Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

25 | 2019
Varia

Recherche d'indices lexicosyntaxiques de segmentation et de liage par une analyse automatique de corpus

Yves Bestgen

Édition électronique
URL : http://journals.openedition.org/discours/10256
DOI : 10.4000/discours.10256
ISSN : 1963-1723

Éditeur :
Laboratoire LATTICE, Presses universitaires de Caen

Référence électronique
Yves Bestgen, « Recherche d'indices lexicosyntaxiques de segmentation et de liage par une analyse automatique de corpus », *Discours* [En ligne], 25 | 2019, mis en ligne le 30 décembre 2019, consulté le 04 avril 2020. URL : http://journals.openedition.org/discours/10256 ; DOI : https://doi.org/10.4000/discours.10256

discours
Revue de linguistique, psycholinguistique et informatique
http://journals.openedition.org/discours/

Recherche d'indices lexicosyntaxiques de segmentation et de liage par une analyse automatique de corpus

Yves Bestgen
Université catholique de Louvain

Yves Bestgen, « Recherche d'indices lexicosyntaxiques de segmentation et de liage par une analyse automatique de corpus », *Discours* [En ligne], 25 | 2019, mis en ligne le 30 décembre 2019.

URL : http://journals.openedition.org/discours/10256

Titre du numéro : *Varia*
Coordination : Laure Sarda & Denis Vigier

Date de réception de l'article : 29/05/2019
Date d'acceptation de l'article : 06/11/2019

Recherche d'indices lexicosyntaxiques de segmentation et de liage par une analyse automatique de corpus

Yves Bestgen

Université catholique de Louvain

Cette étude emploie une technique automatique d'analyse de corpus pour tenter d'apporter un point de vue complémentaire à celui d'études plus qualitatives des indices de segmentation et de liage, tels que les expressions adverbiales, les connecteurs et les anaphores. L'étude vise tout particulièrement à déterminer s'il est possible de distinguer automatiquement dans des textes les phrases en situation de rupture de celles en situation de continuité et d'identifier les indices qui le permettent. L'identification des phrases en situation de (dis)continuité a été effectuée sur la base de la structuration configurationnelle des textes telle qu'elle est rendue « vi-lisible » par les sections et les paragraphes. Les indices potentiels analysés sont composés des *n*-grammes de lemmes et d'étiquettes morphosyntaxiques. Les analyses ont été effectuées sur trois collections de textes de genre différent : des entrées de Wikipédia, des articles de journaux et des romans. D'une manière générale, l'apprentissage supervisé s'est révélé relativement efficace, obtenant une exactitude allant de 64 % à 74 % alors que le hasard seul obtiendrait 50 %. Les indices les plus utiles pour la discrimination sont pour la plupart interprétables dans le cadre des travaux linguistiques sur les marques de segmentation et de liage. Si les performances de détection des paragraphes sont équivalentes dans les trois genres, on observe des différences importantes lorsqu'on compare les indices les plus utiles dans chaque genre. Après avoir discuté quelques-unes des limites de l'étude, la conclusion envisage la possibilité de prendre en compte d'une manière plus complète les indices liés à la coréférence, qui se sont révélés particulièrement utiles.

Mots clés : linguistique textuelle, marqueurs du discours, approche onomasiologique, adverbiaux, expressions coréférentielles, connecteurs, genre de textes, apprentissage supervisé

This study uses an automated corpus analysis technique to try to provide a complementary point of view to that of more qualitative studies of segmentation and linking indices, such as adverbial expressions, connectors and anaphora. The study is specifically aimed at determining whether it is possible to automatically distinguish in texts sentences opening or not a discourse segment and to identify the indices that allow it. The identification of sentences in (dis)continuity situation was carried out on the basis of the segments made visible in the texts by means of the sections and paragraphs. The potential indices were n*-grams of lemmas and part-of-speech tags. Analyses were conducted on three collections of texts of different genres: Wikipedia entries, newspaper articles and novels. In general, supervised learning has been relatively effective, with accuracy ranging from 64% to 74%, while chance alone would get 50%. The most useful indices for discrimination are for the most part interpretable in the context of the linguistic theory on segmentation and linking marks. While paragraph detection performance is equivalent in all three genres, there are significant differences when comparing the most useful indices in each genre. After discussing some of the limitations of the study, the conclusion considers the possibility of taking more fully into account the coreference indices, which have proved particularly useful.*

Keywords: *text linguistics, discourse markers, onomasiological approach, adverbials, co-referential expressions, connectors, text genre, supervised learning*

Yves Bestgen est chercheur qualifié du Fonds de la recherche scientifique (FRS-FNRS). Le présent article est issu d'une présentation orale effectuée lors de la journée d'étude « Référence, coréférence et structure textuelle » organisée par le projet ANR DEMOCRAT (« Description et modélisation des chaînes de référence : outils pour l'annotation de corpus et le traitement automatique ») le 27 novembre 2017 à Lyon. L'auteur tient à remercier les organisateurs de cette journée. Il tient aussi à remercier les rapporteurs pour leurs nombreuses remarques et propositions d'amélioration.

1. Introduction

1 Depuis de nombreuses années, la linguistique textuelle s'intéresse aux indices lexicosyntaxiques de segmentation et de liage (p. ex. Adam, 1990 et 2016 ; Charolles, 1988 et 1993 ; Charolles *et al.*, 2005 ; Le Draoulec et Péry-Woodley, 2005 ; Chafe, 1984 ; Grimes, 1975 ; Longacre, 1979 ; Marcu, 2000 ; Sarda *et al.*, 2014 ; Virtanen, 1992). Il s'agit principalement d'expressions adverbiales (*un matin*, *vers neuf heures*, *en France*), de connecteurs (*et*, *mais*, *alors*), de marqueurs métadiscursifs (*pour en revenir*, *plus généralement*) et d'expressions anaphoriques (*elle*, *son*, *celui-ci*). Ces indices retiennent également l'attention de chercheurs en psychologie cognitive (Anderson *et al.*, 1983 ; Bestgen et Vonk, 2000 ; Colonna *et al.*, 2014) et en traitement automatique des langues (Ferret *et al.*, 2001 ; Jurafsky et Martin, 2009 ; Passonneau et Litman, 1997) parce qu'ils sont susceptibles de faciliter la compréhension d'un texte et d'améliorer les performances d'algorithmes de segmentation automatique. Dans le champ de l'enseignement des langues, leur maîtrise lors d'une production en langue maternelle comme en langue seconde est un signe important de compétence discursive (Garcia-Debanc, 2010 ; Piérard et Bestgen, 2008 ; Vigier, 2008).

2 Deux grandes approches peuvent être employées pour étudier ces expressions (Hansen, 1997 ; Jacques et Poibeau, 2010). L'approche la plus classique est l'approche sémasiologique qui consiste à rechercher les occurrences d'une expression linguistique donnée dans des textes afin de déterminer les raisons de sa présence. De telles analyses peuvent être effectuées sur quelques textes, mais également et aisément sur de grands corpus puisque « lorsque l'objet d'analyse est une forme, le repérage des occurrences ne pose guère de problème : il s'agit simplement de rechercher cette forme et ses éventuelles flexions dans les textes » (Jacques et Poibeau, 2010 : 1). Cette approche s'est révélée particulièrement fructueuse, permettant la description de nombreux candidats-marqueurs de la segmentation du discours. On peut citer les études sur la fonction cadrative, ou plus simplement de segmentation, des expressions adverbiales temporelles et spatiales que ce soit dans de petits nombres de textes ou dans des collections de textes relativement grandes (Charolles, 2007 ; Crompton, 2006 ; Ho-Dac et Péry-Woodley, 2009 ; Piérard et Bestgen, 2006 ; Virtanen, 1992), mais aussi des études centrées sur une expression spécifique comme *à travers* (Stosic, 2012) ou *selon X* (Schrepfer-André, 2005). La limitation principale de l'approche sémasiologique est qu'elle ne peut être employée que lorsqu'on dispose d'une liste de candidats-marqueurs afin de vérifier si ceux-ci fonctionnent bien comme tels. Cette limitation est particulièrement explicite dans l'ambitieuse étude de Marcu (2000) qui

a développé des algorithmes susceptibles d'identifier automatiquement la structure rhétorique d'un texte depuis le niveau phrastique jusqu'à celui du texte complet en passant par celui des paragraphes qui le composent. L'ensemble du système repose toutefois sur une liste initiale de connecteurs («*cue-phrases*») qui ont été analysés lors d'une importante étude de corpus afin de spécifier le plus précisément possible leurs fonctions ainsi que les conditions dans lesquelles ces indices sont valides. Seules les structures signalées par ces connecteurs peuvent être identifiées.

