

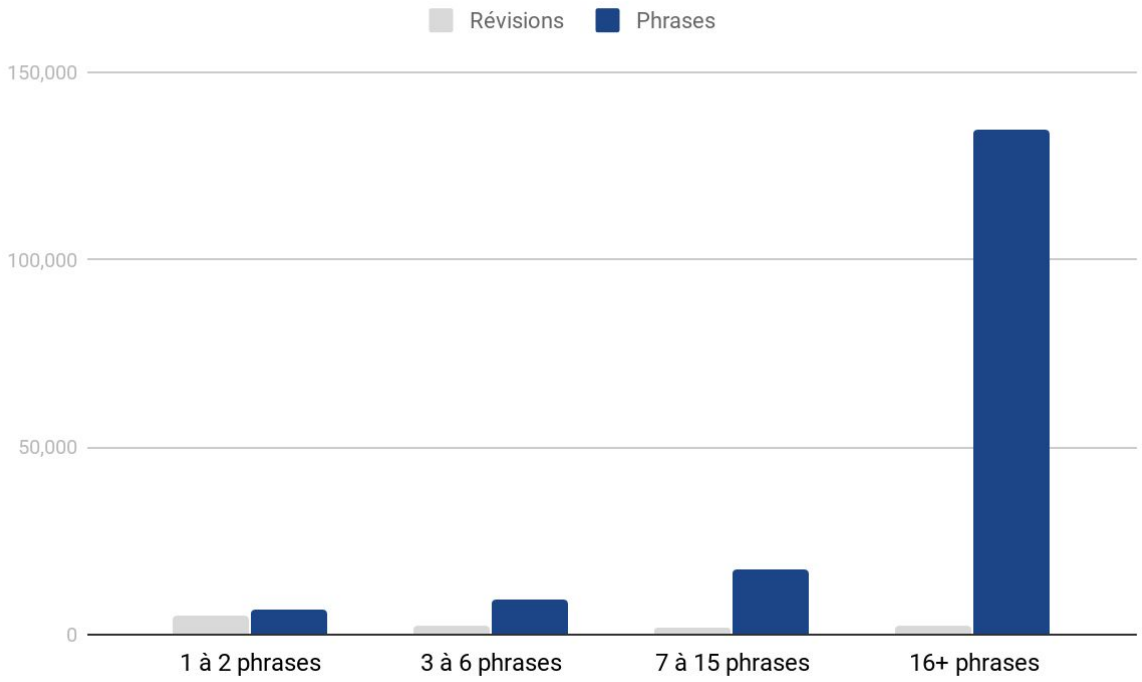
– Bulgare

classe	paires de révisions	phrases
1	105	154
2	94	406
3	90	897
4	117	5742
Total	406	7190



