# TRAITEMENT AUTOMATIQUE DE LA CORÉFÉRENCE POUR UNE APPLICATION DE VEILLE TERMINOLOGIQUE

Mémoire de MASTER I

Clémentine Adam

Sous la direction de Didier Bourigault, Franck Sajous et Ludovic Tanguy

Je remercie Didier Bourigault, Franck Sajous et Ludovic Tanguy, qui m'ont encadrée pour ce travail de master I. Je remercie également Myriam Bras, pour ses conseils bibliographiques, James, qui a contribué à annoter manuellement l'extrait de corpus ayant servi à l'évaluation de l'outil présenté, et toute la promotion de master I TAL de cette année 2006-2007.

# TABLE DES MATIERES

# INTRODUCTION

Le point de départ de notre travail de mémoire est l'observation d'un outil de veille terminologique, LexiMedia2007. Cet outil statistique, qui utilise en amont une analyse syntaxique effectuée par SYNTEX, extrait d'un corpus régulièrement mis à jour et constitué d'articles politiques les syntagmes les plus fréquents, ou ceux dont l'augmentation (ou la diminution) est la plus importante d'une semaine sur l'autre ; cela constitue un point de vue intéressant pour considérer quels sujets ou quelles personnalités politiques font l'actualité dans la presse, et comment ils évoluent.

Nous avons observé que parmi les syntagmes les plus courants, et donc les plus mis à l'honneur par LexiMédia2007, se trouvaient beaucoup de syntagmes coréférents entre eux ; ces syntagmes forment entre eux des classes, et la variation de l'un suit généralement la variation de l'autre. Il nous a paru intéressant de tenter de mettre au jour ces « classes » formées par des syntagmes ayant le même référent, en essayant de résoudre de manière automatique les liens de coréférence.

Notre travail se divise en deux parties :

Dans la première, nous expliquons dans quel contexte prend place LexiMédia2007, et nous détaillons les observations que nous avons faites à partir de cet outil, et qui nous ont conduite à nous pencher sur la coréférence ; nous apportons ensuite quelques notions théoriques sur les phénomènes de référence et de coréférence, afin de nous situer parmi les autres travaux portant sur la résolution automatique de ces phénomènes. Une originalité de notre programme est l'utilisation (conditionnée par les caractéristiques de notre corpus) de mesures de similarité pour résoudre les liens coréférentiels ; nous expliquons les raisons de ce choix et détaillons les différentes mesures de liaison testées.

La seconde partie est consacrée à la description du travail informatique que nous avons effectué ; après avoir caractérisé notre corpus, nous décrivons de manière linéaire (module par module) les traitements informatiques que nous lui faisons subir afin de résoudre les liens coréférentiels ; nous évaluons ensuite les résultats obtenus, et citons quelques limites et perspectives d'amélioration de notre programme.

# PREMIÈRE PARTIE :

# Contexte du travail ; notions théoriques

# 1. Lexicométrie et LexiMedia2007

*« Les mots n'ont pas de sens,*
*ils n'ont que des emplois. »*
*Wittgenstein*

Nous allons dans cette section de notre mémoire parler d'un outil mis au point par Didier Bourigault et Franck Sajous, du laboratoire CLLE-ERSS : LexiMédia2007. LexiMédia2007 permet de suivre au fil des semaines l'évolution de syntagmes utilisés dans des articles de presse traitant des présidentielles de 2007. LexiMédia2007 est donc un outil statistique, mais aussi essentiellement linguistique, conçu pour traiter des corpus appartenant au domaine de la politique.

Notre but est dans un premier temps (1.1.) de présenter le contexte dans lequel s'inscrit LexiMédia2007. On l'a vu, ce dernier se situe au carrefour de différents domaines : linguistique, statistique, politique. Nous allons donc d'abord décrire, dans une perspective diachronique, les liens qui se sont tissés entre ces trois domaines. La linguistique a depuis longtemps été invoquée pour aider à l'étude de textes politiques dans le cadre de l'analyse de contenu puis de l'analyse du discours (1.1.1.). Lorsque vers la fin des années cinquante naît la statistique lexicale (1.1.2.), c'est tout naturellement que les linguistes tentent de l'appliquer aux textes politiques. Ceux-ci semblent avoir beaucoup à révéler à la lexicométrie ; mais ils invitent également à développer les méthodes, le simple « comptage de mots » montrant rapidement ses limites (1.1.3.). Dans une perspective synchronique, LexiMédia2007 prend place dans un terrain où foisonnent les outils statistiques consacrés aux élections présidentielles de 2007 (1.1.4). Il vient en particulier compléter les outils développés par Jean Véronis.

Dans un second temps (1.2.), nous allons décrire plus en détail LexiMédia2007 : les données qu'il utilise (1.2.1.), ses méthodes (1.2.2.), ses résultats (1.2.3.). Nous verrons que sa principale spécificité est liée à l'utilisation en amont de l'analyseur syntaxique SYNTEX, que nous présenterons rapidement. Enfin, nous exposerons les observations qui nous ont conduite à diriger notre travail vers le traitement automatique de la coréférence (1.2.4.).

## 1.1. Politique et linguistique

### 1.1.1. La tradition de l'analyse du discours

L'analyse de contenu est née aux États-Unis au début du XXe siècle. Les premières études portèrent sur la presse écrite. B. Berelson (cité par [Reinert 2006]) définit l'analyse de contenu comme « une technique de recherche, pour la description objective, systématique et quantitative du contenu manifeste des communications, ayant pour but de les interpréter. » La méthode est généralement la suivante : L'analyste commence par définir un ensemble de classes d'équivalence (thèmes, mots, éléments de syntaxe ou de sémantique) dont il repèrera ensuite les occurrences dans les textes. Puis ces classes d'équivalence sont répertoriées dans une grille où l'analyste transcrit les quantités (comptées « manuellement ») qui leur correspondent.

L'analyse du discours serait née, selon [Mayaffre 2005-b], à la fin des années 1960. Elle a été créée pour pallier certaines insuffisances de l'analyse de contenu :

> L'interprétation doit prendre en compte le mode de fonctionnement des discours, les modalités de l'exercice de la parole dans un univers déterminé. [Maingueneau 1991 : 9]

Dominique Maingueneau définit l'analyse du discours comme étant l'analyse de l'articulation du texte et du lieu social dans lequel il est produit. Le texte relève de la linguistique ; le lieu social, de la sociologie ou de l'ethnographie ; l'analyse du discours, qui étudie le mode d'énonciation, se situe à leur charnière.

Un précepte de l'analyse du discours est que « toute énonciation, fût-ce la plus spontanée, apparaît soumise à de multiples codes. » [ibid. : 21] Mais ce qui intéresse le plus l'analyse du discours française, ce sont les textes fortement contraints : ceux qui sont produits dans le cadre d'institutions, inscrits dans un interdiscours serré, ou qui fixent des enjeux historiques ou sociaux. Le texte politique constitue donc un matériau idéal.

### 1.1.2. Analyse de textes politiques et corpus

D'après [Mayaffre 2005-b], l'analyse du discours ne saurait être possible sans en appeler aux corpus :

> Aucune étude qui entend appréhender les discours ou les textes dans leur dimension sociale ou politique n'est indifférente à la question des corpus pour la raison suffisante qu'une production linguistique revêt une dimension politique seulement si elle se trouve effectivement attestée dans la société.

Nous soulignions plus haut que l'analyse du discours était une discipline charnière entre linguistique et sociologie. Les corpus utilisés pour l'analyse seront en conséquence « à cheval » sur ces deux domaines[1] :

> Le matériel linguistique qui a été rassemblé pour être traité a été choisi pour sa valeur extralinguistique. [ibid.]

Damon Mayaffre qualifie le n°13 de la revue Langages, paru en 1969, de « fondateur de l'Analyse du Discours » (*l'Analyse du discours* est d'ailleurs le titre de ce numéro). [Guilhaumou 2002] considère que ce sont les événements de mai 1968 qui avaient alors « promu » le discours politique au rang de « genre ». Le corpus est alors défini comme « un ensemble déterminé de textes sur lesquels on applique une méthode définie » (Jean Dubois, cité par [Guilhaumou 2002]). La méthode proposée par [Dubois 1969 : 6-7], inspirée de Zelig S. Harris, consistait à relever systématiquement les phrases contenant tel ou tel mot-pivot[2], puis à les transformer par l'application de certaines règles (cf. [Harris 1969]) afin de constituer des « classes d'équivalence » plus faciles à observer.[3]

Le principal problème de cette méthode est que le choix des mots-pivots repose sur l'analyste, lequel est comme tout un chacun soumis à l'effet de l'expérimentateur. La lexicométrie qui, nous allons le voir, s'est développée en parallèle, permettra d'apporter des solutions à ce problème, le biais de l'ordinateur assurant l'objectivité de l'analyse.

## 1.2. Statistiques et linguistique : la lexicométrie

### 1.2.1. Naissance de la lexicométrie

La statistique lexicale est née en France vers la fin des années 1950. Les méthodes statistiques avaient alors fait leurs preuves dans de nombreuses autres disciplines des sciences humaines, et ce fut d'abord aux statisticiens[4], et non aux linguistes, que vint l'idée de les appliquer à l'étude de la langue. Pierre Guiraud[5] et Charles Muller[6] soulignent tous deux les réticences

[1] Le corpus politique est selon [Mayaffre 2005-b] « un lieu problématique de rencontre entre la langue et la société ».

[2] [Dubois 1969 : 6] donne l'exemple de l'étude du mot socialisme chez Jaurès : « le corpus est alors fait de toutes les occurrences du mot socialisme compris dans les divers environnements. »

[3] « Les procédures applicables au corpus permettent alors de substituer des termes les uns aux autres dans des environnements équivalents, de les réduire par des transformations inverses à des segments plus simples (ainsi les phrases passives seront transformées en phrases actives afin de repérer des environnements identiques), de constituer des classes d'équivalence. » [ibid.]

[4] On pourrait par exemple évoquer les travaux de G. U. Yule sur la longueur des phrases.

[5] « La linguistique est la science statistique type ; les statisticiens le savent bien ; la plupart des linguistes l'ignorent encore. » [Guiraud 1960 : 15]

[6] « Comme nous nous recrutons souvent parmi les esprits les plus rebelles aux formules mathématiques, comme nos travaux nous portent à une certaine méfiance envers le raisonnement géométrique quand il prétend

éprouvées par les linguistes à faire appel à ces méthodes. Ils furent parmi les premiers (linguistes) à utiliser la statistique, notamment lors d'études portant sur le style de grands auteurs classiques ([Guiraud 1955], [Muller 1964]) – aujourd'hui encore Corneille reste sous les feux de la rampe (bien que « dans l'ombre de Molière ») avec la parution de [Labbé 2003]. En 1967 est fondé à l'ENS de Saint-Cloud le centre de lexicologie politique (qui devint par après le laboratoire « lexicométrie et textes politiques »). On commence à parler de « lexicométrie » dans les années 1970. L'étude effectuée par Tournier *et alii* en 1975, sur les tracts de mai 1968, est qualifiée de « pionnière » par [Guilhaumou 2002]. Le laboratoire de Saint-Cloud assurera pendant plus de trente ans la vitalité de la lexicométrie politique en France ; beaucoup des auteurs dont nous allons parler dans la suite de cette présentation générale en faisaient partie, notamment André Salem et Maurice Tournier.

### 1.2.2. Quelles unités pour l'analyse ?

La statistique, qui consiste à décompter, faire des calculs sur des objets, implique de manière inhérente le choix d'une unité pour ces calculs. Pour pouvoir additionner des objets entre eux, il faut les considérer « au moins le temps d'une expérience comme des occurrences identiques d'un même archétype ou forme plus générale. » ([Salem 1987 : 15]) Toute étude statistique doit donc être précédée d'une cruciale phase de segmentation, qui ramènera le continu de la chaîne textuelle à du discontinu quantifiable. Le résultat de la segmentation sera une succession d'unités distinctes.

En statistique lexicale, l'unité la plus simple à distinguer est la forme graphique, définie comme une suite de caractères non-délimiteurs comprise entre deux caractères délimiteurs. Cela est bien sûr sujet à controverse : on le sait, les unités graphiques ne correspondent pas toujours aux unités sémantiques, loin de là (on peut avoir plusieurs unités graphiques pour une seule unité sémantique, ou inversement plusieurs unités sémantiques en une seule unité graphique). Mais le problème qui a fait couler le plus d'encre est le suivant : faut-il ou non lemmatiser les unités dégagées par l'opération de segmentation ?

Lemmatiser les formes graphiques, c'est les ramener chacune à leur entrée dans le dictionnaire. Il faut pour cela effectuer simultanément deux opérations (détaillées dans [Salem 1987 : 158-159]) :

- regrouper sous un même lemme ses diverses flexions ;

s'appliquer à nos disciplines, il est peut-être téméraire de vouloir créer cet instrument [une méthode d'initiation à la statistique linguistique]. » [Muller 1968 : 6]

- désambiguïser[7] les homographes relevant de lemmes différents.

La lemmatisation se heurte en conséquence à de nombreuses difficultés liées à l'ambiguïté de la langue, et est restée pendant très longtemps une opération très coûteuse (nous verrons en 1.3. qu'elle est aujourd'hui automatisable) ; la question était alors de savoir si ses avantages théoriques compensaient ses désagréments pratiques.

Les avantages de la lemmatisation se comprennent à la lumière des deux opérations que nous donnons plus haut. (i) Le regroupement permet de rassembler derrière un même lemme les différentes formes qui lui correspondent ; il est ainsi plus facile de les rechercher dans le texte ou de tirer des conclusions de leurs fréquences cumulées ([Brunet 2000], [Muller 1984]). Un exemple souvent donné est l'absurdité qu'il y a à traiter séparément les deux graphies « je » et « j' » du pronom personnel de la première personne du singulier. (ii) Le dégroupement permet d'éviter de tirer des conclusions faussées sur la fréquence d'un terme, qui aurait été « gonflée » par l'existence de termes homographes. Par exemple, [Mayaffre 2005-a] revient ainsi sur son travail doctoral dans lequel il avait entre autres comparé l'emploi du mot « parti » dans divers discours politiques de l'entre-deux-guerres :

> Il n'est pas question de remettre en cause l'interprétation donnée, mais l'honnêteté invite à préciser qu'elle repose, tout entière, sur une description linguistique dangereusement bancale du corpus. En effet, dans le traitement lexicométrique effectué, la forme «parti» qui a été indexée et décomptée et sur laquelle repose toute l'analyse, regroupait indifféremment - faute de lemmatisation - le substantif «parti» et le participe passé du verbe partir!

Finalement, pour les partisans de la lemmatisation, c'est toute la validité scientifique de la démarche qui est remise en question si l'on s'arrête à la matérialité graphique des textes : la forme graphique ne correspondant pas à une réalité linguistique, on ne peut tirer de l'étude de celle-ci que des conclusions superficielles sur le texte.

Les partisans de la non-lemmatisation avancent bien sûr l'argument pratique ; mais ils ajoutent qu'au-delà des motivations d'ordre pratique, il s'agit bien également d'un choix théorique. La raison la plus simple de ce choix est que l'objet des recherches lexicométriques n'est pas uniquement la thématique : on peut aussi s'intéresser à la syntaxe ou à la morphologie, auquel cas la forme des mots devient importante. De plus, on considère que la lemmatisation travestit le texte, sans même vraiment résoudre la question du sens : le fait de ramener une forme à un lemme ne lève pas toutes les ambiguïtés, le lemme pouvant lui-même être polysémique. Ces arguments sont développés par [Tournier 1985], mais c'est plus généralement tout le laboratoire de Saint-Cloud qui s'est longuement opposé à la

[7] [Mayaffre 2005-a] utilise le verbe « dégrouper », ce qui permet de souligner la nature opposée de ces deux opérations

lemmatisation ; leurs publications à ce sujet sont nombreuses depuis les années 1970 (notamment le « Plaidoyer pour la non-lemmatisation » [Collectif Saint-Cloud 1974]) jusqu'aux années 1990.

Finalement, certaines motivations théoriques venant étayer un choix très avantageux sur le plan pratique, la forme graphique est longtemps resté l'unité privilégiée des études lexicométriques, malgré une importante remise en cause de la pertinence linguistique (et donc scientifique) de cette approche.

### 1.2.3. Méthodes

Une fois la segmentation effectuée, les possibilités offertes par la statistique sont nombreuses. Les premières recherches en lexicométrie ont principalement porté sur la richesse, la spécificité, l'accroissement et l'évolution chronologique du vocabulaire. Plus généralement, on peut distinguer, selon [Salem 1987 : 36], deux grands types de méthodes lexicométriques :

- Les méthodes qui opèrent, pour chaque texte pris isolément, des comptages et des calculs d'indices statistiques ;
- Les méthodes statistiques contrastives qui produisent des résultats portant sur le vocabulaire de chacun des textes par rapport à l'ensemble des textes réunis dans un même corpus à des fins de comparaison.

### 1.2.4. Les segments répétés

D'après [Salem 1987 : 264], les premiers travaux français cherchant à extraire des unités complexes sont dus à G. Gorcy, R. Martin, J. Maucourt et R. Vienney, qui étudièrent les « groupes binaires » à partir du Trésor de la Langue Française. Il s'agissait ici de repérer des associations de mots sémantiques, séparés ou non par des mots fonctionnels (environ 300 formes listées préalablement). On procédait par « traitement séquentiel du texte » : toutes les paires possibles de mots binaires étaient extraites, puis celles qui dépassaient un certain seuil étaient sélectionnées.

C'est André Salem qui introduit la notion de « segments répétés ». La problématique qui a conduit au développement de cette notion apparaît dans la présentation qu'en fait André Salem particulièrement liée à la prise en compte de textes politiques. En effet, c'est d'après lui « dès les premières études statistiques effectuées sur des textes produits par des organisations politiques ou syndicales » que « les cadres de recherche de la statistique lexicale se sont révélés tout à fait insuffisants ». Lorsque l'on est confronté à des textes politiques, un certain

nombre de questions se poseraient plus particulièrement que si l'on avait affaire à d'autres types de textes :

> Quelles sont les unités qui circulent d'un texte à l'autre dans un corpus de textes politiques ? Sous quelle forme et avec quels contextes les formes du vocabulaire transitent-elles d'une formation discursive à une autre ? Peut-on objectiver par des comptages cette sensation que le discours politique est un genre où les répétitions sont très abondantes, que certaines séquences reviennent souvent, martelées au détriment du sens de chacune des formes qui les composent, allant même jusqu'à produire cette impression, souvent décrite de manière polémique, que le discours est produit par une « langue de bois » ?

Le fait qu'André Salem travaille sur un corpus de textes syndicaux a donc en partie conditionné sa théorie ; ce sont les segments répétés, on l'aura compris, qui permettront d'apporter des réponses à ces questions. Leur relevé se fait comme suit : dans chaque portion de texte contenue entre deux caractères délimiteurs, on extrait automatiquement toutes les suites de formes possibles, d'une longueur allant donc de 2 à N formes (avec N le nombre total de formes de la portion considérée), et sans exclure les mots fonctionnels. Lorsqu'une de ces suites de formes apparaît plus d'une fois dans le corpus considéré, il s'agit d'un segment répété. Les segments sont ensuite listés dans des inventaires alphabétiques, hiérarchiques (établis pour l'ensemble du corpus et pour chacune de ses parties), ou distributionnels (qui pour une même forme pôle sont ordonnés alphabétiquement ou par fréquences selon la forme qui suit ou qui précède). [Salem 1987 : 254] dit avoir été lui-même surpris par les résultats obtenus : « Loin de constituer des faits isolés les répétitions segmentales semblaient constituer la règle pour peu que le texte étudié soit assez long. »

Si nous avons décidé ici de parler des segments répétés, c'est parce qu'ils pouvaient sembler annoncer les unités complexes que traite LexiMédia2007. Mais nous tenons à souligner que si les segments répétés font apparaître des unités supérieures (les segments de taille 2, 3, etc.), l'unité de base n'en reste pas moins la forme graphique. [Salem 1987 : 20] justifie ce choix par son « caractère pratique et généralisable ». Aucun traitement linguistique n'est effectué en amont du processus. Par contre, l'inventaire des segments répétés permet de « mettre en évidence des unités qui constituent souvent des syntagmes autonomes » [Salem 1987 : 51]. Il s'agit donc d'un traitement essentiellement quantitatif, les éléments qualitatifs n'étant pris en considération que lors de l'analyse des résultats.

## 1.3. Vers une approche plus « qualitative » pour l'étude des textes politiques

### 1.3.1. Un genre : les textes politiques

Comme nous le disions au sujet des problématiques d'André Salem, le simple décompte des mots montre peut-être plus encore son insuffisance avec les textes politiques qu'avec les

autres. [Labbé 1990 : 26-27] évoque une comparaison faite entre des allocutions de J. Chirac et de F. Mitterrand : 91% des mots leur étaient communs. Il ne faut pas trop se hâter d'en conclure « le glissement vers le centre de la vie politique française, la mort des idéologies et l'unification du discours public en France au-delà des clivages partisans », pense Dominique Labbé, qui propose une autre interprétation :

> La langue et la société – qui délimite la scène politique et fixe les enjeux – surdéterminent le discours politique et lui imposent certains matériaux de base. Le fait que les matériaux soient les mêmes pour tous ne préjuge pas pour autant du contenu. (…) Dès lors, on voit bien que les substantifs ou les adjectifs étudiés isolément – même en prenant en compte la fréquence – ne sont guère utiles pour juger de la singularité d'un auteur ou d'un orateur.