3 La seconde approche, dite onomasiologique, «prend comme point de départ un certain sens ou un certain effet pragmatique et comme objectif la mise en évidence des formes linguistiques susceptibles de produire ce sens ou cet effet pragmatique» (Jacques et Poibeau, 2010 : 1). Elle peut être illustrée par l'étude de Jacques et Poibeau (2010) sur les textes procéduraux qui met en lumière toute une série d'éléments linguistiques qui ont été décisifs pour les annotateurs lors de l'identification des segments spécifiquement procéduraux dans les textes. Elle l'est aussi par les recherches du CLLE (Cognition, langues, langage, ergonomie) de l'université de Toulouse (Ho-Dac *et al.*, 2012 ; Péry-Woodley *et al.*, 2017) à propos du signalement des structures énumératives, qui s'appuient sur ANNODIS («Annotation discursive»), un corpus de français écrit de plus de 600 000 mots enrichi d'annotations discursives. Cette approche est particulièrement heuristique. Non seulement elle ne nécessite pas une liste initiale de candidats-marqueurs, mais elle permet aussi d'identifier des non-occurrences de marqueurs, par exemple des situations de discontinuité qui ne semblent pas signalées. Cette approche est toutefois difficilement applicable à de très grands corpus en raison de la lourdeur de l'annotation nécessaire pour localiser les structures textuelles cibles. Or, l'analyse de grands corpus est souvent nécessaire compte tenu de la rareté d'emploi d'un grand nombre d'indices linguistiques de la segmentation (Piérard et Bestgen, 2006 ; Stosic, 2012). La situation devient encore plus problématique lorsque l'étude porte sur la comparaison de genres de textes puisque celle-ci a pour effet de multiplier l'ampleur du travail d'annotation à effectuer.

4 L'objectif de la présente étude est d'évaluer une procédure, fondée sur des techniques de traitement automatique des langues (TAL), qui permet l'extraction d'indices de segmentation et de liage dans de très grandes collections de textes de genre différent. Il s'agit d'une étude exploratoire qui s'apparente bien plus à une tentative de preuve du concept qu'à un aboutissement. Néanmoins, les résultats obtenus lors de la mise à l'épreuve de la procédure ont permis de confirmer des hypothèses linguistiques qui n'avaient pas été prises en compte explicitement lors de la mise au point de la procédure.

2. Procédure proposée et travaux antérieurs

5 Une procédure susceptible d'identifier automatiquement des indices de segmentation et de liage doit nécessairement remplir les deux conditions suivantes : être capable d'identifier automatiquement des segments dans des textes et être capable de sélectionner parmi des candidats-marqueurs identifiés automatiquement ceux qui sont les plus efficaces. En ce qui concerne la première condition, j'ai choisi de

centrer l'étude sur les segments au sens d'Adam (1990 et 2016), appelés séquences par Charolles (1988), qui résultent du découpage du matériau discursif en blocs textuels « vi-lisibles » au moyen des titres et des alinéas principalement. Ces segments, aisément identifiables automatiquement, résultent « d'un travail explicite d'organisation de l'énonciation visant en particulier à faciliter la tâche de l'interprétation » (Charolles, 1988 : 9), ce qu'Adam (2015 : 2) formule de la manière suivante :

> la segmentation en paragraphes facilite et programme la lecture en donnant, par les encoches ou entailles entre paragraphes et entre sections regroupant des ensembles de paragraphes, des instructions de maintien temporaire d'informations en mémoire de travail et de mise en relation des informations textuelles par étapes ou boucles de traitement.

6 Même s'il ne fait pas de doute que la mise en paragraphes d'un texte peut résulter de divers objectifs (Stark, 1988), nombre d'auteurs ont souligné que leur principale fonction est de segmenter le texte en unités de sens (p. ex. Bessonnat, 1988 ; Ferret *et al.*, 2001 ; Hofmann, 1989 ; Longacre, 1979 ; voir Adam, 2018, pour une analyse approfondie). Il s'agit d'ailleurs d'un indice relativement fiable de fin de la portée cadrative d'un adverbial (Schrepfer-André, 2005 ; Vigier, 2005). La fonction de segmentation du texte des titres des sections et des sous-sections est évidemment elle aussi bien établie (Jacques, 2005 ; Lemarié *et al.*, 2012 ; Rebeyrolle *et al.*, 2009).

7 Pour remplir la seconde condition, des techniques classiques en TAL ont été utilisées : l'étiquetage morphosyntaxique, l'extraction automatique de *n*-grammes et l'apprentissage supervisé. La procédure proposée repose donc sur la construction, par un algorithme d'apprentissage supervisé, de modèles prédictifs capables de discriminer dans des textes des paires de phrases contiguës initiant ou non des blocs textuels « vi-lisibles ». Les indices potentiels analysés sont composés des unigrammes, bigrammes et trigrammes de lemmes et d'étiquettes morphosyntaxiques présents dans les phrases. L'hypothèse sous-jacente à cette approche est que les indices les plus efficaces pour discriminer les phrases en situation de rupture de celles en situation de continuité seront des candidats à la fonction de marques de segmentation et de liage. Cette procédure s'inspire très largement des travaux sur la détection automatique de paragraphes.

8 Sporleder et Lapata (2006) se sont intéressées au découpage automatique en paragraphes de textes au moyen de traits de surface comme la ponctuation, la présence de guillemets, la longueur de la phrase et sa position dans le texte, la présence de mots spécifiques et la proportion de mots communs avec la phrase précédente. Ces traits sont employés par une procédure d'apprentissage supervisé dont la fonction est de distinguer dans des textes les phrases en début de paragraphe des autres phrases. Sporleder et Lapata ont analysé trois domaines différents (fiction, actualités, parlement) dans trois langues différentes (anglais, allemand et grec). L'efficacité de leur système est élevée puisqu'elle se situe fréquemment à quelques pour-cent à peine de la performance de juges humains, mais certains traits sont problématiques pour l'objectif de la présente étude, comme ceux qui permettent de détecter les alinéas de discours direct, qui sont considérés comme des débuts de paragraphe, et le recouvrement lexical. De plus,

leur objectif est de construire le meilleur système possible pour prédire le début des paragraphes, alors que, dans la présente étude, il s'agit d'identifier des indices lexicaux et morphosyntaxiques de liage et de segmentation. L'approche proposée par ces auteures semble toutefois très prometteuse puisqu'elles observent parmi les mots les plus utiles pour détecter un début de paragraphe des adverbiaux comme *in early trading* et des connecteurs comme *in addition to*. Filippova et Strube (2006) ont poursuivi cette ligne de recherche, mais en analysant exclusivement des biographies extraites de Wikipédia, ce qui leur a permis d'employer des traits spécifiques à ce type de textes comme la forme employée pour mentionner la personne en question. Ils notent toutefois que les connecteurs et autres marqueurs du discours sont peu utiles. Plus récemment, Lai *et al.* (2016) ont employé la même approche pour segmenter en paragraphes des discussions orales. Même si les performances du système proposé sont loin d'être parfaites, elles ont pu souligner l'importance des indices prosodiques, mais aussi des indices lexicaux de rupture. Ces études suggèrent que la recherche d'indices de segmentation et de liage au moyen de l'analyse de grandes quantités de texte par une procédure d'apprentissage supervisé pourrait se révéler efficace. Il est aussi nécessaire de mentionner que Piérard et Bestgen (2006) et Ho-Dac et Péry-Woodley (2009) ont analysé l'occurrence d'expressions adverbiales en fonction du découpage en paragraphes, mais l'approche employée est sémasiologique, cherchant à évaluer des hypothèses linguistiques à propos du statut d'indices de segmentation de ces expressions.

9 Les deux sections suivantes présentent une évaluation de la procédure proposée au moyen de l'analyse de trois collections de textes de taille et de genre différents. Elles mettent en lumière les potentialités, mais aussi les limitations, de la technique, qui sont discutées dans la dernière section.

3. Méthode

3.1. Collection de textes et prétraitement

10 Les analyses ont été effectuées sur trois collections de textes de taille et de genre très différents : des entrées de l'encyclopédie Wikipédia francophone (« Wiki »), des articles parus dans quatre années du journal belge francophone *Le Soir* entre 1995 et 1998 (« *Le Soir* ») et des romans des XIX^{e} et XX^{e} siècles rédigés en français (« Roman »). Tous ces documents ont été lemmatisés et étiquetés morphosyntaxiquement par le TreeTagger (Schmid, 2003) en employant le modèle de langue standard pour le français, fourni sur le site de ce logiciel. Comme les traitements nécessaires pour transformer ces documents en un matériel utilisable pour répondre aux objectifs de l'étude sont partiellement différents selon la source de ceux-ci, ils sont décrits séparément ci-dessous.

