Le repérage de nominations dans les corpus textuels: de l'exploitation de l'analyse des données textuelles à l'exploration des chaînes de coréférence par le TAL

Julien Longhi, Jean-Yves Antoine, Mehdi Mirzapour, Agata Jackiewicz, Anaïs Lefeuvre-Halftermeyer

Le repérage de nominations dans les corpus textuels: de l'exploitation de l'analyse des données textuelles à l'exploration des chaînes de coréférence par le TAL

Julien Longhi[1], Jean-Yves Antoine[2], Mehdi Mirzapour[2], Agata Jackiewicz[3]Anaïs Lefeuvre-Halftermeyer[4]

[1]AGORA/IDHN, CY Cergy Paris université, IUF – Julien.Longhi@u-cergy.fr

[2]LIFAT, Université de Tours – {Jean-Yves.Antoine, Mehdi.Mirzapour}@univ-tours.fr

[3]Praxiling, Université Paul-Valéry, Agata.Jackiewicz@univ-montp3.fr

[4]LIFO, Université d'Orléans, anais.halftermeyer@univ-orleans.fr

Abstract 1

This article focuses on the specific category of nominations, a lexical process that allows the speaker to lead the building of a referent and to engage his/her audience towards the argumentative orientation of his/her choice. We present three strategies of nomination detection that make an increasing use of digital tools: manual exploration, textometric approach, and coreference resolution.

Keywords: nomination, coreference, textometry, discourse analysis.

Abstract 2

Cet article s'intéresse à la catégorie des nominations, un procédé lexical qui permet au locuteur de forger la construction d'un référent et d'engager son allocutaire vers l'orientation argumentative de son choix. Nous présentons trois stratégies de repérage des nominations qui font un recours croissant aux outils numériques : exploration manuelle, textométrique, et résolution des coréférences.

Mots clés : nomination, coréférence, textométrie, analyse du discours.

1. Introduction

Dans le cadre du projet ANR TALAD (*Traitement Automatique des Langues et Analyse du Discours*), nous nous intéressons à la catégorie spécifique des « nominations » (Jackiewicz et al. 2019). Le terme « nomination » renvoie à l'acte d'attribution d'un nom à une entité et ainsi qu'au résultat de cet acte. La nomination est une opération linguistique et cognitive, indissociable des processus d'appréhension, de catégorisation et de configuration des réalités. Elle possède une dimension discursive et dialogique, car l'expression choisie pour nommer un référent reflète la position que le sujet parlant adopte à son égard. Les nominations s'inscrivent plus largement dans une dynamique des relations sociales et révèlent des représentations que les locuteurs construisent, négocient et font circuler. Elles permettent au locuteur de forger la construction d'un référent – posant ainsi une coloration axiologique et un réseau associatif/stéréotypique - et d'engager d'emblée son allocutaire vers l'orientation argumentative de son choix. Ceci a été particulièrement saillant dans les récentes productions, politiques et médiatiques, autour des termes *migrants*/ *immigrants*/ *réfugiés*/ *déracinés*. Notons que ces termes peuvent constituer des entités polylexicales telles que *demandeurs*

d'asile / *candidats à l'asile* mais aussi *réfugiés climatiques*/ *migrants économiques*. On remarque que ces nominations entrent également dans des constructions régulières qui permettent de renforcer ou préciser la coloration comme avec les syntagmes *réfugiés et migrants* / *migrants mineurs* / *migrants désespérés*. L'enjeu général du projet est notamment que le Traitement Automatique des Langues (TAL par la suite) fournisse une aide exploratoire au chercheur en analyse du discours (AD) pour différents types de tâches. Par exemple, partant d'une forme de nomination, les outils TAL élaborés devraient proposer, sous forme de concordance, toutes les unités lexicales qui sont susceptibles d'être une reformulation de la nomination. Parmi ces formes, certaines seront originales et permettront au chercheur d'aller plus loin dans son étude des variations de nomination.

2. Corpus, méthodes, et hypothèses

Ce travail, qui peut concerner tous les domaines d'intervention de l'AD, s'applique, dans le cadre spécifique du projet, à un corpus d'interviews politiques de matinales. Sur de telles données, trois types de méthodologie d'exploration des nominations en corpus, qui font respectivement appel à un recours croissant au TAL, sont abordées :

A) En référence avec les travaux initiaux sur la nomination, une première méthodologie issue de l'analyse du discours de tradition française (Longhi 2015) consiste à constituer des corpus spécifiques en lien avec une problématique spécifique, et à chercher manuellement les nominations existantes sur un même sujet (voir les travaux de Siblot, Moirand, etc.). Si des relevés très précis peuvent être effectués, cette approche a deux écueils : limitation au corpus initialement constitué, et recours à une méthodologie descriptiviste qui ne permet pas d'envisager une systématisation des repérages effectués.

B) Dans la lignée de ces travaux, mais avec une approche plus « outillée », on peut avoir recours à des outils de textométrie, afin de procéder à des explorations de corpus (Longhi et Salem 2018). Si cela permet de traiter des corpus de plus grande envergure, la mise en discours de nominations s'incarne dans une grande variété de formes linguistiques (néologismes, unités polylexicales, termes qui se créditent d'un sens nouveau), et la textométrie ne peut consister qu'en une aide à ce repérage, à travers l'usage de différentes fonctionnalités, et le prélèvement ponctuel d'unités ou de contextes à l'intérieur de contextes plus larges. Ainsi, cet article présente dans une première partie une méthodologie d'analyse qui a été établie, dans le but de saisir les nominations spécifiques à un objet discursif précis : 1) filtre dans tout le corpus de l'ensemble des textes qui contiennent un terme initialement ciblé (repéré comme le plus utilisé ou le premier employé, dans un contexte donné) ; 2) création d'un sous-corpus avec le concordancier ciblé sur le terme, avec l'hypothèse que dans l'environnement du terme on aura des nominations alternatives ; 3) élaboration d'une classification hiérarchique descendante (CHD) qui donne des classes/mondes lexicaux, et qui permet de voir effectivement des alternatives au terme, et les lier à des domaines/thèmes.