Pour dépasser cette insuffisance, Dominique Labbé utilise notamment les co-occurrences : pour déterminer par exemple quel est « l'univers lexical » de la première personne du singulier, il cherche les mots qui apparaissent le plus souvent dans les mêmes phrases. On peut noter aussi qu'il effectue ses analyses sur un corpus entièrement lemmatisé (de manière semi-automatique).

D'autres auteurs ont remarqué la spécificité des textes politiques. [Chetouani 1998], par exemple, fait la remarque suivante :

> Le langage politique fonctionne avec une manipulation de signes linguistiques et d'outils symboliques. Cela explique pourquoi la fabrication des textes est un acte où interviennent des enjeux et où s'élaborent des automatismes et des répétitions.

Pour résumer, nous remarquerons que les textes politiques sont à double tranchant : d'une part, ils se prêtent bien à la statistique, car ce sont des textes très contraints, riches en routines et donc en répétitions qui pourront être mises à jour par des méthodes quantitatives. Mais d'autre part, ils nous engagent à approfondir les méthodes ; si les répétitions sont caractéristiques du genre, alors elles ne suffisent plus à caractériser tel auteur par rapport à tel autre.

### 1.3.2. La lemmatisation : un dilemme aujourd'hui dépassé

Le « dilemme » autour de la lemmatisation (le terme est de [Brunet 2000], qui intitule son article « Qui lemmatise dilemme attise ») se posait, nous l'avons vu plus haut, en ces termes : la lemmatisation apparaît trop coûteuse et engendre une perte d'informations, mais la non-lemmatisation remet en question la scientificité de la méthode. Les linguistes se sont donc trouvés pendant longtemps confrontés à l'alternative approche quantitative / approche qualitative, le choix résultant le plus souvent de la taille du corpus traité.

Mais depuis quelques années, les résultats de la lemmatisation automatisée se sont considérablement améliorés, et la marge d'erreur est à présent faible. On peut désormais avoir

accès en parallèle aux formes et aux lemmes, par exemple grâce à un logiciel comme *Hyperbase*. Ce logiciel de lexicométrie développé par Etienne Brunet a intégré un lemmatiseur automatique, et traite simultanément lemmes et formes pour toutes ses fonctionnalités (l'écran est divisé en deux colonnes, et pour chaque opération, comme par exemple la recherche des spécificités, les deux résultats sont donnés : lemmes spécifiques et formes spécifiques). Le dilemme est dépassé ; il s'agit d'après [Mayaffre 2005-a] d'un tournant dans la lexicométrie, qu'il propose de rebaptiser « logométrie », puisqu'elle ne traite plus uniquement des « lexies ».

La lemmatisation est à présent rapide et fiable, et intégrée à certains logiciels de lexicométrie : reste à savoir l'exploiter. En effet, on constate que les résultats de la statistique appliquée aux lemmes sont finalement les mêmes que ceux qui étaient obtenus en l'appliquant aux formes.

> Cela est toujours confortable de doubler ou tripler les résultats mais il existe une forme de trompe l'œil à prétendre confirmer des constats lorsqu'il s'agit, en fait, seulement de les répéter. [Mayaffre 2006]

[Pincemin 2004] considère que l'adaptation des logiciels de lexicométrie aux corpus étiquetés n'est pas satisfaisante. D'après elle, les informations apportées par l'étiquetage (catégories grammaticales et lemmes) sont traités exactement comme l'étaient les mots. On considère le texte soit comme une succession de graphies, soit comme une succession de catégories grammaticales, soit comme une succession de lemmes ; sans qu'aucun lien ne soit fait entre ces différentes vues. Par exemple, pour *Hyperbase*, le calcul des spécificités est fait d'une part sur les formes, d'autre part sur les lemmes, et le fait que les résultats soient donnés en parallèle ne permet pas pour autant de croiser les informations.

> Les calculs lexicométriques considèrent les textes comme une suite de positions, chacune de ces positions réalisant une occurrence d'une unité type. L'étiquetage s'intègre très bien à ce modèle en considérant que chaque position est multidimensionnelle, c'est-à-dire porteuse de plusieurs informations (graphie, lemme, code grammatical, etc.) [Pincemin 2004]

[Mayaffre 2006], partant d'un constat similaire, propose une première solution qui consisterait à étudier l'une des « dimensions » à la lumière d'une autre (c'est-à-dire en la pondérant par l'autre). Par exemple, la spécificité d'une lexie pourrait se calculer à par rapport à sa classe grammaticale. Cela rendrait visible des subtilités qui échappent aux spécificités traditionnelles, comme par exemple le fait que tel locuteur, qui utilise peu la catégorie verbale, sur-utilise néanmoins tel verbe.

### 1.3.3. Propositions de conciliation entre approches quantitative et qualitative pour l'étude de textes politiques

Dans le dernier numéro de la revue Corpus (2005), consacrée aux corpus politiques, plusieurs auteurs font part de propositions visant à concilier une approche quantitative et une approche qualitative pour le traitement de tels corpus ; il s'agit notamment de [Desmarchelier 2005] et [Rouveyrol 2005].

[Rouveyrol 2005] explique les principes d'une « logométrie intégrative » qui vise à quantifier certains traits essentiellement qualitatifs comme « [les] stratégies de positionnements énonciatifs ou bien le recours à des activités cognitivo-discursives telles que la narration ou l'argumentation » afin de leur appliquer des méthodes statistiques.

[Desmarchelier 2005] s'interroge sur la manière de concilier la lexicométrie et des « analyses sémantiques et argumentatives propres à la linguistique de l'énonciation ». D'après lui, une analyse qualitative doit être insérée entre deux phases quantitatives : une première analyse lexicométrique fournit « une vision globale du corpus », sur laquelle se basera l'analyste pour émettre des hypothèses argumentatives. Puis, à partir de ces hypothèses, le corpus est à nouveau « interrogé » avec des méthodes lexicométriques.

## 1.4. De la lexicométrie, du Web et des présidentielles…

### 1.3.1. Un même « lieu », le Web

Nous allons parler dans les lignes à venir d'outils lexicométriques développés en vue de suivre l'évolution des élections présidentielles de 2007, outils auxquels est venu se joindre, depuis janvier 2007, LexiMédia2007. Dans la mesure où ces outils partagent un même terrain de diffusion – la Toile –, et qu'un tel choix ne peut être considéré comme parfaitement anodin, il nous a semblé utile de dépeindre en quelques grandes lignes le décor dans lequel ils prennent place.

> Après les pouvoirs exécutif, législatif, judiciaire et médiatique, les citoyens fédérés grâce aux technologies de communication récentes forment un nouveau pouvoir : le cinquième pouvoir. Alors que les pessimistes se plaignent que rien ne change, ce sont les fondements de notre société eux-mêmes qui sont réinventés, à commencer par les règles qui régissent nos démocraties.

Voici l'enthousiaste mise en bouche, telle qu'elle nous est livrée en quatrième de couverture, du livre Le Cinquième pouvoir – ou « Comment Internet bouleverse la politique ? » – de Thierry Crouzet. Cet enthousiasme est de mise parmi les internautes, qui voient ce mouvement de l'intérieur, et y participent – Thierry Crouzet est lui aussi blogueur avant d'être écrivain. Mais il nous paraît difficile de mesurer l'impact réel d'Internet sur la vie politique.

Le Web est un milieu par trop mouvant pour que l'on puisse chiffrer l'ampleur de tels phénomènes.

Sans nous engager trop avant sur ce terrain glissant, nous nous contenterons donc de mentionner quelques faits qui ne sont pas contestés.

Il est indéniable que de plus en plus nombreux sont les Français qui consultent des blogs, y postent des commentaires ou créent leur propre blog.[8] Mais il faut préciser que les blogs politiques ne constituent qu'un aspect très minoritaire de cette imposante blogosphère. Pour ceux qui s'y adonnent, ces blogs sont perçus comme des lieux d'échange et de libre expression « citoyenne », ainsi que comme des moyens de diffusion hors pair.

En effet, Internet peut relayer très vite des informations, cela a été constaté à plusieurs reprises. Mais un tel phénomène n'atteint une véritable ampleur que lorsque d'autres médias s'emparent de l'information véhiculée par le Net pour la relayer à leur tour.

Il paraît par contre difficile d'affirmer qu'Internet peut avoir une influence réelle sur les résultats d'une élection. Nous signalons toutefois que des études ont été menées à ce sujet, visant à montrer le rôle d'Internet lors des élections présidentielles de 2004 aux Etats-Unis (*Cf.* [Adamic 2005]), ou lors du referendum sur la Constitution européenne en France en 2005 (*Cf.* [Fouetillou 2006]).

Les sites Internet offrant des informations à ce sujet, ou présentant des fonctionnalités intéressantes, sont extrêmement pléthoriques. Nous nous contenterons d'en mentionner deux.

### 1.3.2. L'Observatoire-Présidentielles 2007

Un premier site mérite qu'on lui accorde un moment d'attention : il s'agit de l'Observatoire-Présidentielles. Ce « lieu de veille et d'analyse du web consacré aux prochaines élections présidentielles » propose principalement deux outils aux internautes :

- Le *Tendançologue* utilise comme données les résultats (quantitatifs) fournis par différents moteurs de recherche proposant des sections *actualités en ligne*, *blogs* ou *newsgroup*. Il permet de mesurer le « bruit médiatique » généré par un candidat dans ces trois sphères.

- La *Blogopole* utilise des robots qui parcourent automatiquement la « blogosphère » en allant de lien hypertexte en lien hypertexte. Les sites répertoriés peuvent être visualisés sur une carte dans laquelle chaque blog est représenté par un nœud et chaque lien hypertexte par

[8] Une étude de Médiamétrie est disponible sur :
http://www.mediametrie.fr/resultats.php?rubrique=net&resultat_id=348

une flèche entre deux nœuds. Les nœuds ont une couleur qui dépend de l'appartenance politique du blog, et une taille proportionnelle au nombre de liens entrants.
Sur le blog associé au site, on trouve des liens intéressants, notamment vers un répertoire des blogs politiques ou vers un moteur de recherche qui leur est consacré (*Cf.* webographie).

### 1.3.3. Combat pour l'Élysée

Jean Véronis a développé divers outils, disponibles sur son blog, et les a regroupés sous les noms de « Presse 2007 » ou « Discours 2007 » selon les corpus qu'ils utilisent.
Les articles de *Presse 2007* sont automatiquement sélectionnés à partir des fils RSS de plusieurs grands journaux : Le Monde, le Figaro, Libération, l'Humanité, Les Echos, Marianne et le Parisien ; le critère de sélection est la présence d'au moins une citation, dans l'intégralité de l'article, d'un des présidentiables listés au préalable.
Les discours regroupés dans *Discours 2007* proviennent de différents sites sur lesquels ils étaient disponibles en format numérisé.
A partir de ce corpus, les outils proposés par Jean Véronis sont notamment :

- Un moteur de recherche permettant de trouver des articles en fonction du candidat, du journal, de la période et de mots-clés pour le corpus *Presse 2007*, ou d'afficher les concordances d'un mot dans les différents discours du corpus *Discours 2007* ;
- Des "palmarès" : graphiques permettant de visualiser au fil des semaines quels candidats ont été le plus cités (en pourcentages de citations) ;
- Des nuages de mots, faisant apparaître les mots (pleins) les plus employés dans les articles de la semaine, ou dans le corpus *Discours 2007* : les mots sont classés par ordre alphabétique, le mot comptant le plus d'occurrences apparaît en gras et en rouge, les suivants en orange, etc.

Nous avons très rapidement balayé les types d'applications qui sont mises à la disposition de tout un chacun sur Internet. Ces applications présentent l'avantage d'être facilement accessibles et de présenter leurs résultats de manière claire et imagée. On pourrait néanmoins regretter que les techniques utilisées restent limitées sur le plan du traitement linguistique. En effet, on a vu que le simple comptage des mots, s'il permet certes de quantifier certaines tendances, gagne à être dépassé.

## 1.5. LexiMédia 2007

### 1.5.1 Données

LexiMédia utilise pour données le flux RSS mis au point par Jean Véronis pour son outil Presse 2007. Pour LexiMédia, il a été décidé de ne conserver que les articles du Monde, du Figaro et de Libération. Inclure, par exemple, les articles de l'Humanité, comme cela avait été envisagé dans un premier temps, aurait faussé les résultats. En effet, ceux-ci étant beaucoup moins nombreux que les articles des autres journaux, la normalisation des fréquences aurait conduit à les sur-pondérer.

LexiMédia travaille donc sur trois journaux. Si les critères qui l'ont guidé paraissent en premier lieu pratiques (disponibilité de suffisamment de textes en format numérisé), ce choix ne s'en justifie pas moins d'un point de vue théorique. Ces trois quotidiens nationaux sont parmi les plus lus en France, et jouissent d'un certain prestige dû à leurs histoires ; Le Monde apparaît généralement comme la référence en matière d'information ; Le Figaro et Libération assurent la diversité des orientations politiques représentées dans le corpus, le premier étant plutôt de droite et le second plutôt de gauche (selon le spectre politique français tel qu'il est communément défini).

### 1.5.2. Méthodes

#### *1.5.2.1. Analyse syntaxique par SYNTEX*

Le corpus sur lequel travaille LexiMédia2007 est dans un premier temps traité par SYNTEX, un analyseur syntaxique de corpus développé par Didier Bourigault ([Bourigault & Fabre 2000], [Bourigault et al. 2005]). C'est l'utilisation de SYNTEX en amont de tout traitement statistique qui fait l'originalité principale de LexiMédia2007. En effet, les unités de l'analyse statistique ne sont alors plus des mots ou des successions de mots, mais des syntagmes, unités linguistiques opérantes.

Après un pré-étiquetage basé sur des règles et des lexiques le corpus est confié à l'étiqueteur TreeTagger (Université de Stuttgart) ; puis l'analyse en dépendance est effectuée par une série de modules prenant chacun en charge une relation syntaxique (comme par exemple la relation sujet) qui reliera des unités régies à des unités rectrices. Pour résoudre les ambiguïtés de rattachement, SYNTEX utilise des procédures d'apprentissage endogène, mais aussi un lexique de probabilités de sous-catégorisation.

#### *1.5.2.2. Extraction des syntagmes*

A partir des résultats des relations de dépendances identifiées par SYNTEX, un module d'extraction de syntagmes, UPERY ([Bourigault 2002]), construit le réseau de mots et de syntagmes qui sera utilisé par la suite par LexiMédia2007. Cette étape est importante pour comprendre pourquoi les syntagmes analysés par LexiMédia ne sont pas forcément constitués de mots contigus dans les textes. En effet, le module d'extraction de syntagmes, après avoir identifié des syntagmes maximaux et les avoir agencés dans un réseau de dépendance, enrichit ce réseau grâce à des opérations de réduction et de simplification qui permettent de générer de nouveaux syntagmes. Par exemple, à partir du syntagme maximal « organisations islamiques de France », l'opération de réduction génèrera les syntagmes « organisations islamiques » et « organisation de France » en associant à la tête du syntagme chacune de ses expansions. Ou encore, à partir du syntagme maximal « position électorale de la candidate », l'opération de simplification, qui remplace les expansions syntagmes par leur tête, génèrera le syntagme « position de la candidate ».

#### *1.5.2.3. Sélection des syntagmes*

Les syntagmes et mots obtenus sont donc très nombreux, et tous ne seront bien sûr pas affichés par LexiMédia2007 ; des critères de sélection ont été mis en place.

Un syntagme ou mot n'est pris en compte que s'il dépasse une certaine fréquence (fréquence relative 10) sur l'ensemble du corpus. Ainsi, chaque semaine, des syntagmes qui existaient précédemment mais qui étaient sous-représentés peuvent dépasser le seuil fixé ; leur évolution est alors retracée depuis la première semaine. Par contre, aucun syntagme ne peut « disparaître ». Le nombre des syntagmes affichés par LexiMédia2007 va donc croissant, et le seuil est de moins en moins astreignant.

Les syntagmes et mots sont également classés par catégories, et seules certaines de ces catégories sont disponibles pour l'utilisateur (*Cf.* 2.3.1.).

La normalisation des fréquences se fait selon des coefficients recalculés chaque semaine en fonction l'évolution de la taille du corpus (et des tailles comparées des sous-corpus *Le Monde*, *Le Figaro* et *Libération*). Chaque fréquence relative affichée pour une semaine donnée a été calculée avec les coefficients de cette semaine.

### 1.5.3. Résultats

Les résultats sont triés par catégories : syntagmes nominaux ayant pour tête un nom commun, syntagmes nominaux ayant pour tête un nom propre, syntagmes verbaux, noms, adjectifs,

verbes. Les éléments extraits par SYNTEX et appartenant à d'autres catégories (par exemple les syntagmes adverbiaux) ont donc été exclus. L'utilisateur peut choisir d'afficher les mots ou expressions appartenant à l'une des catégories, ou de les afficher toutes catégories confondues. Il peut aussi choisir différents critères de tri pour cet affichage : tri par variation (positive ou négative), par augmentation, par diminution, ou par fréquence.

Un moteur de recherche permet de chercher une chaîne (un mot ou plusieurs mots contigus) dans le corpus, avec utilisation possible du joker « * ». En résultat, les différents syntagmes correspondant à la requête sont affichés avec leur fréquence absolue, leur fréquence relative et un graphique montrant leur évolution à travers les semaines.

A partir de l'affichage par tri ou de l'affichage correspondant à la réponse du moteur de recherche, il est possible de cliquer sur les différents syntagmes. On obtient alors, pour le syntagme considéré, son évolution comparée (une courbe par journal) à travers toutes les semaines sous forme de graphique et de tableau, ainsi que les liens vers les originaux des articles concernés, classés par semaine (le second critère de tri étant le nombre d'occurrences).

Ainsi, de nombreuses facilités sont offertes à l'utilisateur, afin de mettre à sa portée des données pertinentes dont il pourra tirer parti selon l'analyse qu'il veut mener. L'interprétation des résultats obtenus n'est plus de notre ressort, même si, bien sûr, sélectionner ce qui sera pertinent pour l'analyse, c'est déjà commencer à analyser.

### 1.5.4. Observation des phénomènes de coréférence à partir de LexiMedia2007

Lorsque l'on observe l'interface de LexiMedia2007, on constate qu'une grande partie des syntagmes présentés (donc des syntagmes les plus fréquents du corpus) sont référentiels à des personnes. C'est le cas bien sûr de la plupart des noms propres (sur cent syntagmes affichés, tous sauf deux ou trois sont des noms de personnes), mais aussi d'une large partie des noms communs (la moitié environ). Parmi ces syntagmes référentiels à des personnes, on observe beaucoup de phénomènes de coréférence : c'est-à-dire que plusieurs syntagmes désignent la même personne. En effet, pour citer une même personne, on peu faire appel soit à son nom, soit, pour présenter cette personne selon un certain jour ou simplement pour éviter les répétitions, à sa fonction. Noms et fonctions peuvent eux-mêmes être déclinés de plusieurs manières (*prénom + nom*, *M. +nom*, etc.).

Les syntagmes coréférentiels forment ensemble des classes qu'on voudrait mettre au jour. Is subissent par exemple les mêmes variations, comme on le voit avec l'exemple de « candidat

de l'UDF », dont la courbe, malgré la moindre fréquence de ce syntagme, est très similaire à celle de « François Bayrou ».