3.1.1. « Wiki »

11 Les documents analysés ont été extraits au moyen de l'outil WikiExtractor d'Attardi, un script en Python qui extrait et nettoie des textes issus de Wikipédia. Quelques prétraitements complémentaires ont été nécessaires afin de conserver les balises

signalant les sections. Les phrases utilisées ont été sélectionnées après suppression des documents se distinguant nettement des valeurs moyennes (sur la base des scores *z*) dans la collection quant à leur longueur, au nombre de phrases très brèves ou au nombre d'étiquettes morphosyntaxiques des catégories « abréviation », « nom propre », « nombre » et « signe de ponctuation ». À l'issue de ces prétraitements, l'ensemble de données Wiki était composé d'approximativement 260 millions de « tokens » (mots, mais aussi signes de ponctuation, symboles...).

3.1.2. « Le Soir »

12 Les articles du journal *Le Soir* ont été extraits au moyen de scripts *ad hoc* des CD-ROM diffusés par ce journal. Les mêmes critères pour éliminer les articles peu pertinents pour l'analyse que ceux mentionnés ci-dessus ont été employés. Le matériel disponible pour les analyses est composé de 70 millions de tokens.

3.1.3. « *Roman* »

13 Soixante-sept romans des XIX^{e} et XX^{e} siècles, comme *Bouvard et Pécuchet* de Flaubert, *Le Rouge et le Noir* de Stendhal, *Germinie Lacerteux* des frères Goncourt, ont été extraits des bases ABU (bibliothèque de l'Association des bibliophiles universels), IntraText et Wordthèque. Le seul prétraitement a consisté à identifier les alinéas de discours direct sur la base du tiret ouvrant une prise de parole et à les supprimer. Après cette suppression, le matériel est composé de 3 200 000 tokens.

3.2. Procédure d'apprentissage supervisé

14 Pour pouvoir atteindre les objectifs de l'étude au moyen d'une procédure d'apprentissage supervisé, quatre ingrédients sont nécessaires : des catégories à comparer, des exemples de chaque catégorie, des indices potentiels pour la discrimination et une procédure d'apprentissage.

3.2.1. Catégories : position dans un segment textuel

15 L'objectif de l'étude est de comparer des phrases qui se trouvent ou non en début de segments identifiées sur la base de la structuration configurationnelle des documents qui est très variable selon la source. Les documents issus de Wikipédia sont découpés en sections, sous-sections et en paragraphes. La hiérarchie des sections et sous-sections peut couvrir plus de cinq niveaux, mais plus un niveau est bas dans la hiérarchie plus faible est sa fréquence dans l'ensemble de données. Dans « *Le Soir* », on trouve des intertitres et des paragraphes, les premiers ayant une fonction textuelle clairement différente des titres de sections dans « Wiki » (Adam et Lugrin, 2000 ; Rebeyrolle *et al.*, 2009). Dans « Roman », seuls les paragraphes peuvent être analysés parce qu'il n'y a pas assez de chapitres pour utiliser une procédure d'apprentissage supervisé afin d'identifier des indices de rupture pertinents. Face à cette diversité, j'ai choisi d'analyser les paragraphes, présents dans les trois collections de textes, ainsi que les sections principales dans « Wiki » et les passages délimités par des intertitres dans « *Le Soir* ». Dans ces deux derniers

cas, les titres des sections dans « Wiki » et les intertitres dans « *Le Soir* » ont été supprimés du matériel analysé.

16 Au total, l'étude porte donc sur les cinq conditions suivantes : les titres principaux dans « Wiki » (Wiki S), les intertitres dans « *Le Soir* » (Soir I), les paragraphes dans « Wiki » (Wiki P), les paragraphes dans « *Le Soir* » (Soir P) et les paragraphes dans « Roman » (Roman P).

3.2.2. Exemples : des phrases en situation de continuité ou de rupture

17 Pour obtenir les phrases à analyser, tous les quadruplets de phrases contiguës, dont chaque phrase contenait de 7 à 43 mots et se trouvait dans un paragraphe de 3 à 15 phrases (mais pas nécessairement le même), ont été extraits de chaque collection de textes. Ces valeurs ont été choisies sur la base des distributions de longueurs des phrases et des paragraphes dans les trois corpus. Pour la longueur des phrases, elles éliminent les 15 à 20 % des phrases les plus extrêmes dans les trois corpus. Pour la longueur des paragraphes, ceux de moins de trois phrases ne peuvent remplir les critères exposés ci-dessous et la limite à quinze phrases au maximum élimine moins de 1 % des paragraphes dans les trois corpus. Pour être considérée en situation de rupture, la troisième phrase de ces quadruplets devait être précédée par une rupture de section, un intertitre ou par un alinéa et il ne pouvait pas y avoir d'alinéas entre les autres phrases du quadruplet. La troisième phrase d'un quadruplet était considérée en situation de continuité s'il n'y avait aucune rupture de la structuration configurationnelle entre les quatre phrases. Une même phrase ne pouvait appartenir qu'à un seul quadruplet sélectionné.

18 La figure 1 illustre cette procédure de sélection des phrases employées pour l'apprentissage supervisé. Le symbole « <2> » signale une nouvelle section. Comme expliqué ci-dessus, le titre correspondant n'est pas pris en compte. Le symbole « <P> » signale un nouveau paragraphe. Les quadruplets de phrases sont encadrés en bleu. La phrase encadrée en rouge est sélectionnée comme un exemple de rupture de type « Paragraphe », celle en vert comme un exemple de phrase en situation de continuité.

19 Dans chaque collection de textes, le maximum possible de phrases en situation de rupture a été extrait ainsi qu'un échantillon aléatoire d'une même taille de phrases en situation de continuité. Cette répartition équilibrée entre les deux catégories, qui ne se retrouve pas dans les corpus, ni dans les travaux antérieurs sur l'identification des paragraphes (Sporleder et Lapata, 2006 ; Filippova et Strube, 2006), présente l'intérêt d'augmenter très fortement la proportion de phrases en début de paragraphes et donc de rendre *a priori* les indices de rupture aussi utiles que ceux de continuité. Le nombre d'exemples de continuité ou de rupture et le nombre de tokens [1] dans les cinq conditions analysées sont donnés dans le tableau 1.

1. Le token plutôt que le mot a été choisi comme unité de comptage parce que les analyses prendront aussi en compte les signes de ponctuation internes à la phrase comme la virgule. Pour rappel, il y a, par construction, exactement le même nombre de phrases en situation de continuité que de phrases en situation de rupture et 20 % de ces phrases sont employés comme matériel de test.

Figure 1 – Procédure de sélection des phrases employées comme exemples

	Wiki S	Soir I	Wiki P	Soir P	Roman P
Nombre de phrases dans chaque catégorie	19216	6966	138723	44265	2697
Nombre de tokens dans les exemples de rupture	460159	155461	3350517	956344	63045
Nombre de tokens dans les exemples de continuité	451575	149055	3227415	934302	66095

Tableau 1 – Matériel sélectionné pour les analyses

3.2.3. *Indices*

20 Les indices mis à la disposition de la procédure d'apprentissage supervisé sont les séquences contiguës ou *n*-grammes d'un à trois lemmes et d'une à trois étiquettes morphosyntaxiques, présentes dans chaque phrase sélectionnée comme exemple. Lors de l'extraction, les *n*-grammes en début de phrases ont été distingués de ceux

qui ne le sont pas. Cette étape est illustrée dans le tableau 2. La présence d'un indice dans une phrase était codée en « présent *versus* absent » (codage binaire).

21 Comme la longueur moyenne d'une phrase dans les trois corpus est approximativement de 23 tokens, cet encodage des exemples produit 6 indices pour le début et en moyenne 126 indices pour la suite, mais seuls ceux qui passent le seuil de fréquence (voir la section suivante) sont conservés.

Alors, il eut une idée.
Début : alors alors_, alors_,_il ADV ADV_PUN ADV_PUN_PRO :PER Suite : , il ,_il avoir il_avoir ,_il_avoir une ... PUN PRO :PER PUN_PRO :PER VER :simp PRO :PER_VER :simp PUN_PRO :PER_VER :simp DET :ART ...