C) L'approche précédente se base sur les propriétés distributionnelles de cooccurrences lexicales. Le recours au TAL peut explorer des pistes alternatives et complémentaires. Ainsi, l'utilisation de plongement de mots nous permet de détecter certaines nominations dont la teneur thématique ou axiologique est atypique en termes de contexte sémantique d'occurrence. Par exemple, *patriotisme économique* mobilisera un champ lexical relevant de l'économie, là où le terme *patriotisme* sera observé dans d'autres contextes thématiques. La stratégie d'exploration que nous présentons dans la dernière partie de ce papier relève quant à elle d'une analyse purement syntaxico-pragmatique. L'idée est en effet d'identifier les chaines

de coréférence dans lesquelles apparait la nomination étudiée, et de rechercher les reprises coréférentielles du terme lesquelles seront souvent des reformulations de cette dernière.

Plus précisément, concernant le corpus, le jeu de données utilisé dans le cadre du projet TALAD contient actuellement 3 166 interviews correspondant à 561 personnalités politiques interviewées, entre le 10 juin 2016 et le 4 décembre 2017 (ce qui couvre notamment la campagne présidentielle 2017). La taille de ce corpus est de 10 millions de mots. Ces données ont été récoltées par la société *Reticular* (http://www.reticularproject.com/), partenaire du projet, grâce à un outil (développé en interne et exploité par l'entreprise dans le cadre d'une application mobile de veille politique) de collecte semi-automatique des interviews/opinions politiques. Parmi le jeu de données fourni, nous avons relevé 466 interviews dans lesquelles « *protectionnisme* » apparaît. Ce terme a semblé intéressant lors de la réflexion sur le choix des thèmes à explorer, car il convoque différentes dimensions (économie, souveraineté, démographie, immigration, etc.) tant sur le plan intérieur qu'extérieur. Nous avons donc constitué un sous-corpus de ces interviews. Ses caractéristiques sont listées dans le tableau 1.

Type de caractéristique	**#**
Nombre d'items	1 914 248
Nombre d'occurrences	1 492 283
Nombre de formes	30 020
Nombre de hapax	11 444

Tableau 1. Caractéristiques du corpus « Protectionnisme ».

Ce sous-corpus a été étudié par des analystes du discours en appliquant les deux méthodes :

1. Analyse linguistique « manuelle » des séquences pertinentes par rapport à l'objet d'étude, à savoir, l'étude de la nomination « *protectionnisme* » : ses réalisations, ses variantes, ses constructions spécifiques.
2. Analyse linguistique outillée du système d'attitudes associé à la nomination « *protectionnisme* » : cadrage, subjectivité et façon dont l'attitude renseigne sur la nomination.

L'objectif de cette analyse est de fournir une étude approfondie (linguistique et discursive) de la nomination « *protectionnisme* » et d'éclairer sur les méthodes et outils pertinents pour l'étude des nominations sujettes à controverses. Par ailleurs, cet article rend compte du développement d'un système de résolution des coréférences qui servira d'outil supplémentaire d'exploration et de repérage semi-automatique des nominations et de leurs variantes.

3. Le traitement « manuel » des nominations : apport et limites

Une première méthode, dans la tradition de l'analyse du discours "à la française", peut s'intéresser à la recherche et analyse manuelle des séquences jugées pertinentes:

A partir de la transcription de l'interview considérée, cette extraction de séquences jugées pertinentes se fait alors par une consultation manuelle des occurrences par le chercheur, qui observera notamment, par la densité de l'apparition des formes, et par les procédés de mise en discours, ce qui pourra se révéler le plus intéressant pour éclairer son objet d'étude. La prise en compte des tours de paroles, des enchaînements, de phénomènes relevant de la pragmatique, par exemple, pourront être pris en compte en plus de l'analyse de la forme

repérée et de son environnement immédiat: le cotexte est enrichi du contexte d'interaction et peut être interprété au regard des certains paramètres complémentaires.