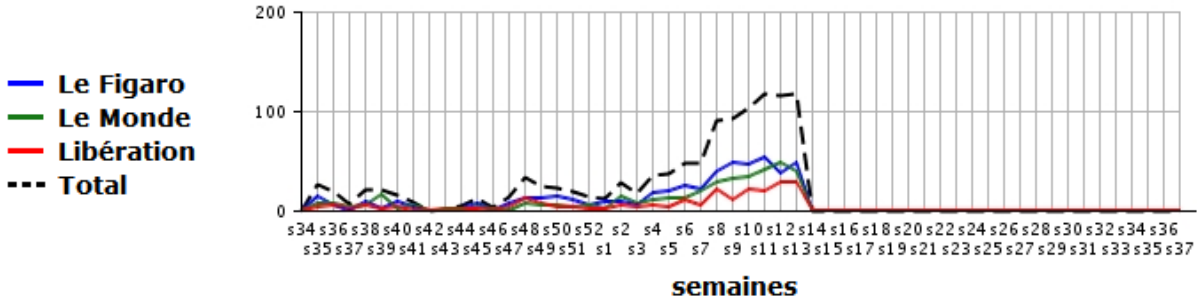



Figure 1 : Courbe du syntagme « François Bayrou »

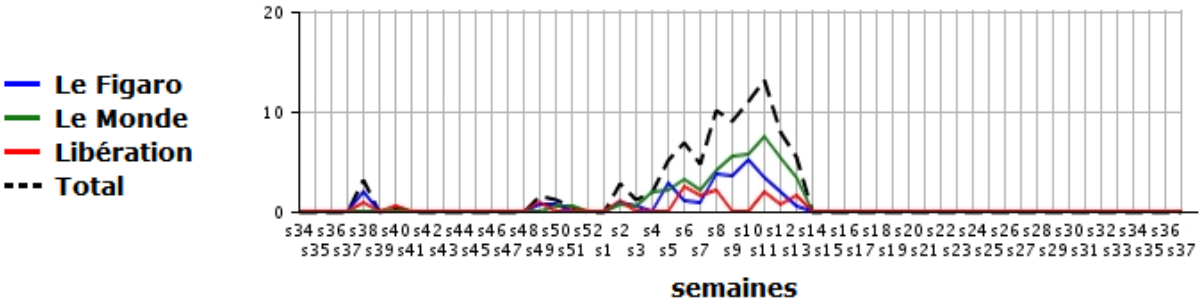
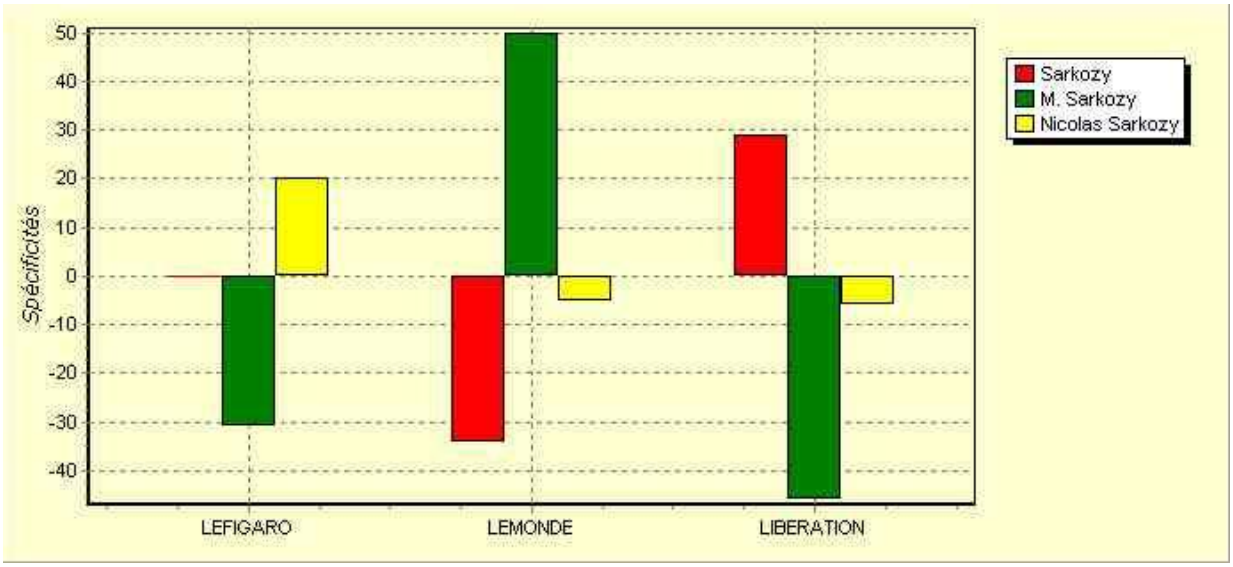



Figure 2 : Courbe du syntagme « candidat de l'UDF »

Il serait peut-être intéressant, si cela est possible, de regrouper dans une même unité les différentes manières de citer un candidat – Nom, M. (ou Mme) Nom, Prénom Nom, fonction – en cumulant leurs fréquences. En effet, il paraît dommage de tracer deux évolutions différentes pour quelque chose qui réfère à une même réalité extralinguistique : cela fausse l'interprétation puisque le nombre total de citations du candidat X sera divisé entre ses différentes variantes. De plus, les noms propres exprimés par un seul mot graphique (par exemple « Royal ») sont pour l'instant traités séparément puisqu'ils ne constituent pas des syntagmes.

Ajoutons à cela que l'on veut étudier l'évolution des expressions considérées tous journaux confondus, ou en comparant explicitement les différents journaux. Or, chaque journal a sa manière privilégiée de citer les personnes : Le graphique ci-dessous montre les spécificités (établies par Lexico) de « Sarkozy », « M. Sarkozy » et « Nicolas Sarkozy » dans le corpus préalablement partitionné par journaux.




Ainsi, en séparant « Nicolas Sarkozy », « Sarkozy » et « M. Sarkozy », on est pour l'instant amené à traiter séparément les informations selon le journal dont elles proviennent. Si au contraire on regroupait tous les syntagmes référant à Nicolas Sarkozy, on pourrait à la fois calculer une fréquence cumulée des différents syntagmes et rendre visible leur répartition selon les journaux – ce qui manifestement ne serait pas sans pertinence.

Avoir accès aux différentes façons de citer un candidat, voir lesquelles sont privilégiées selon l'époque ou le journal, constituerait un point de vue intéressant pour l'analyse.

Nous avons donc, dans la suite de notre travail, tenté d'extraire automatiquement ces « classes » de syntagmes coréférentiels entre eux, en résolvant les liens de coréférence au sein du corpus.

# 2. Référence, coréférence : notions

> *"I don't rejoice in insects at all," Alice explained, "because I'm rather afraid of them – at least the large kinds. But I can tell you the names of some of them."*
> *"Of course they answer to their names?" the Gnat remarked carelessly.*
> *"I never knew them do it."*
> *"What's the use of their having names," the Gnat said, "if they won't answer to them?"*
> *"No use to them," said Alice; "but it's useful to the people that name them, I suppose. If not, why do things have names at all?"*
>
> *Lewis Carroll*

> La référence est un sujet qui met en cause rien moins que les relations entre le langage, la pensée et la réalité. Toute perturbation à même de brouiller les conduites référentielles a des conséquences immédiates et cruciales sur la communication. Dieu ne s'y est du reste pas trompé puisque pour empêcher les hommes d'achever la Tour de Babel destinée à supporter une idole, l'Ancien Testament explique qu'il a tout simplement perverti l'usage des expressions référentielles : « quand l'un demandait une brique, l'autre lui apportait du mortier. Et celui qui voulait de l'eau recevait une truelle ». [Charolles 2002 : 241]

La référence est au cœur du rapport entre les objets réels, l'idée qu'on s'en fait et le nom qu'on leur donne ; nous verrons les problèmes que pose cette notion charnière aux linguistes comme aux philosophes du langage (2.1.).

Comme le fait remarquer [Charolles 2002] à propos du mythe de la Tour de Babel, Dieu, pour arriver à ses fins, n'a pas jugé bon de s'attaquer à autre chose qu'aux noms ; nous expliquerons pourquoi nous aussi, dans ce mémoire, ne nous intéressons qu'à la référence nominale, et examinerons comment la référence peut être réalisée par les différents types de syntagmes nominaux (définis, indéfinis, démonstratifs), ainsi que par les noms propres (2.2.).

Nous nous intéresserons ensuite plus particulièrement à la relation de coréférence, qui lie deux expressions référant à une même entité (2.3.), et à son exploitation dans le domaine du TAL (2.4.) : que peut-elle apporter aux diverses applications de traitement automatique des langues ? (2.4.1.) Quelles méthodes permettent de résoudre les liens coréférentiels dans un texte ? (2.4.2)

## 2.1. De la référence, de la linguistique et de la philosophie du langage

### 2.1.1. Linguistique et référence

La linguistique doit-elle frayer avec la référence ? Non, affirmait Saussure, pour qui « la langue est un tout en soi » [Saussure 1916 : 25] duquel il faut écarter « tout ce qui est étranger à son organisme, à son système, en un mot tout ce qu'on désigne par le terme de "linguistique externe" » [*ibid.* : 40]. Ainsi, le signe linguistique est défini comme unissant « non une chose et un nom, mais un concept et une image acoustique » [*ibid.* : 98], désignés respectivement par les termes de « signifié » et de « signifiant » [*ibid.* : 99] ; une caractéristique primordiale

du signe linguistique est que « le lien unissant le signifiant au signifié est arbitraire » [*Ibid.* : 100-101]. Saussure a ainsi une conception très pure de la langue, basée sur l'arbitraire du signe ([Engler 1968], entre autres, note que ce principe est la « condition essentielle de la conception de la langue comme pure forme »). Ce que [Siblot 1997 : 3] résume ainsi :

> Longtemps le domaine des études linguistiques a paru délimité avec clarté et précision, comme tracé au cordeau. Saussure avait avec la *langue* donné à la linguistique l'objet théorique qu'il lui manquait pour qu'elle puisse devenir « moderne ». Il en avait séparé le bon grain de l'ivraie de la parole.

Mais en délimitant si nettement le champ de la linguistique, Saussure commet une erreur qui fut maintes fois relevée, par [Pichon 1938 : 27 cité par Siblot 1997 : 4] le premier :

> L'erreur de Saussure est à mon sens éclatante. Elle consiste en ce qu'il ne s'aperçoit pas qu'il introduit en cours de démonstration des éléments qui n'étaient pas dans l'énoncé. Il définit d'abord le signifié comme étant l'*idée* générale de bœuf : il se comporte ensuite comme si ce signifié était l'*objet* appelé bœuf ou du moins l'image sensorielle d'un bœuf.

puis, surtout, par [Benveniste 1939 : 50] :

> Il est clair que le raisonnement est faussé par le recours inconscient et subreptice à un troisième terme, qui n'était pas compris dans la définition initiale. Ce troisième terme est la chose même, la réalité. Saussure a beau dire que l'idée de sœur n'est pas liée au signifiant *s-ö-r* ; il n'en pense pas moins à la *réalité* de la notion. Quand il parle de la différence entre *b-ö-f* et *o-k-s*, il se réfère malgré lui au fait que ces deux termes s'appliquent à la même *réalité*. Voilà donc la *chose*, expressément exclue d'abord de la définition du signe, qui s'y introduit par un détour et qui y installe en permanence la contradiction.

Dans la suite de son article, Benveniste pose que la relation entre le signifiant et le signifié est non pas arbitraire, mais nécessaire, et déplace l'arbitraire à la relation entre signe et référent :

> Il apparaît donc que la part de contingence inhérente à la langue affecte la dénomination en tant que symbole phonique de la réalité et dans son rapport avec elle. [*Ibid.* : 55]

Finalement, le réel, la *chose* ne semblent pas pouvoir être exclus du champ de la linguistique, contrairement à ce que le terme « *extra*linguistique » nous suggérait ; du moment que l'on accepte que parler, c'est dire *quelque chose*, on est amené à se pencher sur la référence, qui est la relation qui unit des expressions linguistiques à ce *quelque chose*, qui fait le pont entre le linguistique et l'extralinguistique.

Les problèmes découlant de la référence, du fait de sa position charnière, ne sauraient donc être abandonnés à la seule philosophie du langage ; se pencher sur eux permet également de mieux cerner certaines notions d'un point de vue linguistique. Ces problèmes sont posés depuis longtemps : ils apparaissent déjà dans certains dialogues de Platon[9] ; puis chez les logiciens du moyen âge ; ils ont beaucoup préoccupé par la suite Meinong, Frege, puis Russell, qui expose dans son article *On Denoting* une série d'énigmes liées à la référence,

[9] Notamment dans le Parménide (avec la théorie des formes), dans le Sophiste (dialogue entre Théétète et l'étranger) et dans le Livre VII de la République (avec le mythe de la caverne).

qu'il tente de résoudre à la lumière de la théorie qu'il présente. On pourrait résumer ces problèmes à dimension philosophique en deux questions principales : celle du rapport entre référence et existence et celle du rapport entre référence et sens.

Nous allons examiner ces deux questions rapidement, en essayant de nous cantonner à ce qui concerne le champ de la linguistique et de montrer en quoi ils peuvent toucher à notre problématique.

### 2.1.2. Référence et existence

*… « reality » – one of the few words which mean nothing without quotes.*
*Vladimir Nabokov*

Admettre que des expressions linguistiques puissent référer à « quelque chose », c'est du même coup admettre que ce « quelque chose » existe.[10] Mais se pose alors la question du « type » d'existence des référents :

- Sont-ils des entités du monde réel, indépendantes du langage et que les expressions référentielles permettent de désigner précisément (c'est la thèse objectiviste : le monde préexiste au discours) ? Cette conception se heurte à des énoncés mettant en jeu des entités fictives : peut-on référer à quelque chose qui n'existe *réellement* pas ? Cela constitue l'une des énigmes de Russell, résumée par [Linsky 1967 : 20] par le paradoxe suivant :

  > Soit la proposition « Pégase n'existe pas » qui, pensons-nous, est à la fois, vraie et relative à Pégase. D'une part, il semble que si la proposition est vraie, elle ne peut traiter de Pégase, car elle dit qu'une chose telle que Pégase n'est pas. D'autre part, il semble que si la proposition traite de Pégase, elle ne peut être vraie, car, dans ce cas, il *y a* quelque chose dont elle traite, à savoir Pégase. Donc la proposition doit être fausse, qui dit que Pégase n'existe pas. Ces difficultés nous obligent, semble-t-il, à admettre le contraire de ce que nous croyons tous : à admettre que Pégase doit avoir un mode d'être tel qu'il nous soit permis de nier son existence.

  Pour résoudre ce problème, il faut étendre la référence aux mondes imaginaires ou possibles : elle serait alors, selon la définition de [Dubois 1973 : 414 cité par Kleiber 1997 : 11] « la fonction par laquelle un signe linguistique renvoie à un objet du monde extralinguistique, réel ou imaginaire ». Selon [Kleiber 1997 : 11], cet élargissement de la définition de la référence permet certes de recouvrir « toutes les situations qui peuvent se rencontrer », mais il s'accompagne d'un « effet pernicieux » : celui de « remettre sérieusement en cause le réalisme objectiviste sur lequel s'appuie cette conception standard de la référence ».

[10] « Tout ce à quoi on réfère doit exister. », selon l' « axiome d'existence » de [Searle 1972 : 121].

- De la possibilité de renvoyer à des entités non-existantes, d'aucuns ont déduit que les référents étaient en fait construits par le discours et que, finalement, « le monde réel n'est pas aussi réel que ça et qu'il n'est qu'un univers construit » [*Ibid.* : 11]. C'est là la thèse constructiviste, selon laquelle le monde ne préexisterait pas au discours. Nous ne pourrions pas dire, selon cette thèse, le monde tel qu'il est en soi (puisque nous n'y avons pas accès), mais uniquement un monde perçu, façonné par notre perception.

Kleiber invite à modérer cette conception qui amène à considérer que l'objectivité serait un concept illusoire, et qui substitue à la référence externe une référence interne puisque selon elle, les expressions référentielles ne renverraient finalement qu'à « des entités discursives, à des constructions mentales, à des représentations élaborées par le discours, qui n'ont de validité que dans et par le discours » [*Ibid.* : 16]. Il observe notamment que si **toute** réalité n'est que projection ou construction, alors « l'affaire n'est pas trop grave, parce que tout élément de la réalité se trouvant soumis à une telle conceptualisation, on peut admettre assez sereinement que ce que nous croyons être le monde réel n'est que le monde tel que nous le percevons ou tel que nous croyons qu'il est » [*Ibid.* : 12] ; il est alors superflu de préciser, à chaque fois que l'on parle de réalité, que cette réalité n'est pas la vraie réalité ou réalité objective, mais seulement la réalité expériencée ou réalité phénoménologique. Cela se justifie d'autant plus que si cette réalité n'est pas objective, elle bénéficie tout au moins d'une stabilité intersubjective, qui est d'ailleurs à l'origine du sentiment d'objectivité.

> On peut en conséquence maintenir, sans éprouver un prurit métaphysique trop insupportable, que les expressions linguistiques, si elles réfèrent, réfèrent à des éléments "existants", réels ou fictifs, c'est-à-dire conçus comme existant en dehors du langage : cette existence leur est garantie par cette modélisation intersubjective stable à apparence d'objectivité qui caractérise notre appréhension du monde. [*Ibid.* : 17]

### 2.1.3. Référence et sens

C'est [Frege 1892 : 102-107] qui a imposé la distinction entre *sinn* (sens) et *bedeutung* (référence ou dénotation) dans le but de résoudre le problème de l'identité :

> La relation d'égalité propose à la réflexion quelques questions irrémissibles, auxquelles il n'est pas aisé de répondre. Est-ce une relation ? Une relation entre des objets, ou des noms ou signes d'objets ? (…) Si l'on voulait voir dans l'égalité une relation entre ce que dénotent respectivement les noms « a » et « b », a = b ne pourrait pas, semble-t-il, différer de a = a, à supposer que la proposition a = b soit vraie.

Or, l'énoncé a = b contient une information que ne contient pas l'énoncé a = a ; dire que « l'étoile du matin est l'étoile du soir » ne revient pas à dire que « l'étoile du matin est l'étoile du matin » ; la première proposition constitue une découverte scientifique d'importance, tandis que la seconde ne nous apprend rien. Mais selon une conception

purement référentialiste de la signification telle qu'elle était alors admise, la contribution des deux expressions « étoile du matin » et « étoile du soir » à la signification de l'énoncé devrait être exactement la même puisqu'elles ont le même référent : la planète Vénus. Pour rendre compte de la différence qui existe entre un énoncé de type a = b et un énoncé de type a = a, Frege se sert de la notion de sens (*sinn*) : le sens d'un terme permet de déterminer sa référence, mais il ne s'identifie pas avec cette dernière : il correspond au « mode de donation » de la référence. On le voit, séparer le sens et la référence est nécessaire dès lors que l'on veut rendre compte de la notion de coréférence ; cette distinction sera donc importante pour nous.

Le couple *Sinn-Bedeutung* a joui d'une grande postérité philosophique, mais il n'est pas sans précédents : les logiciens médiévaux distinguaient la *significatio* (sens) et la *suppositio* (référence) ; John Stuart Mill, la connotation et la dénotation ; les grammairiens de Port-Royal, la compréhension et l'étendue, qui ont par après été repris sous les termes intension et extension. Nous utiliserons quant à nous la terminologie de Jean-Claude Milner qui emploie les termes de « référence virtuelle » (qui correspond au sens : « la référence virtuelle d'une unité est bien ce que tente de représenter la définition du dictionnaire » [Milner 1982 : 10]) et de « référence actuelle », ce qui présente l'avantage de mettre en évidence une conception du sens comme étant tourné vers la référence, conditionnant la référence.

> Le segment de réalité associé à une séquence est sa *référence actuelle* ; l'ensemble de conditions caractérisant une unité lexicale est sa *référence virtuelle*. Il ne me semble pas déraisonnable de rapprocher la référence actuelle de la *Bedeutung* et la référence virtuelle du *Sinn*, au sens de Frege. Mais là où la terminologie de Frege oppose deux entités radicalement distinctes et étrangères l'une de l'autre, les termes *virtuel* et *actuel* impliquent une relation systématique. [*Ibid.*]

Une expression linguistique ne peut avoir de référence actuelle que si elle est employée ; hors emploi, elle ne peut comporter que « les conditions d'une éventuelle référence actuelle, c'est-à-dire sa référence virtuelle » [*Ibid.*].

## 2.2. Les moyens linguistiques de la référence

A présent que nous avons mieux cerné la notion de référence, nous allons voir comment elle est réalisée par différents types d'expressions linguistiques.

### 2.2.1. Nature linguistique des expressions référentielles

Classiquement, on a nié que des termes autres que nominaux (ou, bien sûr, pronominaux) puissent avoir la propriété de référer ; puis on a considéré que les noms étaient, du moins, le lieu privilégié de la référence. [Asnes 2004], qui étudie les analogies et interactions entre la

référence verbale et la référence nominale, remarque que de plus en plus, les études actuelles « présupposent non seulement une force référentielle aux unités linguistiques autres que nominales, mais affirment même que la nature de la référence de ces termes est à plusieurs égards comparable à celle des substantifs. » On ne conteste plus aujourd'hui que verbes ou adjectifs aient la propriété de référer.

Le fait que les noms nous paraissent "plus référentiels" que les éléments des autres catégories linguistiques peut s'expliquer à la lumière de la notion de classe référentielle, introduite par [Kleiber 1981]. La classe référentielle est définie comme « l'ensemble conceptuel de tous les particuliers réunis sous l'étiquette de l'item lexical » ; par exemple, la classe référentielle du mot « cheval » est constituée par l'ensemble de tous les chevaux, et la classe référentielle du mot « blanc » est constituée par tout ce qui a la qualité d'être blanc (exemple emprunté à [Boudreau 2004]). On le voit, les éléments de cette seconde classe sont relativement disparates ; c'est la plus grande homogénéité des classes référentielles représentées par les noms qui conférerait à ces derniers un statut particulier dans le domaine de la référence.

Dans ce mémoire, nous ne traiterons que de la référence nominale, dans la mesure où nous nous intéressons uniquement à la référence aux personnes, et que celle-ci est réalisée par les noms ou syntagmes nominaux, et notamment, bien sûr, par les noms propres. Nous allons donc exposer dans les parties qui suivent le fonctionnement référentiel des noms propres et de différents types de syntagmes nominaux : définis, indéfinis et démonstratifs. Nous nous cantonnerons aux emplois spécifiques de ces syntagmes nominaux, et aux cas où ils réfèrent à des entités singulières, concrètes et comptables, puisque nous ne traitons dans notre outil que de la référence singulière aux personnes (qui sont des entités concrètes et comptables), laissant de côté la référence plurielle que l'observe par exemple dans une phrase telle que :

> Ainsi se retrouveront au Parc des expositions de la capitale sarthoise **les candidats à la candidature antilibérale pour 2007**, Marie-George Buffet, secrétaire nationale du PCF, le leader altermondialiste José Bové, Clémentine Autain, adjointe (apparentée PCF) à la mairie de Paris, Patrick Braouezec, député (PCF) de Seine-Saint-Denis et Yves Salesse, président «en congé» de la Fondation Copernic.

avec le syntagme « les candidats à la candidature antilibérale pour 2007 ». Nous ne traiterons donc ni des emplois génériques des syntagmes nominaux, ni du fonctionnement référentiel des pronoms.