Tableau 2 – *N*-grammes employés comme indices par la procédure

3.2.4. Algorithme d'apprentissage supervisé

22 Les modèles prédictifs ont été construits par apprentissage supervisé au moyen de la régression logistique régularisée par la norme L2 [2]. Celle-ci tente de prédire la probabilité qu'un exemple appartienne à une des deux catégories en pénalisant les modèles trop complexes afin d'essayer d'éviter un surapprentissage qui nuirait à la généralisabilité du modèle. Le logiciel LIBLINEAR de Fan *et al.* (2008) a été employé avec l'option « -s 7 ».

23 Pour chaque analyse, les phrases cibles ont été divisées en 80 % pour l'apprentissage et 20 % pour l'évaluation. Deux paramètres ont été optimisés sur 25 % des données d'apprentissage : le paramètre *C* de régularisation de la régression logistique et le seuil de fréquence minimal des indices dans le matériel analysé. Lors de l'application du seuil de fréquence, on a surpondéré les indices présents en début de phrases afin de les rendre en moyenne aussi fréquents que les autres.

24 L'exactitude (« *accuracy* », soit le pourcentage d'exemples bien classés par rapport au nombre total d'exemples à classer) a été employée comme mesure d'efficacité. La précision, le rappel et la mesure F1 ont aussi été calculés, mais ces valeurs étaient quasi identiques à l'exactitude et celle-ci présente l'avantage d'être d'une interprétation évidente.

25 Pour identifier les indices les plus utiles permettant de distinguer les phrases en situation de rupture de celles en situation de continuité, les poids qui leur sont attribués par la régression logistique ont été employés sans aucune transformation étant donné que les différences entre les poids ont un impact direct et proportionnel

2. Cette norme présente l'avantage par rapport à la norme L1 de produire des modèles dans lesquels plusieurs indices utiles et fortement corrélés entre eux ne sont pas pénalisés, une propriété fondamentale pour atteindre les objectifs de l'étude.

sur la fonction de classification du modèle parce que les indices sont codés d'une manière binaire. Les 200 indices ayant reçu les poids les plus élevés en valeur absolue ont été sélectionnés [3] pour les analyses détaillées.

4. Résultats

26 Comme expliqué ci-dessus, les analyses ont été effectuées sur cinq ensembles de données que l'on appellera aussi « conditions » dans la suite : les sections (S) et les paragraphes (P) dans « Wiki », les intertitres (I) et les paragraphes dans « *Le Soir* », et les paragraphes dans « Roman ». Les analyses avaient pour objectif de recueillir deux informations principales : quel est le niveau d'efficacité de la discrimination et quels sont les types d'indices les plus utiles ?

4.1. Efficacité de la discrimination

27 Le tableau 3 présente l'exactitude atteinte sur le matériel de test des cinq conditions pour les ensembles d'indices suivants : tous les indices disponibles (Tout), seulement les indices présents en début de phrases (Début) et seulement les indices non présents en début de phrases (Suite). D'une manière générale, l'apprentissage supervisé s'est révélé relativement efficace puisqu'il obtient, sur la base de tous les indices, une exactitude allant de 64 % à 74 % alors que le hasard seul obtiendrait 50 %. Une exactitude de 64 % pourrait être perçue comme peu élevée, mais, comme l'a souligné un relecteur, il est probable que des ruptures non signalées par un alinéa se produisent à l'intérieur des paragraphes, rendant la discrimination complexe.

Conditions	Tout	Début	Suite
Wiki S	74	67	70
Soir I	64	62	62
Wiki P	65	62	62
Soir P	64	61	61
Roman P	64	59	60

Tableau 3 – Efficacité (exactitude) de la discrimination en pourcentage

28 Comme on peut le voir, ce sont les phrases en début de sections dans « Wiki » qui sont les plus facilement discriminables des phrases en situation de continuité. On observe très peu de différences entre les quatre autres conditions. La très nette différence entre les sections de « Wiki » et les intertitres dans « *Le Soir* » s'accorde

3. On a vérifié que les résultats étaient similaires si on analysait seulement les 100 premiers indices. On a aussi vérifié que ces poids étaient significativement différents de 0 au moyen d'un test de permutation (Howell, 2008 : chap. 18).

avec le statut particulier de ceux-ci, souligné dans la littérature. Les intertitres des journaux, qui ont fréquemment une origine énonciative différente de celle de l'article (Adam, 1997 : 5), remplissent des fonctions d'aération du texte et de mise en évidence d'un élément susceptible d'attirer l'attention du lecteur (Adam et Lugrin, 2000 ; Védénina, 1989).

29 Lorsqu'on limite le type d'indices disponibles, les performances diminuent systématiquement, mais l'effet est surtout manifeste pour les sections dans « Wiki ». On observe peu de différence entre les indices présents en début de phrases et ceux situés plus loin, sauf, de nouveau, dans le cas des sections de « Wiki » pour lesquelles les indices qui ne sont pas en début de phrases sont plus efficaces que ceux qui sont en début.

4.2. Indices les plus importants pour la discrimination

30 Les analyses qui suivent portent sur les 200 indices les plus importants pour chaque discrimination auxquels il est fait référence dans la suite par le terme « indices sélectionnés ». Le tableau 4 présente les 12 indices sélectionnés dans les cinq ensembles de données[4]. Un treizième indice, « NAM », un nom propre, présent dans la suite de la phrase, est sélectionné dans les cinq ensembles, mais il est un indice de continuité dans « Wiki » et de rupture dans les autres[5]. En revanche, positionné en début de phrase, un nom propre est toujours un indice de rupture, une fonction qui peut être mise en relation avec son rôle d'initiateur d'une chaîne de référence (Schnedecker, 1997). On constate que la majorité de ces indices communs sont composés d'étiquettes morphosyntaxiques (10 sur 12), de signaux de continuité (9 sur 12) et sont situés en début de phrase (9 sur 12). On y trouve principalement des reprises anaphoriques (pronoms personnels et démonstratifs, déterminants possessifs). Les trois indices de rupture incluent tous un nom propre.

31 Ces observations s'accordent parfaitement avec les théories linguistiques sur les indices de segmentation et de liage. Le seul indice d'une autre nature est la présence d'un verbe conjugué au conditionnel en tant qu'indice de continuité dans les cinq collections. Tant en situation de continuité que de rupture, les plus fréquents de ces verbes sont (évidemment) des verbes d'emploi très fréquent dans la langue comme les auxiliaires et semi-auxiliaires (*avoir*, *être*, *pouvoir*, *devoir*). Simplement, ils sont observés nettement plus souvent en situation de continuité que de rupture. Le tableau 4 ne contient pas de marqueurs classiques de segmentation que sont, par exemple, les adverbiaux temporels ou spatiaux présents en début de phrases. Ces indices sont néanmoins sélectionnés, mais dans un plus petit nombre de conditions, voire dans une seule, comme on le verra plus loin.

---

4. Pour rappel, ce sont les formes lemmatisées par le TreeTagger qui ont été analysées et qui sont donc présentées dans ce rapport.

5. Seule une analyse qualitative détaillée, qui dépasserait les objectifs exploratoires de la présente étude, permettrait de comprendre l'origine de cette différence. On note néanmoins que la proportion des noms propres parmi l'ensemble des étiquettes morphosyntaxiques dans « Wiki » est une fois et demie plus grande que dans « *Le Soir* » et trois fois plus grande que dans « Roman ».

	Début	Suite
Rupture	NAM	PRP_NAM NOM_PRP_NAM
Continuité	elle il DET:POS PRO:DEM PRO:PER DET:POS_NOM PRO:PER_PRO:PER PRO:PER_VER:impf	VER:cond

Tableau 4 – Indices sélectionnés dans les cinq conditions

32 Il apparaît donc que la très grande majorité des indices sélectionnés ne sont pas communs à l'ensemble des conditions. Deux analyses ont été effectuées afin d'essayer de mieux comprendre leurs caractéristiques et leurs distributions dans les cinq conditions. Le tableau 5 présente le pourcentage que représentent, par rapport aux 200 indices les plus importants, les indices situés en début de phrases, ceux composés de lemmes, les unigrammes et les indices dont la présence prédit une rupture. Afin de limiter la taille du tableau et de faciliter la lecture, la ou les autres conditions qui interviennent dans le calcul de chaque ligne (par exemple, les indices non situés en début de phrases pour ceux en début ou les bigrammes et les trigrammes pour les unigrammes) ne sont pas présentées. Ces valeurs, plus exactement les tables de contingence complètes des fréquences brutes obtenues indépendamment pour chaque type d'indices (position dans la phrase, longueur du *n*-gramme…), ont été analysées au moyen du test du khi-carré de Pearson. Les valeurs données sur la deuxième ligne de chaque type d'indices indiquent le pourcentage observé en plus ou en moins par rapport au pourcentage attendu sous l'hypothèse d'indépendance testée par le khi-carré. Les valeurs suivies d'une étoile sont statistiquement significatives au seuil de 0,01.