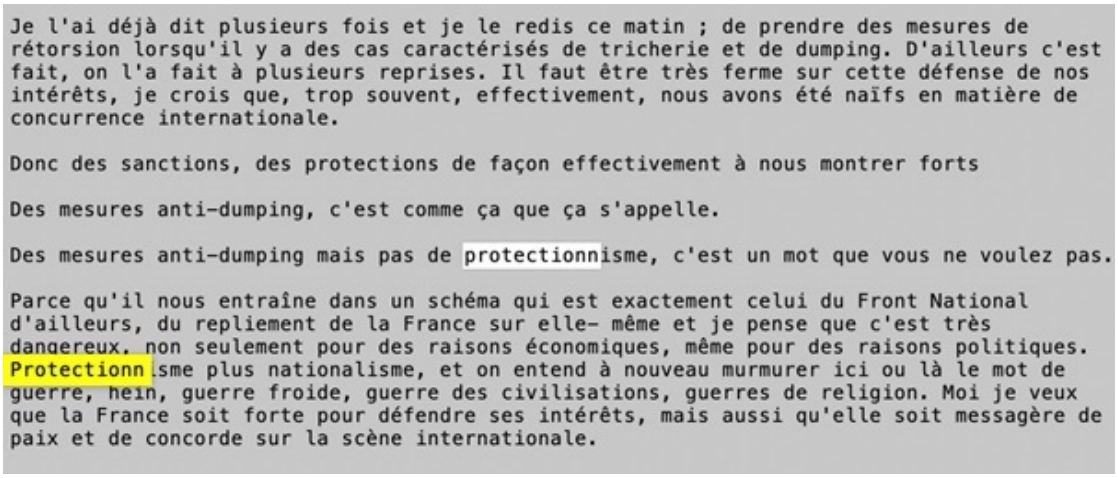

Fig. 1. Recherche simple dans le document (interview d'Alain Juppé)

On repère les enjeux argumentatifs, et le jeu de question/réponse qui induit des précisions et des discussions sur le choix des termes : sur l'exemple de la figure 1, Alain Juppé parle de « mesures de rétorsion » à propos de « tricheries » et de « dumping », puis reformule en « défense des intérêts » ; le journaliste relance en parlant de « sanctions » et de « protections », et Alain Juppé fait œuvre de métadiscours pour écarter les ambiguïtés sémantiques : « Des mesures anti-dumping, c'est comme ça que ça s'appelle ». Le journaliste met en valeur cette négociation dans la nomination : « pas de protectionnisme, c'est un mot que vous ne voulez pas ». La justification de ce choix est faite par Alain Juppé avec un argument intéressant pour l'analyste du discours : ce mot « entraîne dans un schéma qui est exactement celui du Front National ». La négociation sur le choix des mots a donc un double intérêt : « cadrer » le réel en spécifiant le sens de ce qui est décrit (« mesures anti-dumping » présuppose du « dumping », une attitude condamnable qui justifie les mesures, donc une « réponse à »), et « cartographier » les champ politique en lien avec les nominations utilisées, puisque « protectionnisme » est attribué au FN, avec une critique qui l'accompagne. Cette méthode manuelle permet aussi d'observer des phénomènes de dialogisme ou d'intertextualité, comme cela se voit dans l'extrait suivant :

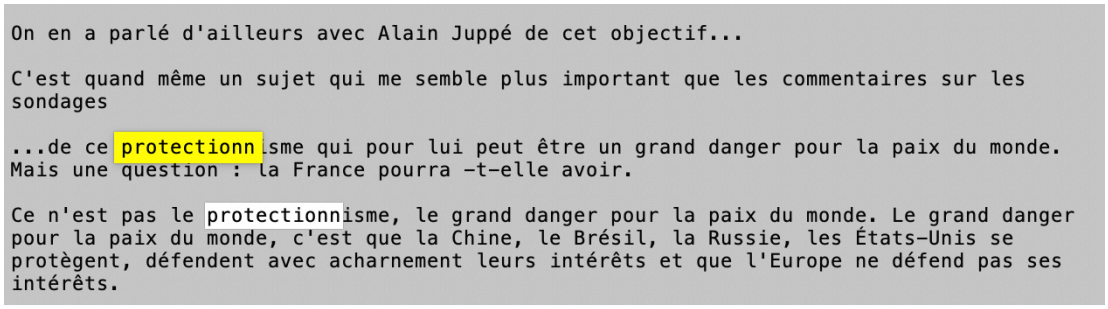

Fig.2 : Extrait d'une interview de Nicolas Sarkozy par Elizabeth Martichoux

Ici, la journaliste fait écho au point de vue d'Alain Juppé sur le protectionnisme (il peut être un grand danger pour la paix du monde), et Nicolas Sarkozy répond avec une structure négative qui contredit le point de vue d'Alain Juppé, puis oppose des pays qui « défendent avec acharnement leurs intérêts » et l'Europe qui « ne défend pas ses intérêts ». On entrevoit ici le « dialogisme de la nomination » (Siblot 2004), avec explicitement la remise en cause de

la portée argumentative du terme (« le protectionnisme n'est pas le grand danger ») et implicitement une négociation du sens (que l'on pourrait reformuler par : « le protectionnisme consisterait à défendre ses intérêts quand nos concurrents le font avec acharnement »). Par cette méthode manuelle, on peut donc mener des analyses fines sur des séquences précises, « sonder » le corpus par des prélèvements d'extraits jugés pertinents, et faire fonctionner des niveaux d'analyse qui mettent en écho ou en confrontation différentes constructions ou points de vue, éclairant des questions politiques (comme ici l'affrontement entre Nicolas Sarkozy et Alain Juppé lors des primaires de la droite). Plusieurs chercheurs confrontés au même corpus ne se focaliseront pas sur les mêmes éléments, aussi, d'un point de vue méthodologique, un recours à des outils qui serviraient de médiation entre le chercheur et son corpus peut constituer un bon moyen pour objectiver le regard du chercheur sur ses observables, et fournir des procédures reproductibles de sélection des données textuelles. Certes, les outils peuvent donner lieu à des lectures différentes des phénomènes, mais la mesure des corpus contribue à faire émerger des facteurs saillants (que le chercheur peut choisir d'étudier ou non).

4. Approche textométrique de la nomination : complémentarités, intérêts et limites

Pour cette analyse, le corpus est un document qui contient l'ensemble des données, mais aussi de métadonnées qui informent sur la nature des textes rassemblés.