### 2.2.2. Les noms propres

La caractéristique première du nom propre est d'être référentiel, au point qu'on a pu nier qu'il ait d'autres propriétés – selon les termes de John Stuart Mill, les noms propres n'ont que la propriété de dénoter, ils ne connotent rien ([Mill 1724 : 33-37]). Cette thèse selon laquelle les

noms propres n'auraient pas de signification a été contestée par la suite, notamment par [Frege 1892 : 102-107] et Russell qui s'appuient sur des énoncés d'identité comme :

*Hespérus est Phosphorus* (Frege)

qui n'auraient aucune valeur informative si l'on n'accepte pas que les noms propres ont un sens (*Cf.* 2.1.3. Référence et sens). Quoi qu'il en soit, on peut admettre que le nom propre a un statut particulier par rapport aux noms communs dans le domaine de la référence.

Les noms propres se distinguent également des noms communs par certaines caractéristiques formelles simples : ils s'emploient généralement sans déterminant (surtout les noms propres de personne), ils sont pauvres morphologiquement (ils ne marquent pas le genre ni le nombre), ils prennent une majuscule à l'écrit.

Les noms propres peuvent référer à tous types d'entités **particulières** du moment que locuteur et destinataires sont en accord sur les conventions dénominatives utilisées. Les procédures d'accointance (cognitivement et socialement complexes) permettent de sceller la convention de dénomination (Cf. [Charolles 2002 : 61-64]).

La convention qui lie les noms propres à leurs porteurs est indépendante des caractéristiques substantielles de ces derniers ; c'est pourquoi les noms propres peuvent continuer à référer à leurs porteurs quels que soient les changements qui affectent ces derniers : ce sont des désignateurs rigides, qui sont en association directe et durable avec leurs référents. La notion de désignateur rigide a été introduite par [Kripke 1980 : 49] :

> Let's call something a rigid designator if in every possible world it designates the same object, a nonrigid or accidental designator if that is not the case. (...) Proper names are rigid designators.

Ainsi, au cours de l'année 2007, le syntagme « le ministre de l'Intérieur » a cessé de désigner *Nicolas Sarkozy*, alors que ce dernier pourra toujours être désigné par son nom. Le nom propre permet de référer à un particulier dans toutes les phases de son existence, et dans ce qui constitue son essence individuelle, alors que les descriptions n'exploitent jamais que des traits accidentels du particulier désigné.

Dans le domaine des chaînes de références (*Cf.* 2.3.2.), le nom propre a également un statut particulier, puisqu'il permet de hiérarchiser les référents (*Cf.* [Schneidecker 1997 : 49]).

### 2.2.3. Les syntagmes nominaux

Les syntagmes nominaux, contrairement aux noms propres et aux pronoms, contiennent des informations sur les entités qu'ils désignent ; c'est pour cela que l'on parlera également de « descriptions définies / indéfinies / démonstratives » – expression introduite par Russell et encore utilisée de nos jours.

Nous avons précisé plus haut que la référence actuelle n'est possible qu'en emploi. Notons aussi que, toujours selon Milner, si l'on considère les emplois, seuls les syntagmes pris dans leur entier peuvent être dotés d'une référence actuelle.

> Ce ne sont pas aux unités lexicales comme telles que sont associés les segments de réalité, mais bien aux groupes nominaux pris dans leur ensemble. Dans ces groupes, plusieurs unités lexicales peuvent intervenir, et les références virtuelles de chacune se combinent pour contraindre une référence actuelle possible ; mais une référence actuelle donnée n'est associée qu'à la combinaison d'ensemble et non pas à chacune des unités combinées. [Milner 1982 : 11]

#### *2.2.3.1. Les syntagmes nominaux définis*

Les syntagmes nominaux définis (ou descriptions définies) sont les syntagmes de type *le N*. La grande majorité des syntagmes que nous aurons à traiter avec notre outil sont de ce type. Une distinction importante pour nous est faite entre les descriptions définies complètes et incomplètes (Cf. [Charolles 202 : 75-83]).

Les descriptions définies complètes sont des syntagmes comme « l'auteur de Crime et Châtiment », qui ne peuvent désigner qu'un seul référent (ici : Dostoïevski) ; elles sont caractérisées par leur autonomie référentielle : elles peuvent désigner un référent indépendamment du contexte. Dans notre corpus, on trouve un grand nombre de descriptions définies qui peuvent être traitées comme complètes, mais seulement à condition de les rapporter au moment de l'énonciation – c'est le cas de toutes les fonctions occupées par une seule personne à un moment donné. Ainsi, dans la mesure où notre corpus couvre une période de temps ne dépassant pas le mois d'avril 2007, le syntagme « ministre de l'Intérieur » peut être considéré comme étant complet, puisqu'il ne peut désigner que l'entité *Nicolas Sarkozy*. Le fonctionnement référentiel des descriptions définies complètes peut être rapproché de celui des désignateurs rigides.

Les descriptions définies incomplètes sont des syntagmes comme « le ministre », qui, bien qu'ils désignent eux aussi une entité particulière – pour un emploi donné – peuvent potentiellement s'appliquer à plusieurs entités ; l'établissement de la référence ne peut ici se faire qu'en contexte.

Ces deux types de descriptions définies, complètes et incomplètes, nécessiteront un traitement différent, comme nous le verrons dans la description de notre outil : si les descriptions complètes peuvent être résolues quasiment sans recours à des indices contextuels, ce ne sera pas (ou moins) le cas des descriptions incomplètes.

### *2.2.3.2. Les syntagmes nominaux indéfinis*

Les syntagmes nominaux indéfinis sont les syntagmes de type *un N* – et, plus généralement, tous les syntagmes nominaux préfixés par un déterminant n'étant ni défini ou possessif, ni démonstratif : *deux, trois, plusieurs, aucun, quelques, chaque* (etc.) *N.* Puisque comme nous l'avons déjà mentionné, nous ne nous intéressons qu'à la référence à des entités singulières et comptables, nous ne traiterons que du premier cas.

Dans notre corpus (et de manière générale), les syntagmes nominaux indéfinis sont souvent employés de manière générique (et donc non-spécifiques), comme dans les phrases :

> **Un ministre ou ancien ministre** ne peut être poursuivi que devant la Cour de justice de la République pour des actes accomplis dans le cadre de ses fonctions.
>
> **Un ministre de l'intérieur** doit sécuriser, pas ajouter au désordre.

et :

> Pour juger **un ministre de l'intérieur** sur son bilan, il existe deux critères : ce qu'il a fait et ce qu'il a fait savoir.

Dans ces deux dernières phrases, on ne fait pas référence à un ministre de l'Intérieur en particulier (et donc pas à Nicolas Sarkozy), mais à la fonction de ministre de l'Intérieur en général. Les emplois génériques sont eux aussi référentiels, mais ils s'opposent aux emplois spécifiques en cela qu'ils réfèrent à toute une classe référentielle et non à un (ou des) individu(s) la composant.

Le fonctionnement référentiel des syntagmes nominaux indéfinis est très différent de celui des syntagmes nominaux définis. Un syntagme nominal indéfini pris seul ne peut désigner qu'une classe référentielle (référence générique) ou un individu quelconque de cette classe référentielle (référence spécifique indéterminée) ; c'est la suite de l'énoncé auquel il participe qui permettra l'établissement – ou non – d'une référence spécifique spécifiée. Il n'est donc pas, contrairement aux syntagmes définis et démonstratifs, un *désignateur*, puisqu'il ne permet pas d'isoler un objet (*Cf.* [Corblin 1987 : 196 -197]. Prenons l'exemple du syntagme « un maire ». Dans la phrase :

> **Un maire** renvoie un chèque au responsable des parrainages du FN.

l'emploi est bien spécifique : on parle d'un individu particulier. Mais la fixation du référent ne se fera qu'avec la phrase suivante :

> Roger Lechevalier, élu âgé de 70 ans, s'était vu remettre le chèque à titre personnel et pour l'Association des amis du patrimoine de la commune, après avoir refusé d'accorder son parrainage à Jean-Marie Le Pen. Il l'avait pourtant fait en 2002, ce qui lui avait rapporté quelque 5.000 francs.

Ainsi, les syntagmes nominaux indéfinis sont généralement utilisés pour introduire un nouveau référent par sa classe référentielle. Intuitivement, on imagine qu'ils seront donc plutôt utilisés pour désigner des personnes peu connues des destinataires (lecteurs des journaux de notre corpus). Cependant, nous avons noté la présence de syntagmes indéfinis faisant référence à certaines des personnes les plus citées de notre corpus, comme dans les phrases suivantes :

> Enfin, cette législature finissante se distingue par l'omniprésence d'**un ministre de l'Intérieur** lancé depuis des mois dans la course à l'Elysée. C'est un homme qui a pris la direction de la justice en initiant des réformes de procédure pénale qui ont échoué sur Outreau, qui n'hésite pas à multiplier les pressions publiques sur les juges, qui s'est servi de la justice dans l'affaire Clearstream pour lutter contre le Premier ministre.

> **Un certain ministre de l'Intérieur et de l'Aménagement du territoire**, actuellement en campagne présidentielle, n'est peut-être pas pour rien dans ce dénouement.

La visée est ici rhétorique ; l'auteur semble laisser dans le flou la personne désignée en employant l'indéfini, mais il est de notoriété publique (du moins pour les destinataires visés) que la classe des « ministres de l'Intérieur » ne contient qu'un seul individu (Nicolas Sarkozy) ; le fonctionnement référentiel du syntagme sera donc ici proche de celui d'un syntagme défini (il n'est pas nécessaire de fixer le référent ou de l'extraire de sa classe en le spécifiant dans la suite de l'énoncé), le choix de l'indéfini résulte essentiellement de la recherche d'un effet plaisant ou de complicité avec le lecteur.

Du point de vue du traitement informatique, la résolution des syntagmes indéfinis demandera un recours au contexte plus important que celle des syntagmes définis ; et surtout, il ne faudra pas chercher à résoudre absolument tous les syntagmes indéfinis considérés comme référentiels, car bien souvent la référence reste sous-spécifiée ou indéterminée, comme cela sera, on s'en doute, le cas pour le ministre de la phrase suivante :

> «Il faut l'aider à se prendre les pieds dans le tapis, c'est une œuvre de salut public», estimait récemment **un ministre** sous couvert d'anonymat.

La spécification du référent sera donc recherchée dans une fenêtre réduite, hors de laquelle le syntagme référentiel ne sera pas attribué.

#### *2.2.3.3. Les syntagmes nominaux démonstratifs*

Les syntagmes nominaux démonstratifs sont les syntagmes nominaux de la forme *ce N*. Ils sont assez proches des définis, car ce sont eux aussi des désignateurs, on les regroupe d'ailleurs quelquefois avec ces derniers comme faisant partie des descriptions définies. Leur principale particularité est qu'ils constituent obligatoirement une reprise d'un référent précédemment introduit (ou introduit juste après eux) ; en cela, ils se rapprochent plutôt des

syntagmes définis incomplets/de reprise que des syntagmes définis complets/autonomes. Leur principale différence réside en ceci : Dans le cas des syntagmes définis incomplets, le contenu du syntagme sert en général uniquement à identifier le référent, dont la dernière citation est parfois assez éloignée ; c'est le cas dans la phrase suivante :

> Bien que coprésident du Parti radical, parti affilié à l'UMP et qui soutient officiellement le candidat UMP, **Jean-Louis Borloo** refuse toujours de dire pour qui son cœur balance. C'est là toute l'astuce, mais aussi toute la limite de l'exercice. « Après 17 ans passés à Valenciennes, au ministère de la Ville puis à celui de la Cohésion sociale, je suis venu à la conclusion qu'il fallait rétablir quatre piliers pour que la société fonctionne : l'éducation, le logement, l'emploi et l'équité territoriale. Aujourd'hui, beaucoup de moyens sont mis sur la table, mais chaque acteur -Régions, syndicats, associations ou services publics - défend sa propre boutique. Tout est figé », explique **le ministre**.

Par contre, les démonstratifs reprennent en général un antécédent très proche ; le contenu du syntagme n'est donc pas essentiel pour établir la référence, et il servira souvent plutôt à introduire une nouvelle caractéristique de la personne désignée, à la présenter sous un nouveau jour, comme dans les phrases suivantes :

> **Maurice Papon**, lui, en 1968, a repris sa brillante carrière. **Ce gaulliste ancien fonctionnaire de Vichy** a eu une carrière exemplaire dans une France schizophrène qui refusait de regarder et d'examiner les années noires de la collaboration.
>
> Alors que les électeurs ont, par deux fois, sanctionné les mandarins isolés dans leur Cité interdite, **Jacques Généreux**, membre du conseil national du PS, estime par exemple : « (...) On serait mieux inspiré d'enseigner les bienfaits de l'immigration pour notre pays et les vertus du métissage et du brassage des populations ». **Ce professeur à Sciences Po** est-il sûr de refléter la réalité vécue ?

Les syntagmes introduits par des démonstratifs seront également souvent plus rares ; on ne trouve par exemple aucune occurrence de « Sarkozy (…) ce ministre de l'Intérieur ».

Cette particularité des démonstratifs pourra être exploitée dans des patrons pour extraire de nouvelles expressions référentielles désignant une personne donnée.

Du point de vue du traitement informatique, l'antécédent (et donc la référence) des syntagmes nominaux démonstratifs sera identifié grâce à des critères essentiellement contextuels (principalement la proximité). La fenêtre dans laquelle on cherchera l'antécédent sera encore plus petite que celle associée aux syntagmes indéfinis (on ne prendra en compte que la phrase dans laquelle apparaît le syntagme et la phrase qui la précède).

## 2.3. Coréférence et reprises anaphoriques

### 2.3.1. Définitions

Nous nous basons sur [Milner 1982 : 9-66] pour les considérations qui suivent.

La coréférence est la relation qui unit deux expressions linguistiques pointant vers un même référent (c'est-à-dire deux référents en relation d'identité). Il ne faut pas confondre identité

des référents et identité lexicale ; dans l'exemple déjà donné plus haut, les référents correspondant aux expressions (résultant de leur référence actuelle) « étoile du matin » et « étoile du soir » sont identiques, mais pas les expressions elles-mêmes (qui détiennent seulement des points communs au niveau de leurs références virtuelles). L'interprétation de l'une des expressions coréférentielles n'est pas affectée par celle de l'autre (c'est ce qui distingue la relation de coréférence de celle d'anaphore) ; la relation de coréférence est symétrique (elle n'a pas d'orientation puisqu'aucun élément n'est dépendant de l'autre), transitive (si *a* est coréférentiel à *b* qui lui-même est coréférentiel à *c*, alors *a* est coréférentiel à *c*) et réflexive (si *a* est référentiel, alors on peut considérer qu'il est coréférentiel à lui-même) ; il s'agit donc d'une relation d'équivalence, ce qui signifie qu'il est possible de répartir les éléments référentiels d'un texte en classes d'équivalence, ces dernières correspondant à des ensembles d'éléments coréférentiels entre eux, et *a priori* interchangeables. Il est tentant de rapprocher ces classes d'équivalences des chaînes de référence (voir plus bas en 2.3.2.), mais il ne s'agit pas tout à fait de la même chose puisque les chaînes de références contiennent également des éléments anaphoriques.

L'anaphore doit en effet être distinguée de la coréférence :

> Il y a relation d'anaphore entre deux unités A et B quand l'interprétation de B dépend crucialement de l'existence de A, au point qu'on peut dire que l'unité B n'est interprétable que dans la mesure où elle reprend – entièrement ou partiellement – A. Cette relation existe quand B est un pronom dont la référence virtuelle n'est établie que par l'interprétation d'un N'' que le pronom « répète ». Elle existe également quand B est un N'' dont le caractère défini – c'est-à-dire le caractère identifié du référent – dépend exclusivement de l'occurrence, dans le contexte, d'un certain N'' – en fait, généralement, le même du point de vue lexical. [Milner 1982 : 18]

La relation d'anaphore est donc fondamentalement asymétrique (elle est orientée puisqu'elle concerne deux éléments dont l'un est anaphorisé (ou antécédent) et l'autre anaphorisant), non-transitive (un élément anaphorisant ne peut, selon [Milner 1982 : 33], être à son tour anaphorisé) et non-réflexive (« il paraît dépourvu de sens de soutenir qu'un terme soit anaphorisant ou anaphorisé de lui-même » [*Ibid.* : 34]). L'anaphore n'est pas l'apanage des pronoms ; l'anaphore nominale existe aussi. L'anaphorisé et l'anaphorisant sont dans ce cas tous deux des syntagmes nominaux, dont l'interprétation référentielle du second dépend crucialement de celle du premier. L'anaphorisant sera, typiquement, une description définie incomplète – on parle également de défini de reprise – (alors que la relation de coréférence préfèrera les descriptions définies complètes – on parlera de défini autonome), comme dans l'exemple suivant :

> **Le Pen** qui se pose en «seul défenseur de l'indépendance nationale» tape ensuite à bras raccourcis sur Ségolène Royal et Nicolas Sarkozy, rebaptisés pour l'occasion «Sargolène et

> Sékozy». « Chez elle, c'est simple, des idées, y en a pas et Sarkozy l'Américain louvoie», constate **le candidat à l'Elysée**.

où « le candidat à l'Elysée » est une reprise anaphorique de « Le Pen ».

Nous garderons à l'esprit qu'anaphore et coréférence sont deux relations différentes ; toutefois, nous continuerons à les rapprocher dans les parties qui suivent puisque leur traitement en TAL présente beaucoup de similarités. D'ailleurs, notre outil ne fait pas vraiment de distinction entre coréférence et anaphore nominale (quoique le traitement diffère légèrement selon le cas, comme nous l'avons souligné à propos de la différence entre descriptions définies et indéfinies).

### 2.3.2. Chaînes de référence

#### *2.3.2.1. Définitions*

Lorsque plus de deux éléments étant en relations anaphoriques les uns avec les autres se succèdent dans un texte, ils forment ce que l'on appelle une chaîne de référence.

C'est [Chastain 1975] (selon [Corblin 1995 : 151]) qui introduit la notion de chaîne de référence, dans un article consacré à la théorie de la référence ; il la définit comme « une séquence d'expressions singulières apparaissant dans un contexte telles que si l'une de ces expressions réfère à quelque chose, toutes les autres y réfèrent également » ([Chastain 1975 : 205] cité par [Corblin 1995]). Mais il faut attendre une dizaine d'années avant que cette notion ne suscite un réel intérêt, grâce aux travaux sur la cohésion textuelle (comme ceux de [Charolles 1988] et [1989]), mais surtout grâce aux recherches en informatique et en intelligence artificielle. [Karttunen 1976 cité par Corblin 1995 : 171] définissait déjà ainsi la tâche principale d'un interpréteur de textes :

> Considérons un automate conçu pour lire un texte dans une langue naturelle donnée, l'interpréter, et en enregistrer en quelque manière son contenu, par exemple être en mesure de répondre à des questions sur ce texte. Pour accomplir cette tâche, la machine devra remplir au minimum les exigences suivantes. Elle devra être en mesure de construire un fichier contenant la liste de toutes les entités, événements, objets, etc. mentionnés dans le texte, et pour chaque entité enregistrer ce qui en est dit.

Lister toutes les entités, événement, objets mentionnés dans un texte, et extraire tous les segments de texte où elles sont impliquées (même sans les interpréter), revient à chercher les chaînes de référence.

Dans la mesure où l'on peut trouver dans des chaînes de référence autre chose que des termes singuliers (des syntagmes nominaux à interprétation générique peuvent être anaphorisés), nous préfèrerons à la définition de Chastain celle, simple, de [Corblin 1995 : 27] :

> On appelle chaîne de référence une suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence.

#### *2.3.2.2. Positions des différents syntagmes*

Les différents types de syntagmes nominaux ont différentes places privilégiées à l'intérieur des chaînes de coréférence, directement liées à leurs fonctionnements référentiels décrits plus haut :

Les noms propres et les syntagmes définis autonomes peuvent apparaître partout dans la chaîne ; on les trouve souvent en tête de chaîne, place que ne peuvent occuper les syntagmes définis de reprise et les démonstratifs, que l'on trouvera plutôt en milieu de chaîne (avec une prédilection pour la seconde position en ce qui concerne les démonstratifs). Les syntagmes indéfinis apparaîtront presque toujours en tête de chaîne.

## 2.4. Le traitement automatique de la coréférence

### 2.4.1. Applications possibles

Le point de départ de notre travail est applicatif, puisqu'il a pour but de répondre à un besoin dans une application de lexicométrie. Mais il nous paraît intéressant de nous pencher sur les autres applications possibles, en TAL, du traitement automatique de la coréférence. Les applications concernées sont toutes celles qui nécessitent un peu de « compréhension » des textes (*Cf.* [Boudreau 2004 : 10]) ; sans être exhaustif, nous allons en donner quelques exemples dans ce qui suit.

L'outil que nous développons dans ce travail de mémoire ne pourrait servir à toutes les applications que nous allons rapidement détailler ci-dessous ; il a été conçu dans une visée applicative particulière, et nécessite un corpus de taille suffisante, et analysé syntaxiquement, alors que pour des raisons de coût, dans des applications pour lesquelles la résolution des coréférence n'est pas un but mais un moyen, on optera plus volontiers pour des méthodes *knowledge-poor* (voir plus bas 2.4.2. Méthodes).

#### *2.4.1.1. Traduction automatique*

[Mitkov 1996] montre le rôle crucial d'une phase de résolution des liens anaphoriques pour un système de traduction automatique, notamment pour la traduction des pronoms. Par exemple, du français à l'anglais, la traduction du pronom « le » (en position COD) peut-être, selon que l'antécédent est une personne ou un objet, « he » ou « it ».