	Wiki S	Soir I	Wiki P	Soir P	Roman P	Khi²	*p*
Début	40 -3	34 -8	56 +14*	42 -1	40 -3	21,3	0,0003
Lemme	38 -10*	47 -1	49 +1	58 +11*	46 -2	17,3	0,0017
Unigramme	31 -2	33 0	36 +3	39 +4	27 -6	9,22	> 0,30
Rupture	48 -5	58 +5	40 -13*	50 -3	70 +17*	40,8	< 0,0001

Tableau 5 – Pourcentages d'indices de différents types dans les cinq conditions

33 On observe que les différents types d'indices ne sont pas présents d'une manière égale dans les 200 indices les plus utiles pour chaque corpus. Dans les sections de « Wiki », ce sont les étiquettes morphosyntaxiques (la catégorie complémentaire des lemmes) qui sont surreprésentées alors que les intertitres dans « *Le Soir* » ne se distinguent pas des autres conditions. Pour les paragraphes, les traits les plus importants pour « Wiki » sont plus fréquemment en début de phrases et signalent la continuité ; ceux dans « *Le Soir* » sont plus souvent des lemmes et ceux dans « Roman », des indices de rupture.

34 On note aussi qu'il n'y a pas de différence statistiquement significative entre les conditions lorsqu'on analyse la longueur des *n*-grammes. Cette variable mérite néanmoins des commentaires supplémentaires. Il y a en effet une très forte relation entre la longueur des *n*-grammes et l'opposition entre lemme et étiquette morphosyntaxique. Plus un *n*-gramme est long et moins il a de chance d'être sélectionné dans le cas des lemmes alors que c'est l'inverse pour les étiquettes morphosyntaxiques. Cette observation s'explique par la relation bien connue : plus un *n*-gramme est long et moins il est fréquent. Dans le cas des lemmes, un *n*-gramme devient vite trop rare pour être utile ; dans le cas des étiquettes morphosyntaxiques, les unigrammes sont trop imprécis (trop fréquents) pour être très utiles. Toutefois, même si on effectue des analyses séparées pour les lemmes et les étiquettes morphosyntaxiques, il n'y a toujours aucune différence entre les conditions quant à la fréquence des différentes longueurs de *n*-grammes.

35 La deuxième analyse porte sur les pourcentages d'indices communs à chacune des paires de conditions possibles. Comme le montre le tableau 6, le plus grand pourcentage est observé pour la paire extraite de « Wiki » et les plus faibles sont clairement observés lorsque la paire de conditions inclut « Roman ». Les pourcentages pour cette condition sont très faibles puisqu'ils incluent les 6 % qui sont communs aux cinq collections. Il apparaît aussi que les intertitres du « *Soir* » sont plus similaires aux paragraphes de « Wiki » qu'aux sous-titres de cette même collection.

36 Il est toutefois important de nuancer ces pourcentages qui sont basés sur les 200 indices les plus utiles dans chaque condition. Un indice sélectionné selon ce critère dans plusieurs conditions peut ne pas l'être dans une autre condition, mais être néanmoins utile, étant rangé un peu après la 200e place.

	Soir I	Wiki P	Soir P	Roman P
Wiki S	31	50	25	10
Soir I		38	37	14
Wiki P			36	13
Soir P				13

Tableau 6 – Pourcentages d'indices communs par paires de conditions

37 La suite de cette section analyse d'une manière plus qualitative les indices sélectionnés. Comme indiqué dans le tableau 6, les indices sélectionnés dans les deux conditions basées sur « Wiki » sont dans 50 % des cas identiques. Il s'agit pour une large part de variantes de ceux qui sont communs aux cinq conditions, mais on y trouve aussi des indices qui peuvent avoir une fonction cadrative comme *sur_le_plan*, qui signale une rupture. Parmi les indices spécifiques aux sections, on trouve principalement toute une série de *n*-grammes liés à la naissance d'un individu ou d'une organisation/institution/entreprise : *naître_le_@card@*[6], *de_un_famille*, *être_fonder_en*, *être_le_fils*. Ces *n*-grammes, qui signalent une rupture, ne sont pas nécessairement situés en début de phrase. Pour les paragraphes, les indices spécifiques sont surtout des prépositions en début de phrase, qui signalent une rupture (*après*, *dans*, *en*, *pour*, *au*...), et des conjonctions et des adverbes, qui signalent la continuité (*et*, *or*, *puis*, *ainsi*, *en effet*, *par exemple*, *de plus*...).

38 Il y a nettement moins d'indices communs dans les conditions « Section » pour « Wiki » et « Intertitre » pour « *Le Soir* ». Il s'agit dans 75 % des cas de *n*-grammes d'étiquettes lexicosyntaxiques incluant principalement des noms propres signalant une rupture ou des pronoms signalant la continuité. Les indices spécifiques aux sections de « Wiki » ont déjà été mentionnés dans le paragraphe précédent. Parmi les indices spécifiques aux phrases qui suivent les intertitres du « *Soir* », les plus typiques sont des *n*-grammes de lemmes qui ne commencent pas la phrase et qui sont liés au contenu des articles comme *Belgique*, *Laurent* (le prénom d'un prince belge), *belge*, *bruxellois*, *contemporain*, *criminel*, *crise*, et *mi-temps*.

39 Les trois conditions les plus intéressantes à comparer sont celles basées sur les paragraphes. La figure 2 présente les nombres d'indices communs et spécifiques aux trois conditions « Paragraphe » sur un total par collection de 200. Elle confirme la forte différence entre « Roman » et les deux autres collections.

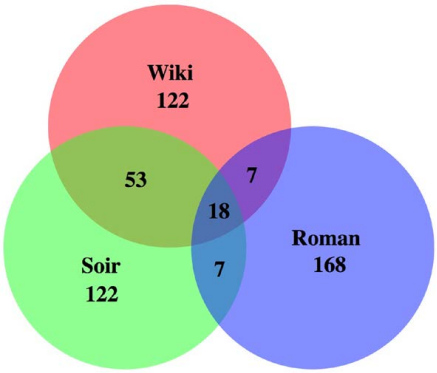

Figure 2 – Nombre d'indices communs et spécifiques aux trois conditions « Paragraphe »

6. Dans les sorties du TreeTagger, « @card@ » remplace un nombre.

40 Peu d'indices communs aux trois conditions s'ajoutent aux douze qui sont communs aux cinq conditions. Pour les lemmes, il s'agit uniquement de signaux de continuité : *son* et *il_se* en début de phrase et *d'ailleurs* et *même* dans la suite. Pour les étiquettes morphosyntaxiques, il s'agit de signaux de rupture, un bigramme incluant un nom propre, mais aussi la séquence *PRP_DET :ART_NOM* en début de phrases, un pattern qui s'accorde avec nombre d'adverbiaux dont la fonction de marqueur de la segmentation est bien établie.

41 Il n'est pas possible d'analyser ici dans le détail l'ensemble des indices sélectionnés dans une ou deux conditions. Un sous-ensemble de ceux-ci est néanmoins donné dans le tableau 7. Il rassemble tous les lemmes présents en début de phrases qui sont sélectionnés dans au moins une des trois conditions en indiquant s'ils signalent une rupture (R) ou une situation de continuité (C) dans chacune de celles-ci, l'absence de lettre signifiant que l'indice en question n'a pas été sélectionné dans cette collection. Les indices sont rangés selon l'ordre alphabétique avec l'exception que certains ont été déplacés de quelques lignes afin de rendre les similarités plus manifestes.

42 Comme le montre le tableau 7, nombre de bigrammes et de trigrammes sélectionnés sont des extensions d'unigrammes également sélectionnés, mais pas toujours dans la même collection et pas toujours avec la même orientation ; par exemple, *un* est un signal de continuité dans « Wiki » et « *Le Soir* », mais *un_autre* est un signal de rupture dans ces deux mêmes collections. On observe aussi quelques cas de désaccord quant à l'orientation d'un indice entre deux collections ; par exemple, *en_effet* est un signal de continuité dans « Wiki » et de rupture dans « Roman ». Comme souligné plus haut, c'est « Roman » qui se distingue le plus des deux autres collections, tout particulièrement par des signaux de rupture. Le tableau 7 montre qu'il s'agit dans une large mesure d'expressions temporelles antéposées (*le_lendemain*, *le_soir*, *un_heure*, *un_jour*, *un_soir*, *pendant_que*, *à_ce_moment*). Il en est de même de *vers* qui est presque toujours employé pour introduire une expression temporelle, un seul cas spatial ayant été identifié dans les textes (« Vers l'orient, l'horizon pâlissait », *Une vie* de Guy de Maupassant). On notera aussi que des adverbiaux temporels sont aussi sélectionnés dans « Wiki » (*en_@card@_*, ou *@card@* remplace une année par exemple), mais en moins grand nombre.