4.1 Méthode et principes

Un balisage est ainsi ajouté pour intégrer ces métadonnées, comme dans la figure 1 dans le cadre d'un corpus formaté pour le logiciel *Iramuteq*. Dans le cas de cet outil, le balisage consiste à faire précéder les métadonnées d'un astérisque *.

```
**** *interview_Alain-Juppe_2016_11-14 *acteur_Alain-Juppe *date_11-14 *annee_2016
Bonjour, Alain Juppé |

Bonjour

Merci beaucoup d'être dans ce studio d'RTL à une semaine maintenant, même moins d'une
semaine d'ailleurs...
```

Fig. 3. Formatage du corpus et insertion de métadonnées.

Ces métadonnées permettent de connaître l'interviewé et la date de l'interview, avec différentes présentations qui permettront ensuite d'interroger le corpus selon telle ou telle dimension (l'interviewé, la date, l'année). Notre premier traitement de ce corpus se fait par une classification lexicale implémentée dans le logiciel *Iramuteq*. Loubère (2016) précise qu'une classification de type Reinert proposée par Iramuteq « permet de mettre en avant les mondes lexicaux », sous forme d'une classification hiérarchique descendante (un dendogramme). Après lemmatisation, un « *tableau à double entrée répertoriant la présence ou absence dans les segments des formes pleines retenues » est construit ; sur ce tableau est effectuée « une série de bi-partitions reposant sur une analyse factorielle des correspondances* ». Un regroupement des segments en classes en fonction des mots qui les composent permettra la constitution d'un dendrogramme qui donne à voir les classes lexicales, composées des mots les plus représentatifs.

Cela nous permet donc de dégager, après analyse et observation des résultats, les grandes thématiques du corpus. Dans le cas du corpus TALAD, qui contient un très grand nombre d'interviews, un premier filtrage a été effectué, en sélectionnant les 466 interviews qui

contenaient *protectionnisme économique*. Sur la figure 4, nous présentons le résultat obtenu après application de la méthode : on trouve des thématiques très larges, parfois éloignées du terme candidat (classe 1 de commentaire de l'élection présidentielle, classe 2 qui concerne des termes spécifiques au genre de l'interview). Le problème possible vient du fait que les interviews sont longues (3165 occurrences par interview en moyenne), et que le *protectionnisme économique* est un élément parmi d'autres.

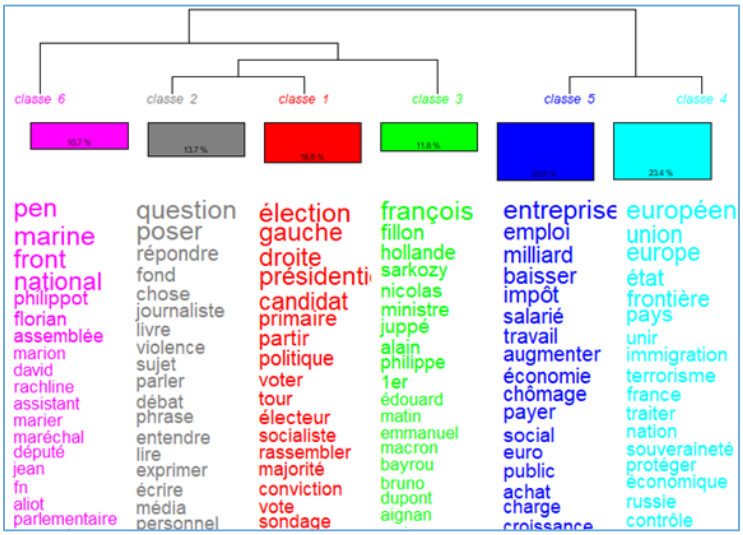

Fig. 4. CHD du corpus global des interviews contenant "*protectionnisme* ***économique***".

La spécificité du genre (des interviews de matinales qui abordent un grand nombre de sujets) nous incite dans un second temps à prélever, dans l'ensemble du corpus, des parties qui concernent plus spécifiquement le sujet abordé.

4.2 Constitution de sous-corpus

On peut donc construire un sous-corpus centré sur ce terme avec l'ensemble des séquences qui contiennent le terme « protectionnisme » avec le « concordancier » (qui est en fait, dans Iramuteq, un relevé des Segments de Textes (ST) contenant la forme candidate), et relancer une analyse sur ce sous-corpus. L'objectif d'une telle démarche est de se concentrer sur les séquences qui évoquent la nomination qui nous intéresse, en faisant l'hypothèse que ces reformulations, négociations, constructions spécifiques, se feront dans l'environnement même de ce terme. On choisit « protectionnisme » plutôt que « protectionnisme économique » afin de saisir les usages elliptiques (sans l'adjectif) de ce terme. On obtient la classification hiérarchique descendante suivante (figure 5). Les classes 3 et 4 semblent indiquer des éléments thématiques qui environnent le sujet du protectionnisme: une dimension "politicienne" dans la classe 3 (avec quelques termes génériques comme "*dépense*", "*emploi*" qui sont en lien avec les discours du Rassemblement National, représenté par [*Le*] "*pen*") et une dimension plus économique avec la classe 4 ("*négocier*", "*commercial*" en lien avec le programme de "*fillon*"). Les classes 1, 2 et 5 sont plus productives pour l'analyse de la nomination, puisque l'on voit, dans les termes des classifications, des concurrents possibles au terme initial. Pour l'étude de la nomination, le croisement des classes et des «segments de textes caractéristiques» offerts par le logiciel permet un retour au texte et l'étude des emplois des termes de ces classes dans leur contexte.