#### *2.4.1.2. Recherche d'informations*

Pour déterminer la pertinence d'un document par rapport à une requête, on peut s'appuyer sur la fréquence des termes communs à la requête et au document, ou encore déterminer si le document contient tous les termes de la requête dans un contexte étroit. En résolvant préalablement les liens coréférentiels des textes, on pourrait :

- ajouter à la fréquence des termes la fréquence de leurs coréférents ;
- reconnaître des situations de présence de tous les termes d'une requête dans un contexte étroit même si l'un des termes est réalisé par une reprise anaphorique.

#### *2.4.1.3. Extraction d'informations*

Dans le domaine de l'extraction d'informations, la résolution des coréférences constitue une phase importante, au point qu'elle a été introduite comme tâche dans les conférences MUC (Message Understanding Conference) depuis 1995 (*Cf.* comptes-rendus des conférences MUC-6 et MUC-7 par [Sundheim 1995] et [Chinchor 1998]).

Les systèmes d'extraction d'informations peuvent notamment utiliser des patrons pour trouver la réponse à une question dans un texte ; si les liens de coréférence ont été résolus, les patrons pourront les utiliser comme points d'ancrage.

Dans le cas où le système doit fusionner des informations provenant de différentes sources, il est utile de détecter des relations de coréférence d'une source à l'autre ; ce problème est expliqué par [Kehler 1997]. Typiquement, un système opère en deux temps (c'est par exemple le cas de FASTUS, de [Hobbs *et al.* 1996]) : premièrement, grâce notamment à l'application de patrons, des *templates* contenant des informations sur les différents événements ou entités de chaque texte sont créés. Dans un second temps, il s'agit de fusionner les *templates* qui se rapportent à une même entité, et sont donc coréférents ; pour déterminer si deux *templates* sont coréférents et doivent donc être fusionnés, [Hobbs *et al*. 1996 : 18] utilise trois critères : la structure interne des syntagmes nominaux désignant les entités, la proximité et la compatibilité des *templates*.

#### *2.4.1.4. Résumé automatique*

Il existe deux sortes de résumés automatiques : l'*extract*, qui est construit à partir de l'extraction de phrases saillantes du texte, et l'*abstract*, qui est généré après une première phase de compréhension du texte. Ces deux types de résumés peuvent nécessiter de faire appel à une phase de résolution des liens anaphoriques.

- Dans le cas de l'*extract*, cela contribuerait à pallier le manque de cohésion du texte obtenu par la simple superposition des phrases extraites. Considérons le texte suivant :

> AU SERVICE de Ségolène Royal, le PS et son premier secrétaire, François Hollande, cherchent désormais leur place dans la prochaine campagne présidentielle. Pas facile, puisque la candidate a surtout construit son succès en s'adressant aux sympathisants de gauche, en dehors du parti. Insistant sur la « synergie » à mettre en œuvre entre le parti et la candidate, Hollande fait l'éloge du sens « collectif ». **Ségolène Royal** a d'ores et déjà laissé entendre qu'elle ne se laisserait pas récupérer par l'appareil. Hier, **il** a rappelé que « sa légitimité procède du vote » des militants du Parti socialiste. Fidèle à son caractère, **Ségolène Royal** n'entend rien négocier avec personne. La victoire de **Ségolène Royal** a donné des aigreurs aux tenants de la gauche du PS.

Il s'agit d'un résumé produit par l'outil de synthèse de textes de Word (2003), à partir d'un article de notre corpus. On peut y relever (en gras) deux types d'erreurs qui pourraient être évitées : le pronom anaphorique « il » semble, telles que sont agencées les phrases, renvoyer à « Ségolène Royal », ce qui est peu probable ; dans les deux dernières phrases, Ségolène Royal est nommée deux fois, alors qu'on s'attendrait à une reprise anaphorique en seconde position, ce qui crée une impression de lourdeur et de manque de cohésion. Il existe bien sûr de meilleurs outils de résumé automatique de type *extract* que celui-ci, mais tous sont concernés par ces problèmes.

- La recherche des chaînes de coréférence peut aussi servir de base pour produire un résumé de type *abstract*. Par exemple, [Bergler 2003] utilise, pour résumer automatiquement un texte, les syntagmes nominaux appartenant aux chaînes coréférentielles les plus longues ; les chaînes les plus longues sont effectivement présumées renvoyer aux entités les plus importantes du texte (et en constituent donc le « sujet »), et les différents syntagmes nominaux qui les composent donnent accès à la manière dont sont envisagées ces entités-sujet.

### 2.4.2. Méthodes

#### *2.4.2.1. Années 80s et 90s : premiers systèmes*

- Hobbs 1978

Selon [Boudreau 2006 : 203], [Hobbs 1978] fut parmi les premiers à se pencher sur la résolution automatique des anaphores (pronominales). L'algorithme qu'il a mis au point repose sur l'ordre selon lequel l'arbre syntaxique de la phrase est parcouru, ainsi que sur des critères de sélection en termes de compatibilité sémantique (nature animée/inanimée de l'antécédent) ou morphologique (concordance du genre et du nombre entre le pronom à résoudre et son antécédent). Lorsqu'un pronom anaphorique est reconnu, la recherche de l'anaphorisé se fait en parcourant l'arbre syntaxique de la phrase selon un ordre prédéterminé,

qui favorise certaines fonctions lexicales (sujet > objet > objet d'une préposition) et préfère une faible distance entre l'anaphorisant et l'anaphorisé. Ce système a joui d'une grande popularité qui fait qu'il a été utilisé jusqu'aux années 90s.

- Lappin et Leass 1994

Seize ans plus tard, le système décrit par [Lappin & Leass 1994] est encore assez proche de celui de Hobbs. Lui aussi nécessite une analyse syntaxique complète du texte source, mais ce n'est pas la façon de parcourir l'arbre syntaxique de la phrase qui détermine quel est l'élément anaphorisé : tous les antécédents potentiels sont pondérés, et l'antécédent potentiel ayant le poids le plus important (on parlera aussi d'antécédent le plus *saillant*) est choisi. Les antécédents potentiels sont les syntagmes nominaux de la phrase et des phrases précédentes (le poids de l'antécédent est divisé par deux pour chaque phrase le séparant de l'anaphorisant) qui ne sont pas exclus par certains critères morphologiques (compatibilité en genre et en nombre) ou positionnels (par exemple, un pronom ne peut renvoyer à un nom s'il se trouve dans le même syntagme nominal que ce dernier) qui tiennent lieu de filtre. Ces antécédents potentiels sont regroupés en classes d'équivalence (les éléments déjà identifiés comme étant coréférentiels entre eux appartiennent à la même classe) ; un poids est affecté au dernier élément de chaque classe d'équivalence, selon divers facteurs, qui peuvent soit augmenter soit diminuer ce poids. Par exemple, le poids des candidats occupant la même fonction syntaxique que le pronom anaphorique (critère du parallélisme des fonctions) sera majoré, alors que le poids des candidats fortement enchâssés sera minoré. Si deux candidats sont également saillants, c'est le plus proche de l'anaphorisant qui sera sélectionné.

L'inconvénient majeur des deux systèmes que nous venons de présenter *était* qu'ils demandaient une analyse syntaxique complète des textes à traiter, ce qui constituait un prétraitement coûteux, l'analyse syntaxique automatique étant peu fiable.

#### *2.4.2.2. Seconde moitié des années 90s : avènement des méthodes knowledge-poor*

De la seconde moitié des années 90s à aujourd'hui, les travaux sur la résolution des liens coréférentiels se caractérisent plutôt par une volonté de réduire au maximum les ressources linguistiques utilisées par les systèmes. [Mira 1990] montre l'importance des marques de « bas-niveau » (marques morphologiques et relations syntaxiques élémentaires, par opposition aux connaissances de haut-niveau : connaissances d'ordre sémantique ou pragmatique, analyse syntaxique complète) pour la détermination de la saillance d'un antécédent.

- Mitkov

Un des algorithmes les plus connus et les plus classiques utilisant ces connaissances de bas-niveau est celui présenté par [Mitkov 1998] pour la résolution des pronoms anaphoriques. Cet algorithme demande en prétraitement non une analyse syntaxique complète, mais le simple balisage des syntagmes nominaux (identifiés à partir de l'étiquetage du texte en parties du discours). Le texte est ensuite parcouru linéairement et chaque pronom anaphorique rencontré est traité. Les antécédents possibles sont les syntagmes nominaux apparaissant dans une fenêtre de trois phrases (celle où apparaît le pronom à résoudre et les deux phrases précédentes). A chaque antécédent est affectée une saillance selon leur conformité aux critères suivants (retranscrits par [Boudreau 2004 : 30]) :

| |
|---|
| - L'antécédent est compatible en genre et en nombre. |
| - L'antécédent est un syntagme défini. |
| - L'antécédent est le thème de la phrase précédente. |
| - L'antécédent est un syntagme nominal suivant un verbe indicateur (discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover). |
| - L'anaphore est une réitération lexicale de l'antécédent. |
| - L'anaphore est une expression qui reprend une partie du titre de la section. |
| - L'antécédent n'est pas un complément indirect (insérer le disque dans le lecteur). |
| - L'antécédent partage le même contexte que le pronom. |
| - L'anaphore est dans une construction intrinsèquement anaphorique. |
| - La distance entre l'anaphore et l'antécédent est petite. |
| - L'antécédent est un terme appartenant au sujet (domaine) du texte. |

L'antécédent obtenant la saillance la plus importante est choisi.

- CogNIAC

Le système CogNIAC, présenté par [Baldwin 1997], s'appuie sur le constat suivant : si la résolution complète des anaphores d'un texte nécessite effectivement de nombreuses connaissances – syntaxiques et sémantique, mais aussi des connaissances du monde, du domaine, de la situation d'énonciation, etc. – il existe de nombreuses anaphores qui peuvent être résolues sans rien de tout cela. A l'instar de l'algorithme de Mitkov, CogNIAC n'utilise comme prétraitement que le découpage du texte en phrases et l'étiquetage des catégories grammaticales des mots les constituant (avec une reconnaissance simple des syntagmes nominaux). Quelques critères de sélection simples permettent ensuite de reconnaître l'antécédent des éléments anaphoriques : S'il n'y a qu'un seul antécédent compatible en genre et en nombre, il est choisi ; si l'anaphorisant est un pronom réfléchi (comme « lui-même »),

l'antécédent le plus proche est choisi ; etc. La spécificité de CogNIAC est que lorsqu'il y a ambiguïté entre plusieurs antécédents, l'anaphorisant n'est pas résolu. CogNIAC résout ainsi les liens anaphoriques avec une précision importante (supérieure à 90%) et, bien sûr, un rappel bien moindre (environ 60%), ce qui peut s'avérer suffisant selon les applications visées.

- Dupont 1998

[Dupont 1998] décrit un système de calcul de la référence opérant sur un corpus homogène, puisqu'il n'est constitué que de constats d'accidents. Des critères classiques permettent de calculer la saillance des antécédents possibles (distance entre l'anaphorisant et l'antécédent, supériorité de la position anaphorique sur la position cataphorique, compatibilité du genre et du nombre, etc.), mais en plus de cela, les différentes entités sont classées en catégories sémantiques (comme par exemple « personne », « véhicule », « route ») ; cela est rendu possible par le fait que tous les textes du corpus appartiennent au même domaine. L'antécédent choisi sera celui ayant la saillance la plus importante parmi ceux dont le type sémantique correspond au type de l'anaphorisant.

- Siddharthan 2003

[Siddharthan 2003] mène plus loin encore que ses prédécesseurs la simplification des connaissances linguistiques nécessaires au traitement des liens anaphoriques. Il propose de n'utiliser comme prétraitement que le balisage des noms, et d'extraire toutes les autres informations nécessaires à un calcul de saillance par des techniques de surface comme l'application de patrons. Par exemple, l'observation des prépositions ou de la position des noms par rapport au verbe peut permettre de deviner la fonction syntaxique d'un syntagme nominal. L'observation des accords des déterminants ou des mots suivant un nom (adjectifs, participes passés) peut permettre de déterminer ses genre et nombre. Une fois ces informations extraites, la résolution des anaphores se fait comme avec les autres systèmes, par sélection de l'antécédent compatible ayant la plus forte saillance.

- Boudreau 2006

[Boudreau 2006] émet l'hypothèse que les phénomènes de coréférence varient selon le type de texte, le domaine. Elle utilise pour extraire les expressions référentielles à résoudre des listes de noms communs propres au domaine.

#### *2.4.2.3. Notre position*

Le choix de notre méthode est conditionné par les caractéristiques de notre corpus et par les ressources que nous avons à notre disposition.

Nous avons à notre disposition un corpus entièrement analysé syntaxiquement ; c'est donc sans trop d'hésitations que nous pouvons recourir à des méthodes laissées de côté par les partisans du *knowledge-poor* pour des raisons pratiques, mais dont l'efficacité n'est pas niée. Nous utilisons donc des critères de saillance basés sur les fonctions syntaxiques des éléments parmi lesquels on cherche à mettre au jour des relations de coréférence.

De plus, notre corpus est à la fois très important (plusieurs millions de mots) et constitué de textes très similaires (appartenant au même domaine – la politique – et de plus, traitant du même sujet – l'élection présidentielle de 2007). De cela, il résulte que les mêmes noms propres apparaissent dans tout le corpus, ainsi que les mêmes syntagmes coréférentiels à ces noms propres (la plupart des syntagmes nominaux coréférentiels à des personnes – surtout dans le monde de la politique – sont des noms de fonctions, et le nombre de fonctions que peut occuper une même personne est limité). C'est cet état des choses qui nous a incitée à utiliser, en plus des critères de résolution en contexte, des scores d'association, qui donnent le degré de « colle » entre chaque couple constitué d'un syntagme nominal référentiel et d'un référent possible (nom propre). Cette pondération de chaque couple, calculée pour tout le corpus (il s'agit donc d'un critère non dépendant du contexte de chaque occurrence particulière) se combinera avec les critères contextuels lors de la phase d'attribution des syntagmes nominaux aux bons référents.

Pour calculer ces scores, nous faisons appel (au choix) à différentes mesures de liaison, que nous allons détailler dans la partie qui suit.

## 3. Mesures de liaison

Les mesures de liaisons sont utilisées pour l'extraction de relations sémantiques, notamment des relations syntagmatiques (collocations), mais aussi paradigmatiques (hyperonymie, synonymie, etc.). Elles mesurent l'intensité du lien entre deux éléments d'un couple. Pour nous, les couples seront formés d'un syntagme nom propre, qui donne le référent, et d'un syntagme nominal potentiellement coréférentiel.

Les mesures de liaison se basent sur les fréquences : fréquence d'association des deux éléments du couple, fréquence des couples où un élément se trouve mais pas l'autre, etc. Ces fréquences peuvent être, pour chaque couple, regroupées dans un *tableau de contingence* :

| | $S_j$ | $S_{j'}$ avec j' différent de j |
|---|---|---|
| $S_i$ | a | b |
| $S_{i'}$ avec i' différent de i | c | d |

**Figure 3 : Tableau de contingence d'un couple de syntagmes, basé sur [Daille 1994]**

Voici à quoi les valeurs *a*, *b*, *c* et *d* correspondent, dans notre corpus :

a = le nombre de fois qu'un couple $S_i$ / $S_j$ (comme par exemple *sarkozy_nicolas* / *ministre de l'intérieur*) a été extrait, par un patron ou une méthode basée sur la cooccurrence ;

b = le nombre de fois où $S_i$ est le premier élément d'un couple sans que $S_j$ n'en soit le second (pour notre exemple, le nombre de fois où *sarkozy_nicolas* est associé à un autre syntagme que *ministre de l'intérieur*) ; ce nombre est obtenu en soustrayant *a* au nombre total de fois où $S_i$ a été extrait au sein d'un couple ;

c = le nombre de fois où $S_j$ est le second élément d'un couple sans que $S_i$ n'en soit le premier (le nombre de fois où *ministre de l'intérieur* est associé à une autre personne que *sarkozy_nicolas*) ; ce nombre est, de la même manière que *b*, obtenu en soustrayant *a* au nombre d'occurrences des couples ou $S_j$ apparaît ;

d = le nombre d'occurrence des couples où ni $S_i$ ni $S_j$ n'apparaissent ; ce nombre équivaut à : nombre total d'occurrences de tous les couples extraits + nombre d'occurrences des couples ou $S_i$ apparaît + nombre d'occurrences des couples où $S_j$ apparaît – nombre d'occurrence du couple $S_i$ / $S_j$.

On note que $a+b+c+d$ = nombre total d'occurrences de tous les couples extraits.

Il existe de nombreuses mesures de liaison ; [Daille 1994] en décrit dix, que nous allons toutes tester dans notre programme. Nous les reprenons telles qu'elle les présente :

1) Coefficient de proximité simple (smc)

$$smc = \frac{a+d}{a+b+c+d}$$

2) Coefficient de Kulczinsky (kuc)

$$kuc = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$$

3) Coefficient d'Ochiai (och)

$$och = \frac{a}{\sqrt{(a+b)(a+c)}}$$

4) Coefficient de Fager et Mc Gowan (fag)

$$fag = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+b}}$$

5) Coefficient de Yule (yul)

$$yul = \frac{ad - bc}{ad + bc}$$

6) Coefficient de McConnoughy (mcc)

$$mcc = \frac{a^2 - bc}{(a+b)(a+c)}$$

7) Coefficient du $\Phi^2$ (phi)

$$phi = \frac{(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

8) Information mutuelle (im)

$$im = \log_2 \frac{a}{(a+b)(a+c)}$$

9) Information mutuelle au cube (im3)

On peut reprocher à l'information mutuelle de sur-pondérer les hapax ; c'est pour pallier ce défaut que [Daille 1994] a introduit cette nouvelle mesure.

$$im = \log_2 \frac{a^3}{(a+b)(a+c)}$$

10) Coefficient de vraisemblance (loglike)

$$a \log a + b \log b + c \log c + d \log d - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + N \log N$$

# SECONDE PARTIE :

# Description du travail informatique effectué

# 1. Corpus et ressources

## 1.1. Caractérisation du corpus

### 1.1.1. Constitution du corpus

Nous avons déjà donné, dans ce qui précède, quelques éléments de description de notre corpus (notamment en 1.5.1. et en 2.4.2.3.). Nous allons ici le caractériser de manière plus systématique, ce qui nous amènera à faire quelques redites.

Le corpus qui nous a été fourni a été constitué automatiquement en sélectionnant des articles pertinents à partir des fils RSS de trois journaux : Le Monde, Le Figaro et Libération. Le seul critère de sélection des articles est la présence d'au moins une citation, dans le corps de l'article, d'un des présidentiables ou personnalités politiques importantes listés au préalable. Pour quelques noms propres ambigus (comme « Hollande »), des règles très simples (absence de déterminant) permettent de s'assurer que l'on a bien affaire à un patronyme. Nous n'avons pas tenté de chiffrer le bruit généré par ce procédé, mais il semble très faible. Sur la totalité des articles que nous avons parcourus, un seul était hors-domaine : traitant des produits de luxe pour animaux, il avait été sélectionné à cause d'une citation de la marque « Royal Canin » (« Royal », patronyme de Ségolène Royal, étant l'un des mots-clés conduisant à la sélection de l'article).

Ce corpus est aujourd'hui encore mis à jour, mais nous n'en avons gardé qu'un corpus stable couvrant une période allant du 21 août 2006 au 20 avril 2007, inclus. La date du 21 août ne correspond à aucun événement particulier, mais en deçà de cette date, la quantité d'articles traitant de l'élection présidentielle de 2007 (qui croît régulièrement sur l'ensemble de notre corpus) a été jugée trop faible. Le 20 avril 2007 se situe à l'avant-veille du premier tour des élections ; or, suite à cet événement (puis plus tard suite au second tour), le statut de nombreuses personnes citées dans notre corpus change : par exemple, François Bayrou n'est plus « candidat UDF à la présidentielle », mais, entre autres, « candidat malheureux au premier tour », et Nicolas Sarkozy devient « Président de la République », ce dernier syntagme ne pouvant donc plus être traité comme un désignateur rigide de Jacques Chirac. Ces changements (on parle de références évolutives) pourraient être traités automatiquement, mais cela sort du cadre de notre travail. Nous observerons toutefois de plus près comment sont traités les quelques cas de références évolutives internes à notre corpus : notamment le syntagme « ministre de l'Intérieur », qui est l'un des syntagmes nominaux les plus fréquents

de notre corpus (voir plus bas) et qui, à partir du 26 mars 2007 cesse de désigner Nicolas Sarkozy pour désigner François Baroin.