43 Plus généralement, les indices sélectionnés sont pour la plupart interprétables dans le cadre des travaux linguistiques sur les marques de segmentation et de liage (p. ex., pronoms personnels, déterminants démonstratifs, connecteurs, adverbiaux temporels, organisateurs textuels). On note néanmoins quelques exceptions liées au contenu ou au style propre à chaque collection. Les plus manifestes sont issues de « Roman » : *Étienne*, *madame* et *Saccard*. Dans « *Le Soir* », on trouve des formulations relativement typiques du style journalistique comme *rester_que*, *rester_le*, *du_côté*, *que_faire*, *autre* et *côté* qui signalent une rupture et *même_chose*, *soit*, et *ici* qui signalent une situation de continuité. Dans « Wiki », on peut citer *le_saison* qui trouve son origine dans toute une série de paragraphes qui commencent par *La saison sèche*, *La saison 1983*, *La saison de terre battue*, *La saison des amours*…

mo	W	S	R
@card@			R
@card@_jour			R
à			R
à_ce			R
à_ce_moment			R
à_le			R
à_le_origine		R	
ainsi	C	C	
ainsi_,	C		
alors			R
alors_,			R
après	R		
après_ce	R		
après_le	R		
après_@card@			R
au	R		
autre		R	
ce	C		
ce_dernier	C	C	
ce_être	C		
ce_être_le	C		
ce_être_comme			C
ceci	C		R
cela	C		C
celui	C	C	
celui_-ci	C		
cependant			R
cependant_,			R

mo	W	S	R
il	C	C	C
il_avoir	C	C	
il_être	C		
il_exister	R		
il_se	C	C	C
il_se_trouver			C
le	R	R	
le_@card@	R		
le_saison	R		
le_lendemain			R
le_soir_,			R
leur	C	C	
madame			R
Maheu			R
mais	C	C	
mais_aussi		C	
mais_il		C	
même		C	
même_chose		C	
notamment		C	
nous		C	
nous_nous			C
on_avoir_dire			C
on_le			C
or	C	C	
or_,		C	
or_,_le		C	
ou		C	

cependant_,_le			R
cependant_le			R
côté		R	
de_plus	C	C	
de_plus_,	C		
début	R		
dès			R
dès_le			R
du_côté		R	
elle	C	C	C
elle_avoir	C	C	
elle_être	C		
elle_se	C	C	
en	R		
en_@card@	R		
en_@card@_,	R		
en_effet	C		R
en_effet_,	C		R
enfin		R	R
enfin_,		R	R
et	C	C	
et_,		C	
et_le		C	
Étienne			R
Étienne_,			R
ici		C	

par_exemple	C		
par_exemple_,	C		
pendant			R
pendant_que			R
plus		R	
pour		R	
puis	C	C	
puis_,	C		
que_faire		R	
rester		R	
rester_le		R	
rester_que		R	
Saccard			R
seul	C		
si		R	
si_le		R	
soit		C	
son	C	C	C
sur_le_plan	R		
un	C	C	
un_autre	R	R	
un_heure			R
un_jour			R
un_jour_,			R
un_soir			R
vers			R

Tableau 7 – Lemmes présents en début de phrases
sélectionnés dans au moins une des trois conditions « Paragraphe »
(W = « Wiki » ; S = « *Le Soir* » ; R = « Roman »)

5. Discussion et conclusion

44 La présente étude a pour objectif de proposer et d'évaluer une procédure, fondée sur des techniques de TAL, qui permet l'extraction d'indices de segmentation et de liage dans de très grandes collections de textes de genre différent. Les analyses effectuées sur trois collections de textes ont montré que cette procédure est relativement efficace pour distinguer les phrases en situation de rupture de celles en situation de continuité puisqu'elle atteint une exactitude de 64 % à 74 %. De plus, cette procédure met en évidence des indices comme des pronoms personnels, des connecteurs, des adverbiaux, des déterminants, qui interviennent dans les anaphores définies, démonstratives et possessives, dont la fonction de signaux de (dis)continuité a été soulignée dans des études linguistiques antérieures, basées sur des analyses manuelles approfondies. Les analyses ont aussi montré que les sections de « Wiki » étaient les plus facilement identifiables et confirmé le statut particulier des intertitres de journaux. Lorsque les discontinuités sont marquées par un alinéa, l'efficacité est équivalente dans les trois collections. En revanche, les indices les plus importants sont majoritairement différents. Le modèle prédictif pour le corpus littéraire privilégie les indices associés aux ruptures (p. ex., *cependant*, *le_lendemain*, *à_ce_moment* en début de phrase) au détriment de ceux associés à la continuité (*il*, *son*, *ainsi* en début de phrase) et ce contrairement à ce qui s'observe dans les deux autres corpus.

45 Les résultats obtenus, même s'ils semblent encourageants, ne pourront prendre tout leur sens qu'après une analyse linguistique détaillée qui va au-delà des objectifs de la présente synthèse. On peut néanmoins penser que les indices recueillis, y compris ceux qui apparaissent au-delà de la 200^{e} position, pourraient se révéler bénéfiques pour plusieurs applications. La possibilité d'identifier automatiquement un grand nombre d'indices plus ou moins spécifiques à certains genres de textes devrait être profitable aux techniques de segmentation de textes. Cela devrait également être le cas pour l'évaluation de la qualité rédactionnelle de textes produits par des apprenants, un domaine où l'organisation du texte et son marquage linguistique sont souvent négligés en raison des difficultés que leur prise en compte pose (Deane et Quinlan, 2010 ; Yannakoudakis et Briscoe, 2012).

46 Après une discussion des principales limites de l'étude, cette section se conclut par la présentation de plusieurs pistes de recherche susceptibles d'améliorer et d'étendre l'approche proposée.

5.1. Limitations

47 Les principales limitations de cette étude découlent directement de l'approche employée qui implique une analyse très largement automatisée de grandes collections de textes. Toutes les analyses effectuées reposent sur les sorties du TreeTagger, outil couramment employé en TAL. Il segmente, identifie les frontières des phrases et génère les étiquettes morphosyntaxiques. Il a été nécessaire de prétraiter les textes pour rendre ces opérations plus correctes. À défaut, une séquence aussi étrange que *pour le Riesling* aurait été un marqueur de continuité important dans « Wiki » parce

que toute une série d'articles évoquent les vins d'Alsace en indiquant par exemple «un titre alcoométrique volumique naturel moyen minimum de 12,5 % vol. pour les cépages…» et que le point qui suit «vol» est considéré par le TreeTagger comme un signal de fin de phrase qui évidemment ne se produit jamais au début d'un paragraphe. De même, il a été nécessaire de corriger un certain nombre d'étiquettes morphosyntaxiques[7]. Il n'est évidemment pas possible de garantir qu'il ne reste pas des scories qui ont pu influencer les résultats. Il ne fait aucun doute que ces difficultés auraient été bien moins nombreuses si on avait employé plusieurs outils d'étiquetage morphosyntaxique et fusionné leurs sorties.

48 Les résultats obtenus ont aussi été nécessairement affectés par une série de choix comme celui d'éliminer les documents très courts ou très longs ou encore contenant une proportion importante de non-mots, d'éliminer les phrases les plus courtes et les plus longues, de focaliser l'ensemble des analyses sur des quadruplets de phrases contiguës, de ne pas considérer les *n*-grammes d'une longueur supérieure à trois et d'analyser les lemmes plutôt que les formes graphiques. Il me semble que chacun de ces choix (même si les valeurs sélectionnées sont par définition nécessairement arbitraires) peut être justifié. Il est probable que celui qui a le plus affecté les résultats est la décision d'analyser des quadruplets de phrases. Considérer des séquences plus longues aurait probablement permis d'obtenir des situations de continuité et de rupture plus extrêmes, mais aurait également réduit le nombre de phrases analysables, nombre déjà relativement faible dans la plus petite des trois collections de textes. Une autre décision qui a certainement eu un impact sur les résultats est le choix de la régression logistique comme procédure d'apprentissage supervisé. D'autres possibilités, si elles peuvent être appliquées à la tâche de catégorisation employée ici, auraient pu être évaluées comme les *Conditional Random Fields* (CRF), XGBoost ou une procédure d'apprentissage profond. Plus généralement, combiner les résultats de plusieurs classifieurs pourrait se révéler très efficace pour identifier les indices les plus généraux.