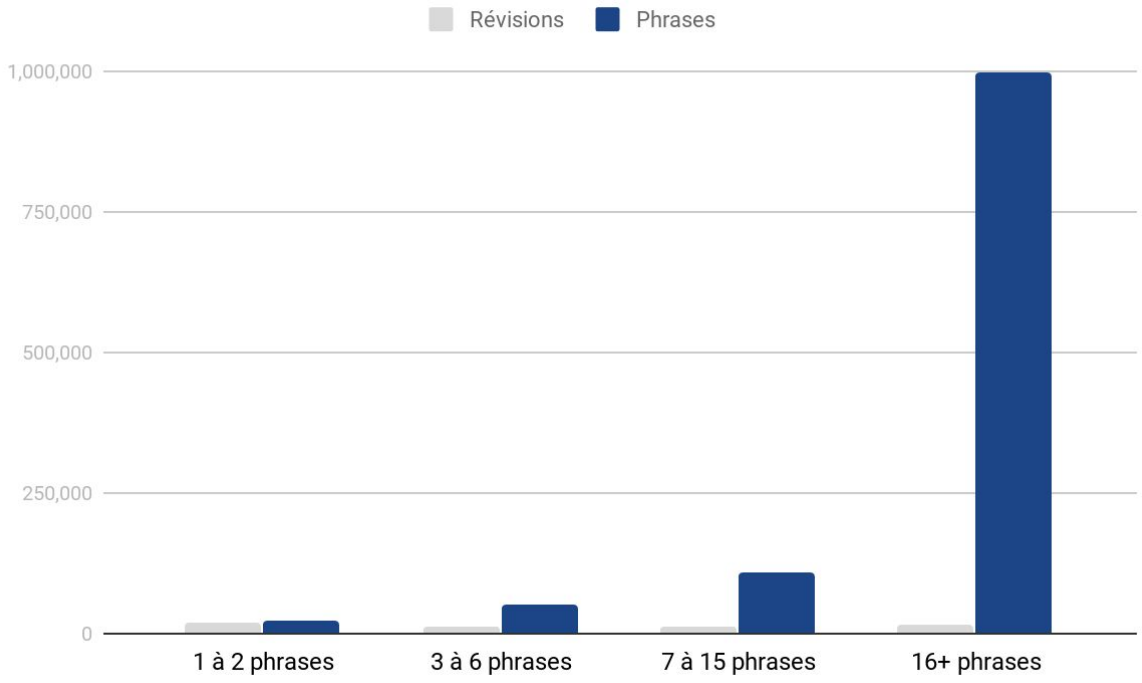
– Français

classe	paires de révisions	phrases
1	5279	6826
2	2306	9596
3	1689	17143
4	2122	134851
Total	11396	168416



– Anglais

classe	paires de révisions	phrases
1	18353	24567
2	12197	51718
3	10670	109309
4	15198	998128
Total	11396	1183722



Annexe 5. Protocole d'évaluation

Task	Evaluate the quality/pertinence of sentences extracted automatically from the revision history of Wikipedia.
Question	Is the removed sentence violating the NPOV policy or not?
You'll be given	- multiple Excel files containing tabs - each tab displays two consecutive versions of a Wikipedia article: columns **Before** and **After** - each version is separated in sentences, one sentence per line - unchanged sentences appear side by side in columns **Before** and **After** - removed sentences appear only in the column **Before** - added sentences appear only in the column **After** - the column Annotation indicates the sentence(s) to consider
Your task is to	- familiarize yourself with the aspects of the Neutrality Point of View Policy of Wikipedia (see pages 2 and 3 of this document) - go through each tab of the Excel documents - locate the cell(s) marked with >>>> in the column **Annotation** - consider the question in regards to the sentence on the marked line in column **Before** - use the immediate context of the sentence in the same revision or the corresponding section of the **After** version to guide your decision, if necessary - replace >>>> with **Y** if you agree that the sentence violates the NPOV or with **N** otherwise. - save changes made to the document and return both files back to sender
Context and data	Neutrality of point of view (NPOV) is one of the three core content policies of Wikipedia. It states that all of the significant views on a topic must be represented fairly (objectively), proportionally (not equally!) and without editorial bias. In this sense, bias could be expressed by both the presence or the absence of content, which means that some edits aiming at correcting bias would consist in deletions, while others in insertions, or both.
Principles to observe	**Avoid stating opinions as facts** For example, an article should not state that *"genocide is an evil action"*, but it may state that *"genocide has been described by John X as the epitome of human evil."* ❑ *Abortion is wrong* – **opinion, not a fact.** ❑ *The pro-life movement holds that abortion is wrong, or occasionally that it is only justified in certain special cases* – **fact, not an opinion.** ❑ *God/spiritual energy/[insert your pet concept here] does/does not exist.* – **opinion, not a fact**. *Nietzsche spent much of his life arguing (among*

	other things) that God does not exist – **fact, not an opinion**. ❏ *Scientologists hold the belief that living cells have a memory. This is based on an erroneous interpretation of the work of Crick and Watson in 1955*. – **opinion, not a fact**. ❏ *Scientologists hold the belief that living cells have a memory. This is based on an interpretation of the work of Crick and Watson in 1955. This interpretation has been heavily criticized by notable cell-biologists such as [whoever]* – **fact, not an opinion**. **Avoid stating facts as opinions** The passage should not be worded in any way that makes it appear to be contested. Words such as *supposed, apparent, alleged* and *purported* can imply that a given point is inaccurate. **Prefer nonjudgmental and neutral language** The tone of Wikipedia articles should be impartial, neither endorsing nor rejecting a particular point of view. Also, It should not be promotional or positively loaded, as much as negatively loaded. ❏ *Bob Dylan is the defining figure of the 1960s counterculture and a brilliant songwriter*. — **promotional language** ❏ *Dylan was included in Time's 100: The Most Important People of the Century, in which he was called "master poet, caustic social critic and intrepid, guiding spirit of the counterculture generation".[refs 1] By the mid-1970s, his songs had been covered by hundreds of other artists. [refs 2]* — **just the facts** **Indicate the relative prominence of opposing views** For example, to state that *"According to Simon Wiesenthal, the Holocaust was a program of extermination of the Jewish people in Germany, but David Irving disputes this analysis"* would be to give apparent parity between the supermajority view and a tiny minority view by assigning each to a single activist in the field.
Words/expressions to avoid	**Puffery** (=peacock words are examples of promotional language) To avoid: *legendary, best, great, acclaimed, iconic, visionary, outstanding, leading, celebrated, award-winning, landmark, cutting-edge, innovative, extraordinary, brilliant, hit, famous, renowned, remarkable, prestigious, world-class, respected, notable, virtuoso, honorable, awesome, unique, pioneering, ...* **Contentious labels** (= an example of loaded language, where the use of a term introduces biase) To avoid: *cult, racist, perverted, sect, fundamentalist, heretic, extremist, denialist, terrorist, freedom fighter, bigot, myth, neo-Nazi, -gate, pseudo-, controversial, ...*