Le corpus ainsi formé totalise 7303 articles, ce qui représente plus de quatre millions de mots[11] répartis comme suit :

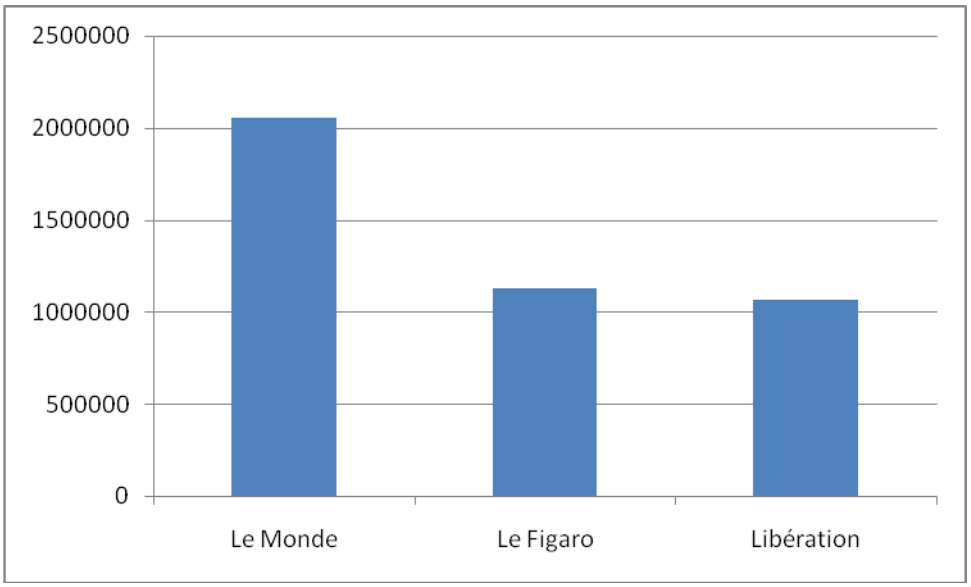



**Figure 2 : Partition par journal**

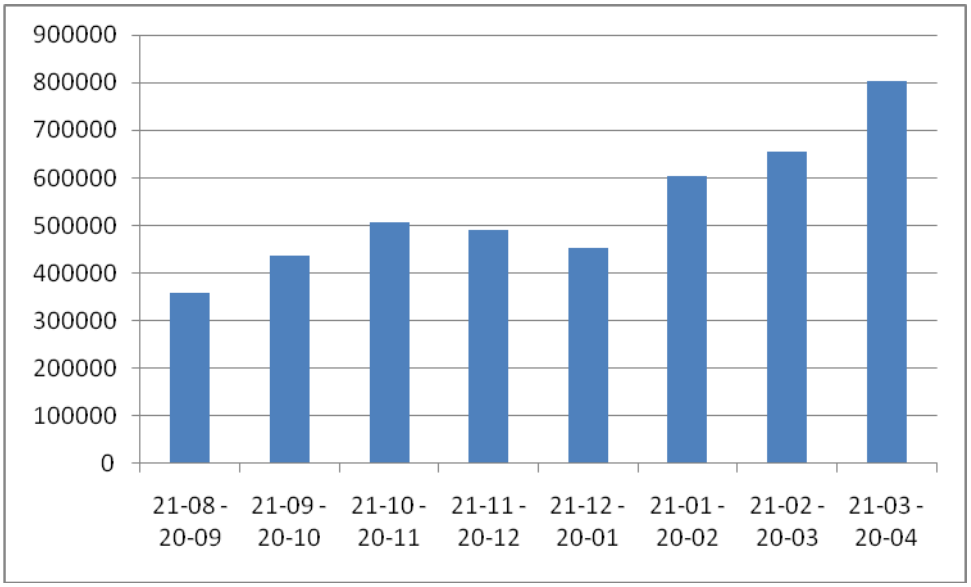



**Figure 4 : Partition par mois**

[11] 4 253 186 mots

Il s'agit donc d'un corpus de taille importante, qui jouit, comme nous l'avons déjà souligné, d'une assez grande homogénéité : les textes qui le constituent appartiennent tous au même genre (ce sont des articles de presse), au même domaine (la politique), et traitent du même sujet (l'élection présidentielle française de 2007) ; le nombre de sources est de plus relativement réduit (3).

### 1.1.2. Expressions référentielles

#### *1.1.2.1. Noms de personnes*

Notre outil a relevé dans le corpus 91 296 noms propres[12], qu'il a associés à 6100 référents différents ; en voici les quinze plus fréquents, avec leurs fréquences par journal.

**Tableau 1 : les 15 noms propres les plus fréquents**

| Colonne1 | total | Le Monde | Le Figaro | Libération |
|---|---|---|---|---|
| sarkozy_nicolas | 13281 | 6260 | 4134 | 2887 |
| royal_ségolène | 11207 | 5409 | 3542 | 2256 |
| bayrou_françois | 4561 | 2465 | 1327 | 769 |
| chirac_jacques | 3869 | 1923 | 1137 | 809 |
| le pen_jean-marie | 2489 | 1288 | 765 | 436 |
| de villepin_dominique | 2477 | 999 | 927 | 551 |
| hollande_françois | 1514 | 664 | 556 | 294 |
| jospin_lionel | 1369 | 548 | 386 | 435 |
| fabius_laurent | 1288 | 577 | 382 | 329 |
| bové_josé | 1195 | 606 | 202 | 387 |
| strauss-kahn_dominique | 1025 | 523 | 302 | 200 |
| hulot_nicolas | 850 | 391 | 275 | 184 |
| buffet_marie-george | 745 | 421 | 146 | 178 |
| voynet_dominique | 739 | 406 | 142 | 191 |
| besancenot_olivier | 716 | 418 | 129 | 169 |

Tous les noms propres ne présentent pas le même intérêt pour nous : on s'intéresse bien sûr particulièrement aux personnalités politiques, et donc aux noms propres les plus fréquents de notre corpus. Toutefois, nous les traitons tous de la même manière, nous expliquerons pourquoi en *2.1.3.1*. Ce n'est que lors de l'évaluation que nous ferons une distinction entre les noms propres selon leurs fréquences.

[12] Le rappel est supérieur à la précision pour cette tâche, donc le nombre réel est légèrement sur-évalué.

### *1.1.2.2. Syntagmes nominaux référentiels*

Nous avons extrait 72 709[13] occurrences de syntagmes nominaux référentiels, correspondant à 24 142 formes différentes ; en voici les quinze plus fréquents :

**Tableau 2 : Les 15 syntagmes nominaux les plus fréquents**

| Syntagme | total | Le Monde | Le Figaro | Libération |
|---|---|---|---|---|
| candidat | 2143 | 1087 | 642 | 414 |
| candidate | 1515 | 708 | 439 | 368 |
| premier ministre | 1415 | 617 | 536 | 262 |
| ministre de l' intérieur | 1399 | 660 | 413 | 326 |
| président | 1085 | 517 | 291 | 277 |
| candidate socialiste | 933 | 517 | 325 | 91 |
| personne | 923 | 439 | 214 | 270 |
| ministre | 898 | 395 | 237 | 266 |
| candidat de l' ump | 821 | 472 | 221 | 128 |
| homme | 812 | 397 | 182 | 233 |
| président de l' ump | 746 | 365 | 279 | 102 |
| président de la république | 730 | 353 | 247 | 130 |
| chef de l' etat | 623 | 347 | 107 | 169 |
| femme | 588 | 268 | 137 | 183 |
| candidat ump | 377 | 155 | 144 | 78 |

Dans la mesure où nous ne traitons pas les syntagmes nominaux tout à fait de la même façon selon leur déterminant, il nous a paru intéressant de regarder leur partition par type de déterminant :

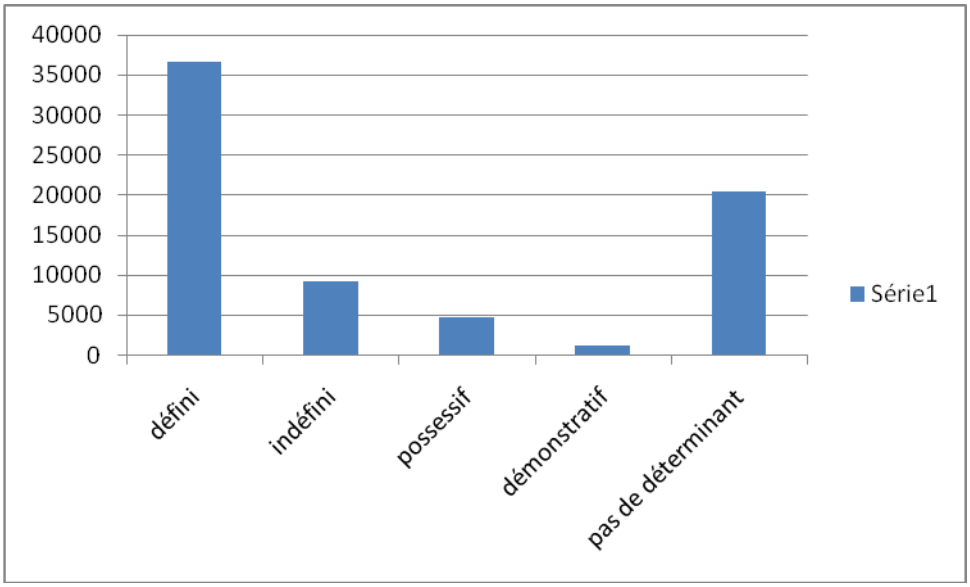



**Figure 5 Distribution des syntagmes nominaux selon leurs déterminants**

[13] Là encore il s'agit d'une légère sur-évaluation.

Parmi les trois journaux, celui qui se démarque le plus est Libération. Il s'agit du journal qui marque le plus son orientation politique, et le style de ses articles est souvent le moins « lissé », le moins normalisé. Cela se répercute sur les expressions référentielles : il s'agit du journal qui utilise le plus d'expressions référentielles différentes pour un même nombre de mots.

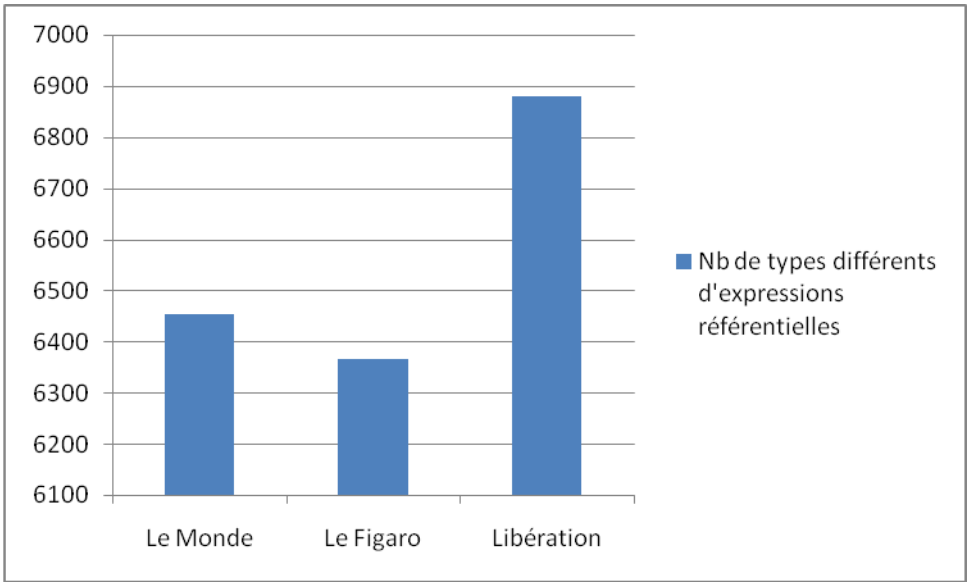



**Figure 6 Types d'expressions référentielles pour les trois sous-corpus ramenés à 1 million de mots**

Certaines expressions référentielles lui sont propres, comme « le locataire de la place Beauvau » pour Nicolas Sarkozy (le ministre de l'Intérieur siège place Beauvau). Nous verrons avec l'évaluation de notre programme que c'est avec les articles de Libération que les résultats sont les moins bons.

## 1.2. Corpus d'entraînement, corpus d'évaluation

### 1.2.1. Partitionnement du corpus

Notre corpus a été partitionné en un sous-corpus d'entraînement, un sous-corpus d'évaluation (annoté manuellement), et un sous-corpus qui n'a pas servi à l'entraînement ni n'a été annoté (la seule utilité de ce dernier est qu'il permet d'observer la progression de résultats avec l'augmentation du corpus).

Le corpus d'entraînement couvre la période allant du 21 août au 18 février. Il représente donc un peu plus de deux millions et demi de mots. Le corpus d'évaluation couvre une petite période de trois jours : du 18 au 20 février. Cela représente 46 644 mots

Ce partionnement n'a pas été réalisé de manière réfléchie, puisqu'il s'est fait avant tout pour des raisons pratiques (au fil des mises à jour de notre corpus). Il aurait peut-être été préférable de constituer le sous-corpus d'évaluation en sélectionnant des articles de manière aléatoire, afin qu'il couvre une plus grande période de temps. Toutefois, cette configuration a également ses avantages : dans la mesure où chaque semaine, de nouveaux syntagmes nominaux apparaissent, au gré de ce qui fait l'actualité (voir pour cela le site de LexiMedia2007, qui met en valeur les syntagmes nouvellement apparus), le fait d'avoir un corpus d'évaluation qui n'est pas englobé (chronologiquement) par le corpus d'entraînement assure que notre outil sera confronté à de nouveaux syntagmes référentiels.

### 1.2.2. Annotations manuelles

L'annotation manuelle a été effectuée par un tiers, et relue par nous. Le fichier à annoter était présenté de la manière suivante :

<TXT>On a passé le temps des affrontements binaires " , glisse l' ancien ministre de l' agriculture , François Patriat , conseiller de Mme Royal sur le sujet .
On
a
passé
le
temps
des
affrontements
binaires
"
,
glisse
l'
ancien
ministre
de
l'
agriculture
,
François
Patriat
,
conseiller
de
Mme
Royal
sur
le
sujet
.

**Instructions données à l'annotateur :**

Lorsqu'un syntagme nominal (nom commun ou propre) est référentiel à une personne, l'annotateur précise sa référence au niveau de la tête du syntagme (pour un nom propre, il

s’agira du nom de famille), après une tabulation. Les autres éléments du syntagme sont marqués par un tiret. En cas d’imbrication de syntagme, le syntagme englobé est annoté après une nouvelle tabulation. Lorsqu’un syntagme est référentiel mais que son référent est indéterminé, il est annoté avec pour référence « inconnu ». Lorsqu’un syntagme est en emploi générique, il n’est pas annoté ; par contre, les emplois attributifs sont annotés. D’un point de vue strictement linguistique, ces derniers ne sont pas référentiels : ils caractérisent une entité à laquelle on a fait référence précédemment. Mais dans la mesure où l’on n’a pas cherché à les exclure dans notre outil (les syntagmes non-référentiels mais apportant une information sur une personne particulière nous intéressent aussi), nous avons demandé à l’annotateur de les traiter comme les autres. L’annotation manuelle pour l’extrait ci-dessus est de cette forme :

<TXT>On a passé le temps des affrontements binaires " , glisse l' ancien ministre de l' agriculture , François Patriat , conseiller de Mme Royal sur le sujet .
On
a
passé
le
temps
des
affrontements
binaires
"
,
glisse
l' -
ancien -
ministre patriat_françois
de -
l' -
agriculture -
,
François -
Patriat patriat_françois
,
conseiller patriat_françois
de -
Mme - -
Royal - royal_ségolène
sur -
le -
sujet -
.

### 1.2.3. Caractérisation du corpus d'évaluation

## 1.3. Prétraitement du corpus

Notre outil prend pour entrée le corpus décrit entièrement analysé syntaxiquement par SYNTEX. Dans le fichier correspondant à l'analyse de SYNTEX, chaque phrase est présentée sous cette forme :

```
<SEQ id=LeFigaro-2006-08-21@00:00-afb6e5233b5d989ca35d9a51aa151017_12; analyse=1;>
<TXT>« Nous ne sommes pas des toutous »
<ETIQ>Typo|«|«|1||        Pro|nous|Nous|2|SUJ;4|    Adv|ne|ne|3|ADV;4|
        VCONJP|être|sommes|4||ADV;3,ADV;5,SUJ;2,ATTS;7        Adv|pas|pas|5|ADV;4|
        Det|de|des|6|DET;7|        NomMP|toutou|toutous|7|ATTS;4|DET;6    Typo|»|»|8||
```

- La ligne *SEQ* contient un identifiant unique pour chaque phrase analysée, constitué de l'identifiant de l'article dans lequel elle apparaît, suivi d'un trait de soulignement (« _ ») et du numéro de la phrase au sein de l'article.
- La ligne *TXT* contient le texte de la phrase.
- La ligne *ETIQ* contient toutes les étiquettes résultant de l'analyse de SYNTEX, séparées par des tabulations.

Chaque étiquette comporte six informations, séparées par des *pipes* : la catégorie grammaticale du mot étiqueté, son lemme, sa forme, le numéro de son apparition dans la phrase, le numéro dans la phrase des mots qui le régissent ainsi que le type de relation qui les lie, le numéro des mots qu'il régit ainsi que le type de relation qui les lie :

catégorie|lemme|forme|numéro dans la phrase|mots recteurs|mots régis

Par exemple, l'étiquette :

NomMP|toutou|toutous|7|ATTS;4|DET;6

représente un nom masculin pluriel, dont la forme est « toutous » et le lemme (sa forme de citation, au singulier) « toutou » ; il s'agit du septième élément étiqueté de la phrase ; il régit le déterminant placé en sixième position, et est un attribut du sujet régi par le verbe situé en quatrième position.

Certaines relations nous intéressent plus particulièrement que d'autres :

- La relation NNPR, qui lie un prénom et son nom de famille.
- La relation EPI, qui lie les noms propres aux particules ou noms qui les régissent (« M. », « docteur », etc.) et que l'on retrouve de manière générale lorsqu'un nom en régit un autre.

- Les relations SUJ et OBJ, qui nous serviront avec plus ou moins de succès à influencer le choix d'un antécédent lors de la résolution des syntagmes référentiels en contexte.

## 1.4. Autres ressources

### 1.4.1. Liste de mots-clés

Pour repérer des syntagmes potentiellement référentiels lors du module de balisage, nous nous servons comme seul filtre une liste de mots clés : il s'agit donc d'une ressource très importante dans notre outil. Nous avons constitué nous-même et de manière semi-automatique la liste que nous utilisons. Pour cela, nous avons relevé automatiquement, sur l'intégralité du corpus d'entraînement tous les noms pouvant être en relation EPI avec un nom propre. Puis nous avons manuellement filtré la liste obtenue pour ne garder que les noms potentiellement référentiels à des personnes, en y faisant quelques ajouts systématiques : par exemple, lorsqu'un nom masculin (par exemple « député ») était rencontré, nous ajoutions l'équivalent féminin (« députée »).

Les mots de notre liste sont de deux types. Certains, les plus nombreux, renvoient toujours à des personnes – c'est le cas notamment de la plupart des noms de fonction ; d'autres peuvent aussi désigner une entité d'un autre type. C'est le cas par exemple de « soutien » qui peut désigner soit l'action de soutenir, d'aider (1), soit la personne qui effectue cette action (2).

> (1) Mme Buffet veut encore croire en un hypothétique **soutien** des collectifs locaux.
>
> (2) «Je pense que les socialistes feraient une erreur s'ils voulaient réduire le budget de la Défense», commente Jean-Pierre Masseret, président de l'Assemblée de l'UEO et **soutien de Ségolène Royal**.

Au final, le mot « soutien », comme certains autres, génère beaucoup plus de bruit qu'il n'extrait de syntagmes réellement référentiels à des personnes ; mais il sera malgré tout conservé dans la liste. En effet, ce qui compte avant tout est d'assurer un rappel important, dans la mesure où un post-filtrage aura lieu par la suite (les syntagmes qui ne seront pas résolus seront considérés non-référentiels).

### 1.4.2. Liste d'associations nom propre-identifiant

Les associations entre chaque nom propre et l'identifiant qui lui correspond se font automatiquement lors du module de balisage du corpus ; mais si on spécifie dans cette liste des associations (sous la forme « nom propre *tabulation* identifiant »), elles prennent le pas sur ce qui est calculé automatiquement. Nous n'avons utilisé cette liste que pour associer certaines abréviations (« Sarko », « Ségo » ou « DSK ») à l'identifiant qui leur correspond

(respectivement *sarkozy_nicolas*, *royal_ségolène* et *strauss-kahn_dominique*). En effet, notre module n'est pas capable de faire lui-même ces associations (deux noms propres différents ne peuvent pour lui référer à la même personne), comme nous le verrons dans la discussion sur les limites de notre outil.

# 2. Description de l'outil programmé

Nous décrivons dans cette section l'outil que nous avons réalisé, module par module. La chaîne de traitement complète est résumée dans le schéma de la figure 7 : le corpus prétraité par SYNTEX est d'abord annoté par le module de balisage, qui repère les expressions potentiellement référentielles à des personnes (noms propres et syntagmes nominaux). Des couples NPr / SN sont alors extraits et classés selon leur score d'association. Enfin, le module de projection utilise d'une part le corpus balisé, et d'autre part les associations calculées précédemment, afin d'attribuer les expressions référentielles détectées au bon référent.

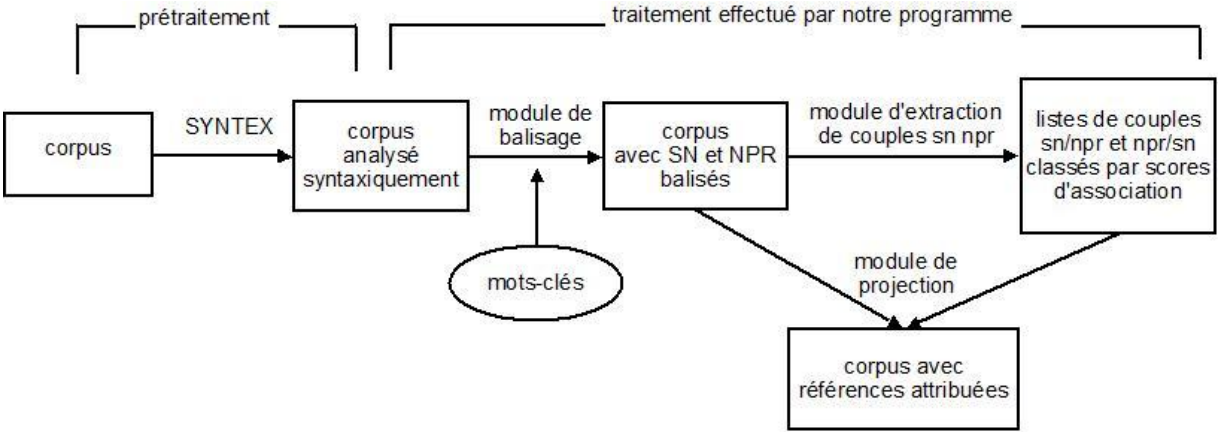



**Figure 7 Chaîne de traitement**

## 2.1. Module de balisage du corpus

### 2.1.1. Principe

Le module de balisage reçoit en entrée le contenu du corpus analysé par SYNTEX, ainsi que la liste de mots clés et la liste d'associations *nom propre-identifiant* spécifiées par l'utilisateur.