49 Il est également aussi nécessaire de mentionner parmi les limitations de cette étude, les collections de textes analysées. Si elles représentent des genres et des domaines très différents, elles se distinguent également sur toute une série d'autres caractéristiques qui ont pu affecter les résultats. Parmi les plus importantes, on doit citer la taille des collections de textes. Ce facteur a pu affecter l'efficacité de la procédure d'apprentissage supervisé et, le cas échéant, son impact ne peut pas être distingué de celui des autres différences, celles-là souhaitées, entre les trois collections. Cette différence de taille pose également problème lorsqu'on essaie de combiner les trois corpus en un seul matériel d'apprentissage. C'est particulièrement regrettable parce qu'une telle approche aurait sans doute permis d'identifier les indices les plus généraux et de contrecarrer la tendance des approches discriminantes à s'appuyer sur les caractéristiques idiosyncrasiques de la collection de textes qui leur sont fournies.

7. En plus des problèmes détectés lors de la vérification manuelle partielle des sorties, cela a été principalement effectué sur la base des trigrammes d'étiquettes morphosyntaxiques relativement fréquentes et difficilement compréhensibles comme celles produites par l'étiquetage de *cette* comme un PRO:DEM.

5.2. Pistes de recherche

50 Parmi les travaux futurs, il serait intéressant de demander à des juges humains d'effectuer la tâche de catégorisation sur un échantillon aléatoire du matériel employé dans la présente étude afin d'estimer un niveau supérieur auquel comparer les performances de la procédure automatique (Sporleder et Lapata, 2006). Il s'agirait donc de leur présenter des quadruplets de phrases et leur tâche serait de décider s'il y a ou non un changement de paragraphe entre la deuxième et la troisième phrase. On sait depuis les travaux de Stark (1988) que retrouver les paragraphes d'un texte entier est une tâche très complexe pour un humain, mais il s'agit là sans doute d'une tâche nettement plus complexe que celle proposée ici.

51 Il serait aussi intéressant d'affiner l'extraction des indices. Par exemple, on a observé qu'il y avait très peu d'adverbiaux spatiaux parmi les indices les plus utiles pour la discrimination. Toutefois, parmi les indices sélectionnés au-delà de la 2000^{e} place pour les paragraphes dans « Wiki », on trouve en début de phrase et signalant une rupture *en_France*, *en_Belgique*, *en_Angleterre*, *en_Suisse* et *en_Espagne*. Pouvoir extraire des textes des motifs comme *en + Nom de pays* ou *en + localisation géographique* permettrait peut-être d'obtenir des indices plus efficaces. Des essais effectués lors du développement de l'approche en employant des *n*-grammes contenant des jokers (« *skipgrams* ») n'ont pas produit de résultats intéressants. Les résultats auraient probablement été plus positifs si des *n*-grammes combinant des lemmes et des étiquettes morphosyntaxiques avaient été employés.

52 Le développement le plus important devrait permettre de prendre en compte d'une manière plus complète les indices liés à la coréférence. Ces indices se sont révélés particulièrement utiles pour discriminer les phrases en début ou non de segments comme l'atteste leur présence parmi les indices les plus importants dans plusieurs ensembles de données. En effet, plus de 75 % des indices sélectionnés dans au moins quatre des cinq ensembles de données relèvent de cette catégorie. Dans les analyses présentées ci-dessus, ces indices sont traités d'une manière isolée (présence ou absence dans une phrase donnée) alors qu'il s'agit par définition d'indices qui peuvent établir des liens avec d'autres phrases que ce soit en amont (les pronoms personnels par exemple) ou en aval (les noms propres par exemple) ou même dans les deux directions. Prendre en compte ces liens, comme cela se fait lorsqu'on analyse les chaînes de référence (Schnedecker et Landragin, 2014), devrait se montrer particulièrement fructueux. Si une procédure automatique, capable d'identifier les chaînes de référence dans des textes non annotés, n'est pas encore à l'ordre du jour, une approche plus rudimentaire pourrait être évaluée. Barzilay et Lapata (2008) ont proposé de suivre la distribution des entités mentionnées dans les phrases d'un texte, en prenant en compte leur rôle syntaxique, et de la représenter par une grille d'entités. Ces grilles sont construites par une procédure automatique au moyen d'outils comme un analyseur syntaxique et un résolveur de coréférence. Elles ont été utilisées pour caractériser les transitions dans des textes (quel type de transition entre entités ? dans quelle proportion ?) et ces informations

ont permis par exemple d'estimer le degré de lisibilité d'un texte. Il devrait être possible d'employer cette approche pour extraire des indices supplémentaires afin de caractériser les phrases à catégoriser. Elle pourra être développée et évaluée sur le corpus en langue française annoté en expressions référentielles et chaînes de référence, issu du projet de recherche DEMOCRAT (Landragin, 2016), qui est téléchargeable à l'adresse suivante : https://www.ortolang.fr/market/corpora/democrat.

Références bibliographiques

Adam, J.-M. 1990. *Éléments de linguistique textuelle : théorie et pratique de l'analyse textuelle*. Liège : Mardaga.

Adam, J.-M. 1997. Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite. *Pratiques* 94 : 3-18.

Adam, J.-M. 2015. Le paragraphe : unité transphrastique et palier d'analyse textuelle. In *3e colloque de l'Association internationale de stylistique : « Méthodes stylistiques. Unités et paliers de pertinence textuelle ? » (Lyon, 31 mars-2 avril 2015)*. 1-25. En ligne à l'adresse suivante : http://www.styl-m.org/wp-content/uploads/2015/03/ADAM_Paragraphe.pdf.

Adam, J.-M. 2016. De la grammaire de texte à la cohérence discursive : un parcours exemplaire. In L. Sarda, D. Vigier et B. Combettes (éd.), *Connexion et indexation. Ces liens qui tissent le texte*. Lyon : ENS Éditions : 55-68.

Adam, J.-M. 2018. *Le paragraphe : entre phrases et texte*. Paris : A. Colin.

Adam, J.-M. et Lugrin, G. 2000. L'hyperstructure : un mode privilégié de présentation des événements scientifiques ? *Les Carnets du Cediscor* 6 : 133-149. En ligne à l'adresse suivante : https://journals.openedition.org/cediscor/327.

Anderson, A., Garrod, S. C. et Sanford, A. J. 1983. The Accessibility of Pronominal Antecedents as a Function of Episode Shifts in Narrative Text. *The Quarterly Journal of Experimental Psychology Section A* 35 (3) : 427-440.

Barzilay, R. et Lapata, M. 2008. Modeling Local Coherence : An Entity-Based Approach. *Computational Linguistics* 34 (1) : 1-34. En ligne à l'adresse suivante : http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.1.1.

Bessonnat, B. 1988. Le découpage en paragraphes et ses fonctions. *Pratiques* 57 : 81-105.

Bestgen, Y. et Vonk, W. 2000. Temporal Adverbials as Segmentation Markers in Discourse Comprehension. *Journal of Memory and Language* 42 (1) : 74-87.

Chafe, W. 1984. How People Use Adverbial Clauses. In C. Brugman et M. Macaulay (éd.), *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society*. Berkeley : Berkeley Linguistic Society : 437-449.

Charolles, M. 1988 Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques* 57 : 3-13.

Charolles, M. 1993. Les plans d'organisation du discours et leurs interactions. In S. Moirand, A. Ali Bouacha, J.-C. Beacco et A. Collinot (éd.), *Parcours linguistiques de discours spécialisés*. Berne – Berlin – Paris : P. Lang : 301-314.

Charolles, M. 2007. *Un jour* (One Day) in Narratives. In I. Korzen et L. Lundquist (éd.), *Comparing Anaphors. Between Sentences, Texts and Languages*. Copenhague : Samfundslitteratur Press : 11-26.

Charolles, M., Le Draoulec, A., Péry-Woodley, M.-P. et Sarda, L. 2005. Temporal and Spatial Dimensions of Discourse Organisation. *Journal of French Language Studies* 15 (2) : 115-130.

Colonna, S., Charolles, M., Sarda, L. et Pynte, J. 2014. Effect on Comprehension of Preposed versus Postposed Adverbial Phrases. *Journal of Psycholinguistic Research* 43 (6) : 771-790.

Crompton, P. 2006. The Effect of Position on the Discourse Scope of Adverbials. *Text and Talk* 26 (3) : 245-279.