	Unsupported attributions (= weasel words are words and phrases aimed at creating an impression that something specific and meaningful has been said, when in fact only a vague or ambiguous claim has been communicated.) To avoid: *some people say, many scholars state, it is believed/regarded, many are of the opinion, most feel, experts declare, it is often reported, it is widely thought, research has shown, science says, scientists claim, it is often said, ...* **Expressions of doubt** To avoid: *supposed, apparent, purported, alleged, accused, so-called* **Editorializing** (= use of adverbs to highlight something over something else or to indicate particular interpretative viewpoints) To avoid: *notably, it should be noted, arguably, interestingly, essentially, actually, clearly, of course, without a doubt, happily, tragically, aptly, fortunately, unfortunately, untimely* **Synonyms for said** Said, stated, described, wrote, commented, and according to are almost always neutral and accurate. To avoid: *reveal, point out, clarify, expose, explain, find, note, observe, insist, speculate, surmise, claim, assert, admit, confess, deny, ...*

Annexe 6. Exemple de fiche d'annotation

Annotation	Before	After
	la notion de race humaine est une tentative d application à l espèce homo sapiens du concept de race terme qui définit des sous groupes dans une espèce animale	la notion de race humaine est une tentative d application à l espèce homo sapiens du concept de race terme qui définit des sous groupes dans une espèce animale
>>>>	la définition zoologique du terme race est la suivante subdivision d une espèce qui hérite des caractéristiques la distinguant des autres populations de l espèce	(removed)
	(added)	la définition biologique du terme race est la suivante subdivision d une espèce qui hérite des caractéristiques la distinguant des autres populations de l espèce
	au sens génétique une race est une population qui diffère dans l incidence de certains gènes des autres populations conséquence d une isolation le plus souvent géographique	au sens génétique une race est une population qui diffère dans l incidence de certains gènes des autres populations conséquence d une isolation le plus souvent géographique
	en l état actuel des connaissances la pertinence scientifique de ce terme appliqué à l espèce homo sapiens est rejetée par l immense majorité des scientifiques	(removed)
	...	...

Annexe 7. Résultats de l'annotation par langue par annotateur

– Bulgare

	classe	phrases positives	nombre de phrases	proportion des phrases positives		
ann 1	1	34	74	0.46		
	2	48	74	0.65		
	3	51	74	0.69		
	4	60	74	0.81		
		193	296	0.65		
ann 2	1	16	74	0.22		
	2	47	74	0.64		
	3	42	74	0.57		
	4	33	74	0.45		
		138	296	0.47	**kappa**	***s***
moyenne	1	25	74	0.34	0.32	0.12
	2	48	74	0.64	0.22	0.01
	3	47	74	0.63	0.32	0.06
	4	47	74	0.63	-0.23	0.18
		166	**296**	**0.56**	**0.16**	**0.17**

– Français

	classe	phrases positives	nombre de phrases	proportion des phrases positives		
ann 1	1	40	74	0.54		
	2	37	74	0.50		
	3	33	74	0.45		
	4	27	74	0.36		
		137	296	0.46		
ann 2	1	30	74	0.41		
	2	30	74	0.41		
	3	23	74	0.31		
	4	24	74	0.32		
		107	296	0.36	**kappa**	***s***
moyenne	1	35	74	0.47	0.67	0.07
	2	34	74	0.45	0.44	0.05
	3	28	74	0.38	0.61	0.07
	4	26	74	0.34	0.68	0.02
		122	**296**	**0.41**	**0.60**	**0.05**

– Anglais

	classe	phrases positives	nombre de phrases	proportion des phrases positives		
ann 1	1	39	74	0.53		
	2	37	74	0.50		
	3	33	74	0.45		
	4	40	74	0.54		
		149	296	0.50		
ann 2	1	26	74	0.35		
	2	26	74	0.35		
	3	25	74	0.34		
	4	31	74	0.42		
		108	296	0.36		
ann 3	1	48	74	0.65		
	2	38	74	0.51		
	3	41	74	0.55		
	4	45	74	0.61		
		172	296	0.58	**kappa**	***s***
moyenne	1	38	74	0.51	0.55	0.12
	2	34	74	0.45	0.58	0.07
	3	33	74	0.45	0.31	0.09
	4	39	74	0.52	0.39	0.08
		143	**296**	**0.48**	**0.46**	**0.10**