En sortie, les expressions référentielles (noms propres et les syntagmes nominaux – potentiellement – référentiels) sont balisées ; des informations sont associées à chaque occurrence de ces expressions référentielles, mais aussi à chaque type. Les syntagmes

nominaux du même type sont les syntagmes identiques à l'exception du déterminant (« un ministre de l'Intérieur », « le ministre de l'Intérieur » et « ce ministre de l'Intérieur » sont du même type) ; les noms propres du même type sont ceux qui ont le même référent (« M. Sarkozy » et « Nicolas Sarkozy » sont du même type).

Pour ce faire, le module parcours deux fois intégralement la structure de données contenant le corpus analysé. Lors du premier parcours, tous les syntagmes nominaux dont la tête est un mot-clé sont annotés, et les noms propres de personnes sont distingués des autres par l'application de patrons et de statistiques simples ; certaines informations les concernant sont déjà collectées. L'annotation des noms propres ainsi repérés se fait lors du second parcours.

### 2.1.2. Annotation des syntagmes nominaux potentiellement référentiels

Le corpus analysé par SYNTEX est parcouru séquence par séquence, et pour chaque séquence, étiquette par étiquette. Lorsque l'on trouve un nom singulier présent dans la liste de mots-clés, on construit le syntagme maximal dont ce nom est la tête, en parcourant les étiquettes de mots régis en mots régis, jusqu'à arriver à des mots n'étant pas eux-mêmes recteurs. Prenons l'exemple du syntagme « l'ancien ministre socialiste de la culture » ; dans le fichier analysé par SYNTEX, il se présente sous cette forme :

```
DetMS|le|l'|5|DET;7|
AdjMS|ancien|ancien|6|ADJ;7|
NomMS|ministre|ministre|7||DET;5,PREP;9,ADJ;6,ADJ;8
AdjMS|socialiste|socialiste|8|ADJ;7|
Prep|de|de|9|PREP;7|NOMPREP;11
DetFS|le|la|10|DET;11|
NomFS|culture|culture|11|NOMPREP;9|DET;10
```

« Ministre » étant un nom singulier présent dans la liste de mots-clés, le module va construire le syntagme dont il est la tête. Pour cela, il recherche ses mots régis : un déterminant (position 5), une préposition (position 9) et deux adjectifs (positions 6 et 8). Le déterminant et les adjectifs sont de nœuds finaux, qui ne sont pas eux-mêmes recteurs. La préposition « de », par contre régit un nom, « culture », qui lui-même régit un déterminant, « la », qui lui n'est pas recteur. Ces mots sont ajoutés au syntagme, qui est à présent maximal.

Pour chaque syntagme nominal construit, certaines informations sont associées :

- à cette occurrence de syntagme en particulier, représenté par un identifiant unique : genre, nombre (toujours singulier puisque l'on filtre pour l'instant les noms pluriels, mais une amélioration de cet outil pourrait consister à rechercher également les

syntagmes référentiels pluriels ; on garde donc la possibilité de modifier facilement ce point), type de déterminant, positions des mots régis ;

- à ce type de syntagme (tous les syntagmes identiques à l'exception du déterminant appartiennent au même type) : notamment, la fréquence est incrémentée, les coordonnées de chaque occurrence sont mémorisées.

La détection de SN potentiellement référentiels ne pose pas de problème particulier, puisque le seul critère est la présence de la tête dans une liste de mots-clés, et que c'est toujours le syntagme maximal qui est construit. Cela se justifie par le fait que le syntagme maximal sera toujours celui qui aura la référence la plus précise ; en construisant un syntagme plus petit que le syntagme maximal, on risque d'ailleurs de modifier la référence : « ministre de l'éducation » et « ancien ministre de l'éducation » ne renvoient pas au même référent. Les seules imbrications de syntagmes se produisent lorsqu'un syntagme contient deux références (car il exprime une relation du premier référent au second), comme avec l'énoncé « le porte-parole de la candidate », dans lequel deux syntagmes référentiels imbriqués seront balisés : *[le porte-parole de [la candidate]]*.

Comme déjà mentionné, le rappel est ici bien plus important que la précision, puisque ces syntagmes nominaux potentiellement référentiels subiront par la suite jusqu'à deux filtrages : lors de l'application des patrons (exclusion des types de syntagmes qui ne sont jamais rapportés par aucun patron) et lors de la résolution en corpus de chaque syntagme (exclusion des occurrences de syntagmes pour lesquelles aucun antécédent n'est trouvé).

### 2.1.3. Annotation des noms de personne

#### *2.1.3.1. Quels noms propres traiter ?*

Nous avons souligné plus haut que, étant donnée la visée applicative de notre outil, tous les noms propres ne présentent pas le même intérêt pour nous. Nous avons songé un moment ne traiter qu'une liste fermée de noms propres (comme la liste des candidats à la présidentielle), puis uniquement les noms propres dont la fréquence dépasserait un certain seuil. Finalement, ce sont tous les noms propres, même les hapax, qui seront balisés par le module, puis qui seront traités par les modules suivants. Nous avons fait ce choix pour réduire les possibilités d'erreurs d'attribution. En effet, aucune discrimination n'étant faite selon la fréquence pour les syntagmes nominaux potentiellement référentiels, on aurait tort d'en faire pour les noms propres : si certains noms propres sont exclus, les syntagmes dont ils sont coréférents ne pourront être correctement résolus, et le programme cherchera pour eux un autre antécédent.

Lors de la projection, on pourra décider de ne projeter sur corpus que les noms propres dont la fréquence dépasse un certain seuil. On évaluera d'ailleurs, en plus des résultats pour tous les référents, les résultats obtenus pour les 50 référents les plus cités.

#### *2.1.3.2. Repérage des noms de personnes*

Lors du premier parcours, pour chaque nom propre rencontré, on cherche s'il est précédé d'un prénom (on utilise pour cela la relation NNPR de SYNTEX) ou d'une particule « M. », « Mme » ou « Mlle ». Si c'est le cas au moins une fois, on considère qu'il s'agit d'un nom de personne, qui pourra être étiqueté lors du second parcours.

Ce faisant, on mémorise quels sont les prénoms et les particules associés aux différents noms propres, et avec quelles fréquences.

Une pierre d'achoppement de cet algorithme se rencontre lorsque SYNTEX ne reconnaît pas le prénom associé à un nom de famille, comme cela arrive assez souvent pour des noms étrangers ; par exemple, « Azouz Begag » et « Rachida Dati » ne sont pas reconnus par notre programme comme étant des noms de personnes.

#### *2.1.3.3. Traitement des noms de personnes ambigus*

Certains noms de personnes reconnus comme tels peuvent aussi désigner parfois autre chose que des personnes. Par exemple, « Hollande » et « France », peuvent désigner, selon les occurrences, des personnes (*François Hollande* et *Cécile de France*) ou des lieux (*la Hollande*, *la France*). Pour repérer ces noms de personnes ambigus, on cherche si les noms propres reconnus comme étant des noms de personnes sont parfois précédés d'un déterminant ou d'une préposition « à » ou « en » (et on compte combien de fois cela arrive). Ce critère ne suffit bien sûr pas à discriminer les noms propres ambigus de ceux désignant exclusivement des personnes : on peut trouver "dire à X", "croire en X", ou X est un nom de personne ; on trouve même des occurrences de « un/le X », comme par exemple dans la phrase :

> Le véritable Sarkozy n'est pas celui du samedi et du dimanche, c'est le Sarkozy du Medef.

Mais si le rapport entre la fréquence du type *déterminant|préposition X* et la fréquence du type *particule|prénom X* dépasse un premier seuil, on considèrera que X est un nom de personne ambigu. Pour chaque occurrence d'un nom propre appartenant à la liste ainsi créée, l'ambiguïté sera levée en faisant appel au contexte (donc lors du second parcours – cf. 2.1.3.4. Etiquetage). Mais on le devine, « Hollande » et « France » ne doivent pas être traités de la même façon. Dans le cas de « Hollande », le référent sera dans la grande majorité des cas une

personne ; pour France, toutes les occurrences à l'exception d'une renvoient à un pays[14]. C'est pourquoi on a défini un second seuil. En deçà de ce seuil, le nom propre est considéré comme un nom de personne sauf dans certains contextes (présence d'un déterminant, de « à » ou de « en ») ; ce sera le cas de « Hollande ». Au delà du seuil, au contraire, le nom propre ne sera considéré comme un nom de personne que s'il se trouve dans certains contextes (présence d'une particule ou d'un prénom) ; ce sera le cas de France.

Ce traitement permet aussi d'exclure certains noms propres qui auraient été classés par erreur parmi les noms de personnes. Par exemple, la présence dans le corpus d'une phrase telle que

> Thierry Solère, M. Internet de l'UMP, (…) estime que 35 % du trafic sur les blogs UMP provient de ces liens sponsorisés.

amène notre programme à ajouter « Internet » à la liste des noms de personnes. Mais lors de la phase de traitement des noms propres ambigus, « Internet » dépasse largement les deux seuils (86 occurrences de « l'Internet » contre une seule de « M. Internet »), et finalement, seule l'occurrence « M. Internet » sera étiquetée comme désignant une personne – certes à tort (on verra que cette façon de traiter les noms propres peut poser quelques problèmes), mais on aura au moins évité de générer une quantité importante de bruit en étiquetant toutes les occurrences d' « Internet » comme référent à ce « M. Internet » créé par l'outil.

#### *2.1.3.3. Association des noms propres aux référents qu'ils désignent*

Le but est d'associer (virtuellement) à chaque nom de personne un ou plusieurs (si plusieurs personnes partagent le même nom de famille) identifiants correspondant à un référent du monde réel. A chaque identifiant correspond un et un seul référent, à chaque référent correspond un et un seul identifiant ; on parlera donc indifféremment d'identifiants ou de référents, selon qu'on aura un point de vue plutôt interne ou externe. Ce dernier est de la forme *nom_prenom* si le prénom apparaît au moins une fois dans le corpus, et de la forme *nom* si aucun prénom n'est spécifié, comme cela arrive parfois pour les noms propres à faible fréquence. A chaque identifiant est associé un genre.

Si un nom de personne n'est jamais associé qu'à un seul prénom, on considère qu'il réfère toujours à la même personne, désignée par l'identifiant *nom_prenom*. On adjoint à cet

[14] Cette seule occurrence est d'ailleurs assez « accidentelle » :
« François Hollande était, tard dans la soirée, mardi 12 septembre, l'invité de Marc-Olivier Fogiel dans la nouvelle émission, "T'empêches tout le monde de dormir", que ce dernier présente sur M6. Il était même si tard qu'on était déjà mercredi. (…) La comédienne Cécile de France attendait son tour en balançant sa jambe sur son tabouret, l'air de s'ennuyer furieusement. »

identifiant des informations, notamment son genre, comme nous le verrons dans la sous-partie qui suit.

S'il est associé à plusieurs prénoms, ce qui est assez souvent le cas[15] (il est rare qu'aucune référence ne soit faite aux membres de la famille d'une personnalité politique importante), un identifiant par prénom différent est construit.

Un problème est à ce stade posé par les noms de famille qui tout en étant associés avec plusieurs prénoms différents continuent de référer à la même personne. C'est le cas par exemple pour George W. Bush (l'initiale W. est bien reconnue par SYNTEX comme étant un prénom), qui peut être désigné par « George Bush » ou par « George W. Bush ». Pour écarter ce problème, l'outil, lorsque deux prénoms sont associés à une occurrence d'un nom de famille, n'utilise que le premier (supposé être le plus couramment utilisé) pour créer l'identifiant. Ainsi, « George Bush » et « George W. Bush » sont tous deux classés comme référent à *bush_george*. Seules d'éventuelles (il n'y en a pas dans notre corpus) occurrences de « W. Bush [sans George] » seraient encore mal attribuées ; dans la mesure où le nombre de cas concernés par ce problème est très faible, nous n'avons pas poussé plus avant sa résolution.

#### *2.1.3.4. Détermination du genre de chaque référent*

Toujours lors du premier parcours, on associe un genre à chaque identifiant/référent, en se fondant sur les particules (M., Mme, Mlle) qui lui sont adjointes. Si un nom de famille n'est jamais associé qu'à une seule particule, le ou les identifiants correspondant à ce nom de famille se verront affecter le genre de cette particule (masculin pour M., féminin pour Mme ou Mlle). Dans le cas où plusieurs genres sont associés à un nom de famille, il faudra associer le bon genre au bon identifiant. Par exemple, au nom de famille Le Pen, deux prénoms sont associés dans notre corpus (« Jean-Marie » et « Marine ») et deux particules (« M. » et « Mme »). Puisque deux prénoms ont été extraits, deux identifiants sont créés : *le pen_jean-marie* et *le pen_marine*, auxquelles il s'agit d'assigner un genre. L'algorithme fonctionne de la manière suivante :

- Si une occurrence du type *particule1 prénom1 nom1* a été vue, le genre associé à l'identifiant *nom1_prénom1* est celui qu'indique la particule *particule1*. Mais cela reste plus rare qu'on serait tenté de le croire. Par exemple, il n'y a aucune occurrence de « Mme Marine Le Pen » dans notre corpus, pour 13 occurrences de Mme Le Pen et 152 occurrences de Marine Le Pen.

[15] C'est le cas d'environ 15% des noms propres

- Sinon, on met en correspondance les fréquences des différents prénoms associés avec celles des différences particules associées. Par exemple, au nom de famille Le Pen, deux prénoms sont associés dans notre corpus (« Jean-Marie » et « Marine ») et deux particules (« M. » et « Mme »). Puisque deux prénoms ont été extraits, deux identifiants sont créés : *le pen_jean-marie* et *le pen_marine*, auxquelles il s'agit d'assigner un genre. Comme « Jean-Marie Le Pen » est beaucoup plus fréquent que « Marine Le Pen » et que « M. Le Pen » et « Mme Le Pen » ont un rapport de fréquences similaire, c'est *le pen_jean-marie* qui se verra attribuer le genre masculin, et *le pen_marine* le genre féminin.

#### *2.1.3.5. Balisage*

Le balisage a lieu lors du deuxième parcours du fichier. On adjoint alors à chaque occurrence de nom de personne l'identifiant qui lui correspond. Si le nom rencontré est sans ambiguïté un nom de personne, et qu'un seul identifiant lui est associé (s'il n'a jamais été vu qu'avec un seul prénom), il est étiqueté sans plus de formalités.

Si la catégorie du nom propre est ambiguë (il peut s'agir d'un patronyme mais aussi, par exemple, d'un toponyme), on désambiguïse grâce au contexte comme nous l'avons décrit en *2.1.3.2*.

S'il y a ambiguïté sur l'identifiant à associer au nom propre, l'algorithme procède ainsi :

- Si, pour l'occurrence considérée, le prénom est spécifié, l'identifiant sera celui qui correspond à ce prénom ;
- Si le prénom n'est pas spécifié :
    - Si un identifiant a été associé à la dernière occurrence du même nom de famille au sein du même article, il sera également associé à cette nouvelle occurrence ;
    - Sinon, le référent sera la personne la plus citée dans le corpus parmi celles qui portent le même nom de famille.

Le syntagme à baliser est construit par un parcours à peine plus complexe des étiquettes données par SYNTEX que pour les syntagmes nominaux. Seules deux relations nous intéressent : la relation NNPR et la relation EPI : si le nom de famille considéré régit un prénom ou plusieurs prénoms, ces derniers sont ajoutés au syntagme nom propre ; s'il est régi par un nom commun (relation EPI), et que ce nom fait partie de la liste de mots-clés, ce nom, ainsi que tous les éléments qu'il régit, sont également ajoutés au syntagme construit. Par exemple, pour le syntagme « la candidate socialiste Ségolène Royal » :

```
DetFS|le|la|12|DET;13|
NomFS|candidat|candidate|13|NOMPREP;11|DET;12,EPI;16,ADJ;14
```

AdjFS|socialiste|socialiste|14|ADJ;13|
NomPrXXPrenom|Ségolène|Ségolène|15|NNPR;16|
NomPr|Royal|Royal|16|EPI;13|NNPR;15

l'algorithme part de « Royal », qui régit un prénom : « Ségolène », et qui est régi par un nom : « candidate ». Les mots appartenant au syntagme nominal dont « candidate » est la tête sont eux aussi adjoints au syntagme ; il s'agit ici du déterminant « la » et de l'adjectif « socialiste ».

Comme pour les syntagmes nominaux, deux structures de données sont créées : l'une adjoignant des informations comme le genre et la fréquence totale aux identifiants *nom_prénom*, l'autre précisant pour chaque occurrence de syntagme nom propre sa forme et l'identifiant auquel il a été associé.

## 2.2. Module de calcul des scores d'association

### 2.2.1. Principe

Ce module prend en entrée la sortie du module de balisage ; son but est d'extraire des couples formés d'un nom propre (qui donne la référence) et d'un syntagme nominal (qui sera potentiellement coréférent au nom propre) et de chiffrer leur degré d'association. Deux types de méthodes sont utilisés pour l'extraction des couples : des patrons et des méthodes basées sur la cooccurrence. Toutes les mesures de liaison précédemment décrites sont applicables, selon le choix de l'utilisateur. Le module fournit en sortie deux nouvelles structures de données : l'une donnant pour chaque syntagme nominal les noms propres qui lui sont associés, avec leur fréquence et score d'association, et l'autre donnant pour chaque nom propre les syntagmes associés et leur fréquence et score d'association

### 2.2.2. Méthodes d'extraction des couples

L'utilisation des patrons, des cooccurrences ou des deux types de méthodes est paramétrable par l'utilisateur. L'avantage des patrons utilisés est qu'ils sont suffisamment précis pour ne générer que très peu de bruit. Si l'on recherche une très grande précision, on pourra les utiliser seuls. Les cooccurrences au sein d'un contexte large (fenêtre de trois phrases ou article entier) permettent de « ramasser » beaucoup plus de syntagmes potentiellement coréférentiels pour chaque nom propre), mais avec aussi beaucoup de bruit.

#### *2.2.1.1. Extraction par patrons*

- **SN *virgule* NPR *Typo*** et **NPR *virgule* SN *Typo***

Ce patron permet d'extraire les couples contenus dans les phrases suivantes :

> **Le candidat de l'UDF, François Bayrou,** a jugé que les propos de Lang relevaient d'une «grave imprudence» (…)
>
> **Corinne Lepage, la candidate de Cap 21,** dans son projet de Constitution, souhaite qu'un tiers des députés soient désormais élus à la proportionnelle (…)

Ces constructions sont extrêmement courantes dans notre corpus ; elles sont souvent utilisées pour introduire de nouveaux référents en déclinant nom et fonction, mais on les trouve également pour les personnes les plus citées de notre corpus (par exemple, il y a 112 occurrences de « ministre de l'intérieur, Nicolas Sarkozy » dans le corpus).

En demandant un signe typographique (qui ne soit pas un guillemet) après le second élément du couple, on se limite aux appositions et on réduit beaucoup le bruit, en évitant de construire des couples tels que *sarkozy_nicolas* / *candidate socialiste* ou avec des phrases comme :

> S'emparant du thème de l'identité nationale, avancé d'abord par **Nicolas Sarkozy, la candidate socialiste** a en effet appelé vendredi les Français à avoir chez eux un drapeau tricolore et à l'exposer à leurs fenêtres le jour de la fête nationale.

L'automate qui correspond au patron **SN *virgule* NPR *Typo*** fonctionne comme suit :

- Si l'étiquette *j* de la séquence *i* appartient à un syntagme nominal balisé ;
- que l'étiquette qui suit directement le syntagme nominal balisé (étiquette *j+longueur du SN* de la séquence *i*) est une virgule (est reconnue par l'expression régulière /^Typo\|,\|**/) ;**
- que l'étiquette *j+longueur du SN+1* de la séquence *i* appartient à un syntagme nom propre balisé ;
- et que l'étiquette *j+longueur du SN+longueur du SNPR+2* de la séquence *i* est un signe typographique (sauf guillemet) ;

⇨ alors, le SN et le SNPR rencontrés sont en possible relation de coréférence ; la fréquence de leur association est incrémentée.

L'automate pour le patron **NPR *virgule* SN *Typo*** est similaire.

- **NPR (SN) / NPR – SN –** et **SN (NPR) / SN – NPR –**

Ce patron est très similaire aux précédents, et est d'ailleurs appliqué en même temps (avec le même automate) ; la seule différence est que si le premier signe de ponctuation rencontré est non une virgule mais une parenthèse ou un tiret, alors le second devra être identique.