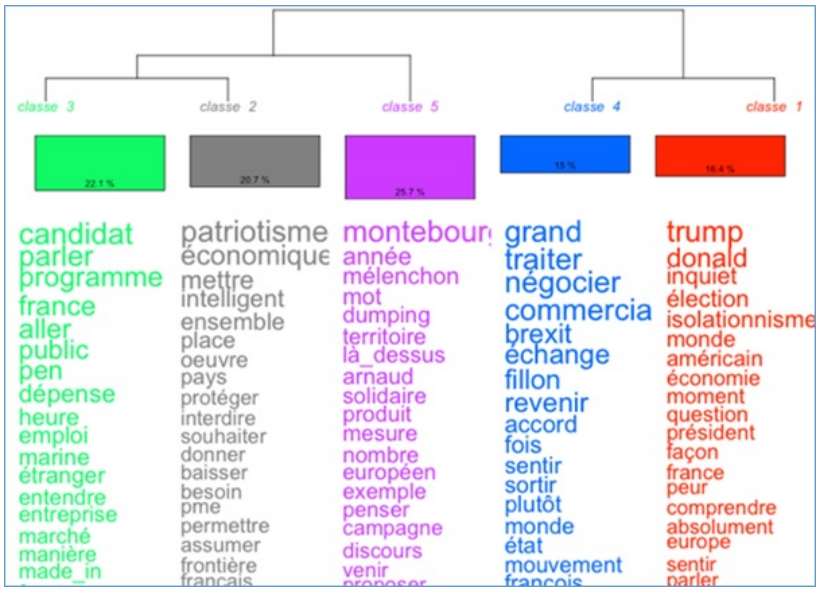

Fig. 5. CHD du sous-corpus des ST contenant protectionnisme

Regardons par exemple la classe 1 (figure 6). Dans tous ces contextes, il est question de Donald Trump, et nous observons que lorsqu'il s'agit du contexte américain, "*isolationnisme*" peut être privilégié, et que les énoncés témoignent d'une inquiétude. Cela signifie que ce qui conçu comme de la "protection" par le président américain est appréhendé par les acteurs français comme une attitude d'isolation vis-à-vis du reste du monde, notamment l'Europe.

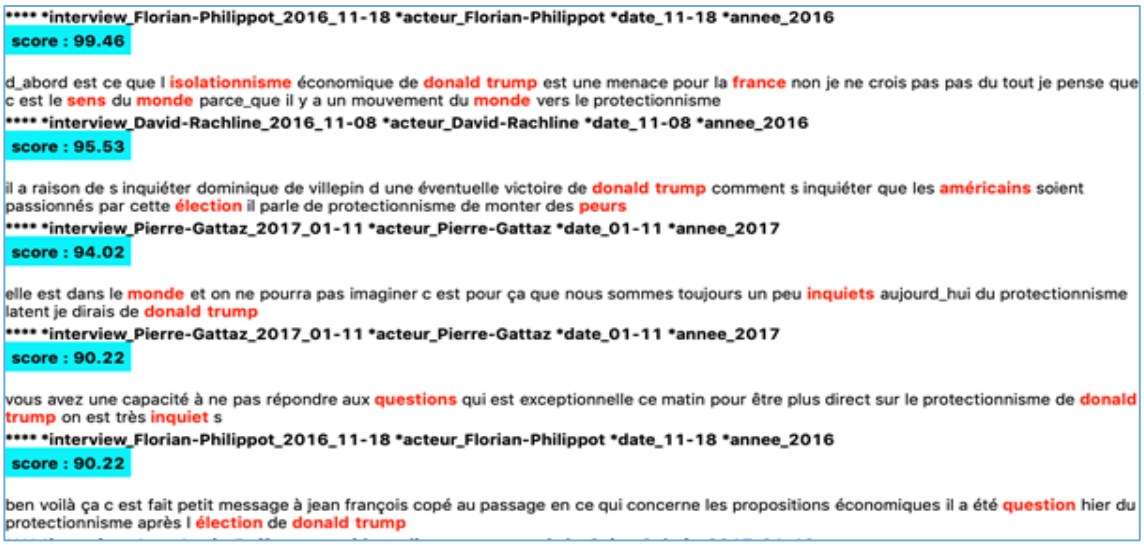

Fig. 6. Segments de textes caractéristiques de la classe 1.

Dans un autre contexte, celui de la classe 2 (figure 7), où il est question des solutions proposées (c'est Marine le Pen qui se distingue dans cette classe, comme nous le voyons avec des scores les plus élevés), la nomination passe par une qualification ("*protectionnisme intelligent*"), une spécification ("*patriotisme économique*") et une association à une autre notion ("*patriotisme* économique »).

Cette approche statistique a l'avantage de pouvoir aborder le corpus de manière plus globale, en procédant au repérage de nominations alternatives ou de structures spécifiques. L'originalité de la méthode est ici de « détourner » la CHD, en la considérant comme un bon

moyen d'accéder à différentes nominations. Cela suppose une procédure de constitution d'un sous-corpus, qui n'épuisera pas l'analyse discursive du corpus, mais se concentrera sur le phénomène spécifique de la nomination.

Fig. 7. Segments de textes caractéristiques de la classe 2.

Grâce au retour au corpus, on peut contribuer à appréhender la discursivité de la nomination, en liant certains emplois à des thématiques, voire à des idéologies. Néanmoins, l'aspect textuel qui était présent dans l'approche manuelle (travailler finement sur une séquence) est difficilement utilisable ici, et une méthode automatisée qui prendrait en compte les aspects syntaxiques et pragmatiques des extraits est nécessaire. C'est ici que peut intervenir le TAL, avec l'étude des chaînes de coréférence.