Deane, P. et Quinlan, T. 2010. What Automated Analyses of Corpora Can Tell Us about Students' Writing Skills. *Journal of Writing Research* 2 (2) : 151-177. En ligne à l'adresse suivante : http://www.jowr.org/Ccount/click.php?id=24.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. et Lin, C.-J. 2008. LIBLINEAR : A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 : 1871-1874. En ligne à l'adresse suivante : http://www.jmlr.org/papers/volume9/fan08a/fan08a.pdf.

Ferret, O., Grau, B., Minel, J.-L. et Porhiel, S. 2001. Repérage de structures thématiques dans des textes. In *TALN – RECITAL 2001 : 8e conférence annuelle sur le Traitement automatique des langues naturelles (2-5 juillet 2001, Tours)*. Paris : Association pour le Traitement automatique des langues (ATALA) : 163-172. En ligne à l'adresse suivante : https://www.atala.org/sites/default/files/actes_taln/AC_0036.pdf.

Filippova, K. et Strube, M. 2006. Using Linguistically Motivated Features for Paragraph Boundary Identification. In D. Jurafsky et É. Gaussier (éd.), *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing – EMNLP 2006*. Stroudsburg : Association for Computational Linguistics (ACL) : 267-274. En ligne à l'adresse suivante : https://www.aclweb.org/anthology/W06-1632.pdf.

Garcia-Debanc, C. 2010. Segmentation, connexion et indexation dans des productions écrites d'élèves de 9 à 13 ans de deux genres textuels. *Synergies Pays Scandinaves* 5 : 81-96. En ligne à l'adresse suivante : https://gerflint.fr/Base/Paysscandinaves5/claudine.pdf.

Grimes, J. E. 1975. *The Thread of Discourse*. La Haye – Paris : Mouton.

Hansen, M.-B. M. 1997. *Alors* and *Donc* in Spoken French : A Reanalysis. *Journal of Pragmatics* 28 (2) : 153-187.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J. et Tanguy, L. 2012. An Empirical Approach to the Signalling of Enumerative Structures. *Discours* 10 : 1-27. En ligne à l'adresse suivante : https://journals.openedition.org/discours/8611.

Ho-Dac, L.-M. et Péry-Woodley, M.-P. 2009. A Data-Driven Study of Temporal Adverbials as Discourse Segmentation Markers. *Discours* 4 : 1-20. En ligne à l'adresse suivante : https://journals.openedition.org/discours/5952.

Hofmann, T. R. 1989. Paragraphs, and Anaphora. *Journal of Pragmatics* 13 (2) : 239-250.

Howell, D. 2008. *Méthodes statistiques en sciences humaines*. V. Yzerbyt et Y. Bestgen (éd.) ; M. Rogier (trad.). Bruxelles : De Boeck.

Jacques, M.-P. 2005. Structure matérielle et contenu sémantique du texte écrit. *Corela* 3 (2) : 1-27. En ligne à l'adresse suivante : https://journals.openedition.org/corela/560.

Jacques, M.-P. et Poibeau, T. 2010. Étudier des structures de discours : préoccupations pratiques et méthodologiques. *Corela* 8 (2) : 1-22. En ligne à l'adresse suivante : https://journals.openedition.org/corela/1855.

Jurafsky, D. et Martin, J. H. 2009. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River : Pearson Prentice Hall [2e éd.].

Lai, C., Farrús, M. et Moore, J. D. 2016. Automatic Paragraph Segmentation with Lexical and Prosodic Features. In *Proceedings of Interspeech 2016*. Baixas : International Speech Communication Association (ISCA) : 1034-1038. En ligne à l'adresse suivante : https://www.isca-speech.org/archive/Interspeech_2016/pdfs/0992.PDF.

Landragin, F. 2016. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association française pour l'Intelligence Artificielle* 92 : 11-15.

Le Draoulec, A. et Péry-Woodley, M.-P. 2005. Encadrement temporel et relations de discours. *Langue française* 148 : 45-60.

Lemarié, J., Lorch, R. F. Jr. et Péry-Woodley, M.-P. 2012. Understanding How Headings Influence Text Processing. *Discours* 10 : 1-22. En ligne à l'adresse suivante : https://journals.openedition.org/discours/8600.

Longacre, R. E. 1979. The Paragraph as a Grammatical Unit. In J. P. Kimball et T. Givón (éd.), *Syntax and Semantics*. New York : Academic Press. Vol. 12 : *Discourse and Syntax* : 116-134.

Marcu, D. 2000. The Rhetorical Parsing of Unrestricted Texts : A Surface-Based Approach. *Computational Linguistics* 26 (3) : 395-448. En ligne à l'adresse suivante : https://www.aclweb.org/anthology/J00-3005.pdf.

Passonneau, R. J. et Litman, D. J. 1997. Discourse Segmentation by Human and Automated Means. *Computational Linguistics* 23 (1) : 103-139. En ligne à l'adresse suivante : https://www.aclweb.org/anthology/J97-1005.pdf.

Péry-Woodley, M.-P., Ho-Dac, L.-M., Rebeyrolle, J., Tanguy, L. et Fabre, C. 2017. A Corpus-Driven Approach to Discourse Organisation : From Cues to Complex Markers. *Dialogue and Discourse* 8 (1) : 66-105.

Piérard, S. et Bestgen, Y. 2006. Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL* 47 (2) : 89-110. En ligne à l'adresse suivante : https://www.atala.org/sites/default/files/TAL-2006-47-2-05-Pierard.pdf.

Pierard, S. et Bestgen, Y. 2008. Use of Temporal Adverbials as Segmentation Discourse Markers by Second Language Learners. *Archives de Psychologie* 73 : 209-230.

Rebeyrolle, J., Jacques, M.-P. et Péry-Woodley, M.-P. 2009. Titres et intertitres dans l'organisation du discours. *Journal of French Language Studies* 19 (2) : 269-290.

Sarda, L., Carter-Thomas, S., Fagard, B. et Charolles, M. 2014. Adverbials : From Predicative to Discourse Functions. In L. Sarda, S. Carter-Thomas, B. Fagard et M. Charolles (éd.), *Adverbials in Use : From Predicative to Discourse Functions*. Louvain-la-Neuve : Presses universitaires de Louvain : 17-34.

Schmid, H. 2003. Probabilistic Part-of-Speech Tagging Using Decision Trees. In H. L. Somers et D. B. Jones (éd.), *New Methods in Language Processing*. Londres : Routledge : 154-164.

Schnedecker, C. 1997. *Nom propre et chaînes de référence*. Paris – Metz : Klincksieck – Université de Metz.

Schnedecker, C. et Landragin, F. 2014. Les chaînes de référence : présentation. *Langages* 195 : 3-22.

Schrepfer-André, G. 2005. La portée phrastique et textuelle des expressions introductrices de cadres énonciatifs : les syntagmes prépositionnels en *selon X*. Thèse de doctorat. Université Sorbonne Nouvelle – Paris 3.

Sporleder, C. et Lapata, M. 2006. Broad Coverage Paragraph Segmentation across Languages and Domains. *ACM Transactions on Speech and Language Processing* 3 (2) : 1-35.

Stark, H. A. 1988. What Do Paragraph Markings Do ? *Discourse Processes* 11 (3) : 275-303.

Stosic, D. 2012. Le pouvoir cadratif des compléments introduits par *à travers* : des cadres de discours pas comme les autres ? *Travaux de linguistique* 64 : 55-78.

Védénina, L. G. 1989. *Pertinence linguistique de la présentation typographique*. Louvain : Peeters.

Vigier, D. 2005. Les adverbiaux praxéologiques détachés en position initiale et leur portée. *Verbum* 27 (3) : 293-312.

Vigier, D. 2008. La gestion des cadres de discours dans une tâche rédactionnelle en FLE. In H. Hilton (éd.), *Acquisition et didactique 1. Actes de l'atelier didactique, AFLS 2005*. Chambéry : Université de Savoie : 113-129.

Virtanen, T. 1992. *Discourse Functions of Adverbial Placement in English : Clause-Initial Adverbials of Time and Place in Narratives and Procedural Place Descriptions*. Åbo : Åbo Akademis Förlag.

Yannakoudakis, H. et Briscoe, T. 2012. Modeling Coherence in ESOL Learner Texts. In J. Tetreault, J. Burstein et C. Leacock (éd.), *Proceedings of the Seventh Workshop on the Innovative Use of NLP for Building Educational Applications – NAACL-HLT 2012*. Stroudsburg : Association for Computational Linguistics (ACL) : 33-43. En ligne à l'adresse suivante : https://www.aclweb.org/anthology/W12-2004.pdf.