5. Repérage des nominations par détection des chaînes de coréférence

5.1. Objectifs

Les approches outillées présentées reposent sur des techniques textométriques. Ces outils permettent un repérage efficace d'unités d'intérêt dans de grands corpus, mais aussi d'analyser leurs situations d'usage en discours. Toutefois, ils ne permettent pas une recherche exploratoire de variantes de nomination. Ils proposent essentiellement d'explorer les corpus à partir de formes lexicales déjà isolées (listées dans Lexico ou à l'aide d'expressions régulières pour TXM). Certains prétraitements peuvent concerner un étiquetage morphosyntaxique ou une lemmatisation (Alceste, Hyperbase). Mais, à l'exception de Tropes (qui construit les classes sémantiques d'un discours à l'aide d'un dictionnaire), une analyse plus profonde, qu'elle soit syntaxique ou sémantique, n'est jamais envisagée. C'est ainsi que les outils dédiés à l'AD ne permettent pas de traiter les ambiguïtés de sémantique lexicale. Ces outils imposent en outre de définir a priori les dénominations disponibles pour un référent donné : la stratégie est alors d'ordre confirmatoire et non exploratoire. Dans le cadre du projet TALAD, nous nous proposons précisément d'adapter à l'analyse du discours des modules de traitement issus du TALN, pour fournir une aide exploratoire au chercheur en AD. La stratégie d'exploration que nous présentons dans cet article relève d'une analyse purement syntaxico-pragmatique. L'idée est d'identifier les chaines de coréférence dans lesquelles apparait une nomination, et de rechercher les reprises coréférentielles qui sont des reformulations de cette dernière. La notion de coréférence, commune en sciences du langage comme en TAL, est directement reliée à celle de référence, c'est-à-dire la relation qui existe entre les termes employés dans le

discours et les entités auxquels ces termes réfèrent dans l'univers du discours. Considérons par exemple l'énoncé suivant :

(1) *il a raison de s'inquiéter Dominique de Villepin d'une éventuelle victoire de Donal Trump*

Dans cet énoncé, le pronom *il* et le patronyme *Dominique De Villepin* réfèrent à une même entité du discours, à savoir un ancien premier ministre français. On dit que les deux termes partagent à ce titre relation de coréférence, et forme donc une **chaîne de coréférence** de deux termes. Ces termes qui réfèrent à une entité du discours sont appelés **mentions**. Au sein d'une chaîne de coréférence, on distingue enfin différents types de relations de coréférence :

- Directe : reprise du même terme *la victoire ... cette victoire*
- Indirecte : reprise par un groupe nominal de tête différente *la victoire … l'élection*
- Pronominale : reprise par un pronom anaphorique *la victoire ... elle.*

Le scénario d'utilisation prototypique de la résolution des coréférences dans le projet est le suivant : partant d'une forme de nomination proposée par le chercheur en AD, la détection des chaînes de référence va proposer, sous forme de concordance, toutes les unités lexicales qui sont susceptibles d'être une reformulation de la nomination (ex : *protectionnisme*/ *patriotisme économique* ; *revenu universel/ revenu d'existence*). Parmi ces formes, certaines seront originales et permettront au chercheur d'aller plus loin dans son étude des variations de nomination. L'analyse pourra être conduite dans diverses perspectives, en se donnant pour champ d'observation le texte, mais aussi les discours circulants envisagés suivant une certaine logique (par époque, scripteurs…). Le chercheur en AD pourra ainsi atteindre le paradigme nominationnel d'une entité donnée. Du point de vue du TAL, l'originalité de l'approche consiste à se focaliser sur les coréférences indirectes (ex : *hidjab … voile islamique*). Cette focalisation va de soi côté AD mais elle est peu explorée en TAL, car les indirectes, qui ne représentent qu'une faible proportion des coréférences observées, ne sont pas ciblées par les techniques actuelles de résolution. Observons l'exemple suivant (interview avec M. Le Pen) :

(2) *à mettre en œuvre du **protectionnisme intelligent** à mettre en œuvre du **patriotisme économique** pour donner un avantage aux entreprises françaises dans la commande publique voilà tout cela. Le **patriotisme économique qui** n'a jamais été mis en œuvre.*

Avec la coréférence infidèle entre *protectionnisme intelligent* et *patriotisme économique*, on observe une redéfinition axiologique de la notion de *protectionnisme* : à la /protection/ s'ajoute, en gommant la dimension stratégique, un trait sémantique /défense de la patrie/. L'analyse discursive des relations sémantiques sous-jacentes à ce travail sur la dénomination (par ex : sommes-nous en présence d'une synonymie pure, ou y-a-t-il glissement référentiel) suppose une compréhension fine du discours. Un modèle linguistique portant sur les indices des plans (ontologique, langagier), des procédés d'élaboration (introduction, ajustement, rejet) et des attitudes (prise en charge, interaction, cadrage), qui renvoient aux modalités spécifiques de catégorisation et de nomination en discours, a été élaboré (Jackiewicz 2020) et opérationnalisé sous la forme d'un système de détection à base de règles. A l'opposé, la détection des coréférences ne vise pas une analyse aussi fine. On ne cherche pas en particulier à distinguer la coréférence de ce que (Recasens et al. 2010) a qualifié de `NEAR_IDENTITY`, mais seulement de détecter des variantes nominationnelles qui soumises à l'AD.

5.2. Mise en œuvre

Notre chaîne de détection des coréférences permet la recherche exploratoire de nouvelles variantes à partir d'un terme graine. Plutôt que de rechercher l'ensemble des chaînes de

coréférence dans un corpus donné, nous nous limitons à présenter au chercheur en AD les mentions qui sont potentiellement coréférentes avec ce terme d'amorçage, ceci sous la forme d'un concordancier. Si l'on reprend l'exemple (1) avec le terme graine *protectionnisme*, on obtient en sortie l'affichage donné en figure 8.

ID	Thème gauche	Contexte	Thème droit
1	$, mettre_en_œuvre	du **protectionnisme** intelligent	mettre_en_œuvre
1	intelligent, à mettre_en_oeuvre	du **patriotisme économique** pour	donner, un avantage
1	commande, publique	le **patriotisme économique** qui	mettre_en_œuvre, $
1	publique, patriotisme économique	patriotisme économique **qui** n	mettre_en_œuvre, $

Fig. 8. Affichage des chaînes de coréférence sous forme de concordances

Les chaînes de coréférence sont présentées sous forme de concordances dans un tableau à 4 colonnes. La colonne `ID` est un indicateur du corpus concerné, et le cas échéant du locuteur concerné dans le cas d'un dialogue (débat, interview…). La colonne `Contexte`, donne ligne par ligne les mentions de la chaîne de coréférence avec un contexte gauche et droit de longueur paramétrable (1 mots sur la gauche et sur la droite dans cet exemple, faute de place) dans un concordancier classique. En outre, deux autres colonnes (`Thème gauche` et `Thème droit`) donnent le contexte d'occurrence des mentions d'un point de vue thématique. On affiche ici le lemme des mots de contenus situés à gauche et à droite de la mention (2 mots de contexte dans cet exemple). L'expert a ainsi une idée du contexte thématique d'occurrence des mentions de la chaîne. Cela l'aide à comprendre dans quelle thématique la nomination est employée, ce qui constitue par exemple un aspect essentiel dans l'analyse du discours politique. L'expert peut découvrir immédiatement des variantes nominationnelles du terme graine (ici *protectionnisme*) via la coréférence. C'est le cas ici de *patriotisme économique*. Il peut également focaliser l'affichage sur les seules coréférences indirectes (figure 9), car notre système type les relations qu'il découvre. Cela lui permet de cibler les seules variantes de nomination, mais fait perdre en contrepartie le contexte d'occurrence des formes qui ne sont pas des variantes. L'analyste peut ajuster l'empan des contextes affichés, ainsi que certains paramètres de la détection des coréférences, que nous allons brièvement décrire.

ID	Thème gauche	Contexte	Thème droit
1	$, mettre_en_œuvre	du **protectionnisme** intelligent	mettre_en_œuvre
1	intelligent, mettre_en_oeuvre	du **patriotisme économique** pour	donner, avantage
1	commande, publique	le **patriotisme économique** qui	mettre_en_œuvre, $

Fig. 9. Affichage des chaînes de coréférence avec focalisation sur les indirectes

Le cœur de cet outil d'aide à l'exploration est la détection des chaînes de coréférence. Cet objectif applicatif spécifique (trouver des variantes et non pas toutes les relations existant dans un texte) nous conduit à des choix particuliers par rapport à l'état de l'art :

- Nous tenons à construire des modèles qui soient explicables et intelligibles pour les analystes. C'est pourquoi nous n'envisageons pas le recours à des techniques neuronales pour la détection des coréférences, car ils pêchent de ce point de vue. Toutefois, l'étape d'identification des mentions fait appel à ces techniques.

- Dans cette tâche d'exploration, notre cas d'utilisation consiste à proposer à l'analyste des variantes de nomination potentielles, que celui-ci filtrera ensuite par analyse experte. L'important est donc de relever un maximum de termes candidats. On privilégiera donc le rappel à la précision dans l'évaluation des performances. L'analyste ne doit toutefois pas être

noyé par trop de bruit : une évaluation expérimentale sera menée avec des spécialistes de l'AD pour estimer un seuil de précision acceptable pour ces derniers.

Une première chaîne de résolution est désormais opérationnelle, qui demande encore à être optimisée. Elle suit une stratégie d'analyse en trois passes. Dans un premier temps, nous procédons à une détection des mentions coréférentielles, puis nous détectons les paires de mentions coréférentes, le tout par des techniques d'apprentissage automatique. Enfin, nous reconstruisons les chaînes de coréférence par chaînage des relations binaires détectées. Cette stratégie d'analyse séquentielle est restée classique jusqu'à l'apparition récente de systèmes end-to-end. Une telle approche comporte des risques de propagation des erreurs. Toutefois, le système étant amorcé à partir d'une seule nomination, ce risque est limité. Nous ne décrivons pas ici le module neuronal de détection des mentions qui est une reprise du travail de Loïc Grobol (Grobol 2019). Disons simplement que ce module cherche à classifier par force brute toutes les chaînes de 1, 2, … N mots comme des mentions (ou pas), ceci à l'aide d'un RNN bidirectionnel. L'étape suivante va considérer tous les couples de paires de mentions détectées, et les classer comme paires coréférentes, ou non. Cette étape repose indifféremment sur un classifieur à vaste marge (SVM) ou des forêts d'arbres aléatoires (*random forest*) qui ont été appris sur un grand corpus francophone annoté en coréférence : ANCOR (Muzerelle et al. 2014). Les SVM et les forêts d'arbres aléatoires reposent sur des techniques de classification qui ont fait leurs preuves et qui ont l'intérêt d'avoir un comportement paramétrique bien compris : l'optimisation de ces techniques d'apprentissage automatique est donc maitrisable assez aisément. Notre système de classification considère 25 traits d'apprentissage morphosyntaxiques ou syntaxiques. Ces traits sont issus de (Desoyer et al. 2015). Ils décrivent les paires de mentions en termes de distance dans le texte, de catégories morphosyntaxique, d'identité de genre et de nombre, et proximité morphologique (exemple, inclusion d'une mention dans une autre : *patriotisme* / *patriotisme économique*). Le seul trait de nature quelque peu sémantique est le type d'entité nommée rencontré dans le cas d'une telle mention. Ce trait est fourni par le système *mXs,* qui fait appel à une fouille hiérarchique de séquences (Nouvel et al. 2012). Les chaînes de coréférence sont enfin construites par des algorithmes de chaînage classique, actuellement encore en cours d'optimisation.

6. Conclusion

L'intérêt du travail présenté est de montrer ce qu'un outillage allant au-delà de la textométrie pourrait apporter à l'AD. Du côté de l'exploration des variations d'une nomination, l'enjeu est de relier l'analyse des circulations et reprises d'une nomination (AD) et la coréférence (TAL). Cela intéresse directement l'AD car un même référent peut être désigné par différentes unités lexicales (mono- ou polylexicales) qui sont les supports de la nomination. L'étude de la coréférence permet ainsi de dégager les unités lexicales choisies pour désigner une même entité. Du côté du TAL, la reconnaissance automatique des chaînes de coréférence représente un objet d'étude bien identifié, qui est au cœur de grandes campagnes d'évaluation telles que MUC, ACE ou CoNLL. Ces campagnes conservent toutefois parfois un aspect purement technologique, au sens où elles ne sont pas toujours reliées à une tâche applicative supérieure. Dans le cadre du projet TALAD, nous proposons au contraire une tâche totalement originale, l'exploration de nominations, dont on espère qu'elle permettra de définir de nouveaux objectifs scientifiques à la résolution des coréférences, en particulier sur la question des coréférences indirectes – ou coréférences infidèles (Schnedecker et Landragin 2014). Dans l'immédiat, nous travaillons à l'optimisation de notre chaîne de traitement, à la fois en termes de performances pures, et d'utilisabilité par l'AD. Le premier point va nous conduire à utiliser

des modèles de classification plus avancés. Nous envisageons en particulier d'utiliser des forêts d'arbres aléatoires (*random forest*) qui combinent l'intérêt d'un haut niveau de performance avec le caractère interprétable des arbres de décision. La focalisation de la recherche exploratoire sur les coréférences indirectes nous conduira également à développer des classifieurs spécifiques à chaque type de coréférence, mais aussi à intégrer des traits sémantiques pouvant en particulier provenir de plongements de mots (Mirzapour et al. 2020). La question de l'utilisabilité du système nous conduira à réétudier celle de l'évaluation des systèmes de résolutions des coréférences (Luo 2005, Holen 2013).

Références

Desoyer A., Landragin F., Tellier I., Lefeuvre A. and Antoine J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *TAL*, vol. 55 (2) : 97-121.

Grobol L. (2019). Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French. *Proc. of CRAC19*, pp. 8-14.

Holen G.I. (2013). Critical reflexion on evaluation practices in coreference resolution. *Proc. of the 2013 NAACL-HLT Student Research Workshop*, pp. 1-7.

Jackiewicz A., Bebeshina-Clairet N., Cassier M, Frontini F., Halftermmeyer A., Longhi J., Luxardo G. and Nouvel D. (2019). Vers une ontologie de la nomination et de la référence dédiée à l'annotation des textes. ToTH 2019 (Chambéry, France) (https://hal.archives-ouvertes.fr/hal-02269154).

Jackiewicz A. (2020). Un modèle linguistique pour l'étude des nominations émergentes. *Colloque international Le mot dans la langue et dans le discours* (Vilnius, Lituanie).

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

Longhi J. (2015 éd.). Stabilité et instabilité dans la production du sens : la nomination en discours, *Langue Française* 188.

Longhi J. and Salem A. (2018). Approche textométrique des variations du sens. *JADT 2018*, pp. 452-458.

Loubère L. (2016). L'analyse de similitude pour modéliser les CHD. *JADT 2016* http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83440/83440.pdf.

Luo X. (2005). On coreference resolution performance metrics. *Proc. of HLT'2005&EMNLP'2005*, pp. 25-32.

Mirzapour M., Raghed W., Cousot L. and Antoine J.Y. (2020). Introducing RezoJDM Semantic DataSet for Relation Prediction Task. *Proc. of LREC'2020* (submitted).

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-Y., Pelletier A., Maurel D., Eshkol I. and Villaneau J. (2014). ANCOR_Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures. *Proc. of LREC'2014* (Reykjavik, Island), pp. 843-847.

Nouvel D., Soulet A., Antoine J.-Y. and Friburger N. (2012). Coupling Knowledge-Based and Data-Driven Systems for Named Entity Recognition. *Proc. of EACL'2012* (Avignon), pp. 69-77.

Recassens M., Hovy E. and Marti M. (2010). A Typology of Near-Identity Relations for Coreference. *Proc. of LREC'2010* (Malta), pp. 149-156.

Schnedecker C. and Landragin F. (2014). Les chaînes de référence : présentation. *Langages* 195 : 3-22.

Siblot P. (2004). Du dialogisme de la nomination. *Proc. of Dialogisme et Nomination* (Montpellier), pp. 331-337.