- **Phrase ne contenant qu'un seul NPR *point* SN démonstratif**

Nous avons vu en *2.2.3.3.* qu'un syntagme démonstratif permet souvent d'introduire une nouvelle façon de désigner un référent cité précédemment ; son avantage est qu'il permettra d'extraire des syntagmes coréférentiels moins « canoniques » qu'avec les patrons précédents.
L'automate fonctionne comme suit :

- Si une séquence *i* ne contient qu'un seul nom propre ;
- Et que la première étiquette de la séquence *i+1* et un démonstratif appartenant à un syntagme nominal balisé ;

⇨ alors, le nom propre de la séquence *i* et le syntagme nominal de la séquence *i+1* sont peut-être coréférents ; leur fréquence d'association est incrémentée.

- **NPR SUJ – verbe d'état – SN ATTS**
- Si la tête d'un syntagme nom propre est en relation sujet (SUJ) avec un verbe, c'est-à-dire que la partie de son étiquette correspondant aux mots recteurs est reconnue par l'expression régulière /SUJ;([0-9]+)/ ;
- que le verbe pointé par cette relation régit lui-même un nom attribut du sujet ; c'est-à-dire que la partie de son étiquette correspondant aux mots régis est reconnue par l'expression régulière /ATTS;([0-9]+)/ ;
- et que le nom pointé est la tête d'un syntagme nominal balisé ;

⇨ alors, le nom propre et le syntagme nominal extraits sont possiblement coréférentiels.

#### *2.2.1.2. Méthodes d'extraction basées sur la cooccurrence*

Les cooccurrences sont utilisées pour trouver des couples plus éloignés ; elles sont bien moins précises que les patrons, puisque le seul critère est la coprésence d'un syntagme nominal et d'un nom propre dans une certaine fenêtre, sans que les deux éléments n'aient besoin d'être en relation syntaxique ou d'apposition. Nous avons toutefois renoncé aux méthodes qui généraient trop de couples et donc trop de bruit, comme la cooccurrence au sein d'un même paragraphe ; si par exemple un paragraphe contenait trois citations de personnes, chacune accompagnée d'une description définie (ce qui est loin d'être excessif), neuf couples étaient créés, dont les deux tiers à tort. C'est pourquoi la méthode suivante demande non seulement la coprésence d'un nom propre et d'un syntagme nominal, mais aussi l'absence d'un autre nom propre.

- **Cooccurrence au sein des trois premières phrases d'un article**

En observant notre corpus, nous avons constaté que très souvent, dans les trois premières phrases d'un article, le sujet est repris sous différents angles (citation directe / citation par la fonction). Cela est dû à ce que la première phrase est souvent un titre (mais cela n'est pas mis en valeur dans notre corpus), et que la phrase suivante explicite ce titre ; les articles du Figaro, en particulier, suivent souvent ce modèle – les deux exemples suivants en sont tirés :

> **Nicolas Sarkozy** rappelle à l'ordre ceux qui parlent en son nom
> **Le ministre de l'Intérieur** s'agace que certains proches parlent en ordre dispersé, et parfois de façon contradictoire.
>
> 2007 : **Alliot-Marie** fait un pas de plus
> **Le numéro trois du gouvernement** franchit une nouvelle étape vers une candidature en installant, à Paris, son association Le Chêne, en présence de plusieurs dizaines de parlementaires et d'experts.

L'algorithme procède de la manière suivante, pour chaque article :

- Tous les syntagmes balisés des trois premières phrases sont relevés ;
- Si une seule personne est citée dans ces trois phrases (éventuellement plusieurs fois), tous les syntagmes nominaux relevés sont considérés comme potentiellement coréférentiels avec cette personne ; la fréquence de chaque couple est incrémentée.
- **Cooccurrence au sein d'un article dont le NPR est le sujet**

Il arrive souvent qu'un article s'intéresse à une personnalité politique en particulier ; lorsque cela advient, le besoin d'utiliser des expressions coréférentielles est important, pour éviter la redondance. On suppose donc qu'il y aura dans ces articles beaucoup d'expressions coréférentielles à leur sujet. L'algorithme procède ainsi :

- Tous les syntagmes balisés de chaque article sont relevés et comptés ;
- La personne la plus citée de l'article est considérée comme en étant le sujet si sa fréquence de citation est d'au moins deux fois supérieure à celle de la seconde personne la plus citée ; si c'est le cas, la fréquence des couples qu'elle forme avec chaque syntagme nominal du corpus est incrémentée.
- Lorsque tous les articles ont été traités, pour chaque nom propre, les coréférents potentiels trouvés par cette méthode et dont la fréquence d'association avec le nom propre est inférieure à deux sont supprimés, afin de limiter – légèrement – le bruit.

Cette méthode d'extraction a la particularité de ne trouver de coréférents potentiels que pour les personnes les plus « importantes » du corpus, les noms propres à faible fréquence n'étant que très rarement détectés comme étant le « sujet » d'un article ; cela n'est pas gênant, tout d'abord parce que ce sont ces noms propres qui nous intéressent le plus, comme nous l'avons

déjà souligné, et ensuite parce que les noms propres à forte fréquence ont souvent plus de coréférents à trouver. Plus gênant est bien sûr le bruit important qu'elle génère ; cela nous a conduit à la sous-pondérer lorsqu'elle est utilisée.

### 2.2.3. Application des mesures de liaison

On applique ensuite sur l'ensemble des couples la mesure de liaison qui aura été choisie par l'utilisateur. Au final, on obtient pour les deux orientations (SN->NPR et NPR->SN) une structure de données qui associe à chaque couple leur fréquence et leur score d'association, ainsi que le rang du second élément du couple parmi tous les couples (classés par mesure d'association) dont le premier élément fait partie.

## 2.3. Module de projection sur le corpus / résolution en contexte

### 2.3.1. Principe

Ce module tente successivement de résoudre (d'attribuer à un référent désigné par un nom propre de l'article) tous les syntagmes nominaux balisés. Pour chaque syntagme, selon son type, il accordera plus ou moins d'importance aux critères contextuels ou non-contextuels (scores d'association calculés par le module précédent). S'il y a plusieurs antécédents[16] possibles, le module tranche ; si aucun antécédent n'est trouvé, le syntagme est laissé de côté.

### 2.3.2. Indices contextuels utilisés

Les indices suivants peuvent, selon le type de syntagme, soit fonctionner comme des filtres (en excluant les antécédents qui ne répondent pas aux critères), soit augmenter ou diminuer la saillance de l'antécédent.

#### *2.3.2.1. Compatibilité en genre*

Ce critère joue un rôle de filtre pour tous les types de syntagmes dans un cas de figure uniquement : si le nom propre est masculin et l'antécédent féminin – ils sont alors considérés incompatibles. En effet, si le nom propre est féminin et l'antécédent masculin, on ne peut considérer qu'il y a incompatibilité : il arrive bien souvent que la fonction d'une femme soit un nom masculin, comme dans :

> «Dans l'intérêt du service», dit la lettre signée par **le** ministre de la Défense, **Michèle Alliot-Marie**.

[16] Pas « antécédent », on désigne le nom propre de l'article qui a le même référent que le syntagme à résoudre. Ce terme n'est pas toujours très adapté, dans la mesure où le nom propre recherché peut être en position anaphorique comme cataphorique, et où il laisse entendre que la relation n'est pas symétrique – ce qui n'est pas toujours vrai. Nous l'emploierons malgré tout de manière généralisée pour ne pas multiplier les appellations.

La saillance de l'antécédent ne sera dans ce cas ni augmentée ni diminuée ; il en sera de même si l'un des genres est inconnu. En revanche, la saillance est augmentée si les éléments sont tous deux masculins ou tous deux féminins.

#### *2.3.2.2. Eloignement, position anaphorique / cataphorique*

L'éloignement de l'antécédent par rapport au syntagme à résoudre est donné en nombre de phrases. Dans le cas des syntagmes définis, plus l'antécédent est éloigné, plus sa saillance sera faible. Pour les autres types de syntagmes, l'éloignement tient en plus lieu de filtre, puisque les antécédents se trouvant hors d'une certaine fenêtre sont exclus.

#### *2.3.2.3. Fonctions syntaxiques*

De manière générale, un antécédent sujet (SUJ) est préféré à un antécédent objet (OBJ), qui est lui-même préféré à un antécédent qui n'est ni sujet ni objet. En plus de cela, la saillance d'un antécédent qui a la même fonction syntaxique que le syntagme à résoudre est augmentée. Enfin, certaines règles basées sur la syntaxe permettent d'exclure des antécédents : le syntagme à résoudre et son éventuel coréférent ne peuvent se trouver dans certaines constructions syntaxiques :

- Ils ne peuvent être sujet et objet du même verbe ;

  *Nicolas Sarkozy attaque le ministre de l'Intérieur

- Ils ne peuvent être en relation de coordination.

  *Nicolas Sarkozy et le ministre de l'Intérieur…

### 2.3.3. Utilisation combinée de la saillance (critère contextuel) et des scores d'association (critère non-contextuel)

Le but est d'assigner une valeur à chaque antécédent, afin de les classer et de pouvoir sélectionner le meilleur. Cette valeur doit faire intervenir à la fois les critères contextuels et non-contextuels. En faisant intervenir les scores d'association dans des formules, nous avons constaté que selon la mesure de liaison utilisée, ces derniers avaient plus ou moins de poids ; plutôt que de faire directement intervenir les scores d'associations, nous avons donc décidé de prendre en compte le rang. seconds éléments des couples lorsqu'ils sont classés par score d'association avec le premier élément, et réciproquement. On multiplie les deux chiffres (rang du premier élément pour le second et du second pour le premier) pour obtenir un nouveau score ; plus ce score est bas, plus fort est le degré d'association du couple. Par exemple, pour déterminer à qui réfère le syntagme « candidat de l'UMP », on prend en compte les rangs de

chaque personne associée à ce syntagme (classées par scores d'association) – on parlera de $rang_{SN/NPR}$ :

<candidat de l' UMP>
1) sarkozy_nicolas -1.51732262352629
2) hulot_nicolas -3.51898041731854
3) guéant_claude -3.62434093297637
4) hortefeux_brice -3.96081316959758

et le rang de ce syntagme pour chacune de ces personnes – on parlera de $rang_{NPR/SN}$ :

<sarkozy_nicolas>
1) ministre de l' intérieur 2.82571972573927
2) président de l' UMP 1.58477778286878
3) ministre de l' Intérieur 1.09666106853685
4) candidat -1.23046325782598
5) candidat de l' UMP -1.51732262352629

<hulot_nicolas>
1) futur candidat à la présidentielle -0.980829253011726
(…)
11) candidat de l' UMP -3.51898041731854

<guéant_claude>
1) directeur de cabinet du ministre -1.08618976866955
(…)
13) candidat de l' UMP -3.62434093297637

<hortefeux_brice>
1) ministre délégué aux collectivités territoriales 0.0971637484536477
(…)
16) candidat de l' UMP -3.96081316959758

Les scores des différents couples sont les suivants :

candidat de l'UMP / *sarkozy_nicolas* : 1*5=5

candidat de l'UMP / *hulot_nicolas* : 2*11=22

candidat de l'UMP / *guéant_claude* : 3*13=36

candidat de l'UMP / *hortefeux_brice* : 4*16=64

On le voit, *sarkozy_nicolas* est largement favorisé.

Pour que la prise en compte de la saillance soit à peu près similaire à celle des scores d'association, nous avons choisi de classer tous les antécédents possibles par saillance décroissante, et de retenir également leurs rangs ($rang_{saillance}$) ; ainsi, c'est également le score le plus faible qui prime.

### 2.3.4. Résolution des différents types de syntagmes

#### *2.3.2.1. Syntagmes définis*

Le traitement est différent selon que le syntagme soit considéré complet ou incomplet. Pour distinguer (assez sommairement) entre syntagmes complets et incomplets, on utilise le critère suivant : un syntagme est complet s'il fait plus de deux mots, si le nombre de couples auxquels il appartient est inférieur à 5[17] et si la différence entre le score d'association de son premier couple et celui du second est importante ; sinon, il est incomplet. On reprendra cette distinction complet/incomplet pour les syntagmes autres que définis, bien qu'elle n'ait plus de validité théorique, car empiriquement, on a constaté que les syntagmes plus longs et associés à moins de personne avaient une plus grande autonomie vis-à-vis du contexte, même lorsqu'il s'agit de syntagmes non définis.

2.3.2.1.1. Syntagmes définis complets

L' « antécédent » d'un syntagme défini complet peut se trouver n'importe où dans l'article ; on choisit l'antécédent compatible en genre pour lequel le score $rang_{SN/NPR}$*$rang_{NPR/SN}$ est le plus faible. Ainsi, le seul critère « contextuel » utilisé est la compatibilité en genre qui doit être respectée.

2.3.2.1.2. Syntagmes définis incomplets

L'antécédent peut se trouver n'importe où dans l'article, mais on privilégie lors du calcul de la saillance un moindre éloignement et la position anaphorique ; l'antécédent choisi est l'antécédent compatible en genre pour lequel le score $rang_{saillance}$*$rang_{SN/NPR}$*$rang_{NPR/SN}$ est le plus faible.

#### *2.3.2.2. Syntagmes indéfinis*

Les syntagmes indéfinis sont résolus de la même manière que les syntagmes définis (avec les mêmes formules selon qu'ils soient « complets » ou « incomplets »), à la différence notable près que l'antécédent est cherché dans une fenêtre de deux phrases seulement : la phrase dans laquelle apparaît le syntagme et la phrase suivante, et que c'est la position cataphorique qui est préférée à l'anaphorique.

Dans la mesure où la fenêtre est relativement réduite, il arrivera souvent qu'aucun antécédent ne soit trouvé et que le syntagme ne soit pas résolu ; cela est volontaire, car on a constaté que

[17] Ce critère n'est utilisé que si on n'a utilisé que des patrons lors du module précédent ; sinon, le bruit est trop important.

bien souvent, les syntagmes indéfinis sont génériques, et donc ne font référence à aucune entité particulière ; il ne faut donc pas chercher à les résoudre à tout prix.

#### *2.3.2.3. Syntagmes démonstratifs*

Les syntagmes démonstratifs sont toujours résolus en faisant appel à la saillance (pas de distinction complet/incomplet) ; la fenêtre dans laquelle on cherche l'antécédent est de deux phrases : la phrase dans laquelle apparaît le syntagme et la phrase précédente.

#### *2.3.2.4. Syntagmes possessifs*

Les syntagmes possessifs sont résolus de la même manière que les syntagmes indéfinis. Là encore, la petitesse de la fenêtre fait que beaucoup de syntagmes ne sont pas résolus ; cela se justifie par le fait que leur référence est souvent indéterminée.

#### *2.3.2.5. Syntagmes sans déterminant*

Les syntagmes sans déterminant se trouvent surtout dans certaines constructions : en apposition après un nom propre :

> François Bayrou, candidat UDF, (…)

coordonnés avec un autre syntagme nominal coréférentiel :

> (…) candidat UDF et ancien ministre de l'Education

ou en position d'attribut du sujet :

> Sarkozy est ministre d'État depuis quatre ans.

Ces constructions seront privilégiées dans la recherche de l'antécédent ; sinon, ce dernier est cherché dans une fenêtre de deux phrases : la phrase contenant le syntagme et la phrase précédente.

### 2.3.4. Syntagmes non résolus

Les syntagmes non-résolus sont de plusieurs types : certains, bien sûrs, sont référentiels, auraient pu être résolus et ne l'ont pas été ; d'autres sont non-référentiels (à des personnes), ou correspondent à des emplois génériques (pas de référence particulière) ; d'autres enfin sont référentiels à une entité particulière, mais indéterminée.

# 3. Evaluation

Il y a plusieurs choses à évaluer dans notre programme : l'extraction des expressions référentielles et leur attribution au bon référent sont deux choses différentes, qui doivent être évaluées séparément. Dans la mesure où les résultats de ces deux tâches évoluent avec le passage de chaque module, nous avons décidé d'évaluer notre programme module par module, avant d'en donner une évaluation globale. Pour chaque module, nous évaluerons l'apport de certains paramètres pouvant être modifiés (méthodes d'extractions et mesure de liaison utilisées pour le module consacré au calcul des scores d'association, indices contextuels utilisés pour le module de projection sur le corpus).

## 3.1. Evaluation des expressions référentielles extraites

### 3.1.1. Expressions référentielles potentielles

Le module de balisage devait repérer des expressions potentiellement référentielles, notamment à partir d'une liste de mots-clés ; on évalue ici non l'attribution, mais la « référentialité » des expressions extraites par le module.

| | |
|---|---|
| Précision | 88,52% |
| Rappel | 84,44% |

#### *3.1.1.1. Noms propres*

Pour les noms propres en particulier, les résultats sont les suivants :

| | |
|---|---|
| Précision | 95,52% |
| Rappel | 88,91% |

- Silence

Si l'on observe le silence, on se rend compte que c'est en particulier sur les prénoms isolés que bute notre programme. En effet, rien n'a été fait pour détecter les prénoms sans nom de famille comme des noms de personne ; or, dans le corpus d'évaluation, plusieurs articles résument les intentions de vote de personnes interrogées et citées seulement par leurs prénoms.

- Bruit

Il y a peu de bruit pour l'annotation des noms propres ; il est généré par des expressions ayant l'apparence de noms de personnes mais n'en étant pas, comme par exemple « Lady Fitness », nom d'une salle de sport.

#### *3.1.1.2. Syntagmes nominaux*

Les résultats pour les syntagmes nominaux sont les suivants :

| | |
|---|---|
| Précision | 71,12% |
| Rappel | 86,80% |

- Silence

Le silence est généré par des syntagmes nominaux dont la tête ne fait pas partie des mots-clés, comme par exemple « la réincarnation de Mitterrand », utilisé pour désigner Ségolène Royal.

- Bruit

Le bruit très important est généré par le balisage de syntagmes qui sont en emploi générique, comme dans « le statut pénal du chef de l'état », et qui donc n'ont pas été annotés manuellement.

### 3.1.2. Evaluation des expressions référentielles projetées sur le corpus

Les expressions référentielles qui ont résisté aux filtres des modules de calcul des scores d'association et de résolution en contexte sont finalement projetées sur le corpus. Les paramètres par défaut qui permettent d'obtenir les valeurs ci-dessous sont : utilisation de toutes les méthodes d'extraction sauf la cooccurrence par article (qui génère trop de bruit), de l'information mutuelle au cube comme mesure de liaison, et de tous les indices contextuels. Nous avons beaucoup, dans notre programme, favorisé la précision sur le rappel (notamment en ne cherchant pas à tout prix un antécédent pour tous les syntagmes nominaux). Les résultats s'en ressentent :

| | |
|---|---|
| Précision | 94,82% |
| Rappel | 66,26% |

Comme on le voit, le gain en précision est conséquent (+6,30%), mais le rappel est très – trop – affecté (-18,18%). Cela est surtout dû au module de résolution en contexte qui, lorsqu'il ne trouve pas d'antécédent, exclut le syntagme. Or, lors de l'annotation manuelle, les références indéterminées ont été annotées (avec comme mention « inconnu »), ce que ne fait pas notre programme, chacune d'entre elles ajoute donc au silence. De plus, un syntagme qui n'aura pas

été extrait par aucune méthode du module de calcul des scores d'association n'est jamais attribué, et donc jamais projeté sur le corpus ; pour augmenter le rappel, il faudrait donc améliorer le rappel de ces méthodes.

## 3.2. Evaluation de l'attribution

### 3.2.1. Evaluation globale

Toujours avec les mêmes paramètres par défaut, pour une expression référentielle correctement extraite, l'attribution est bonne à 86,06%. Les résultats pour les deux tâches successives d'extraction puis d'attribution (obtenus en multipliant précision et rappel des expressions projetées par le pourcentage de bonne attribution) sont donc :

| | |
|---|---|
| Précision | 82,61% |
| Rappel | 57,03% |
| F-mesure | 67,14% |

Nous allons voir l'apport des modules de calcul des scores d'association et de résolution en contexte à ces résultats.

### 3.2.2. Apport de chaque méthode d'extraction

#### *3.2.2.1. Patrons*

- Patron 1 (basé sur les signes typographiques)

Ce patron est celui qui renvoie le plus de couples, et qui apporte les plus aux résultats. Si l'on **n'utilise plus** ce patron, les résultats globaux sont de :

| | |
|---|---|
| Précision | 79,69% |
| Rappel | 48,84% |
| F-mesure | 60,56% |

Son apport est donc le suivant :

| | |
|---|---|
| Précision | +2,92 |
| Rappel | +8,19 |
| F-mesure | +6,58 |

Pour la suite, on ne donnera que le tableau d'apport.

- Patron 2 : NPR être SN

| | |
|---|---|
| Précision | +0,52% |
| Rappel | +0,33% |
| F-mesure | +0,07% |

- Patron 3 : Phrase contenant NPR. Ce SN

| | |
|---|---|
| Précision | +0,81% |
| Rappel | 0 |
| F-mesure | +0,07% |

#### *3.2.2.2. Cooccurrencess*

- Cooccurrence par article dont NPR est le sujet

Comme nous l'avons dit, nous n'utilisons pas, par défaut, cette méthode, à cause du bruit qu'elle génère. En l'utilisant en plus des autres, les résultats obtenus sont :

| | |
|---|---|
| Précision | 71,86% |
| Rappel | 61,13% |
| F-mesure | 66,06% |

Son « apport » est donc de :

| | |
|---|---|
| Précision | -10,75% |
| Rappel | +4,10% |
| F-mesure | -1,08% |

- Cooccurrence dans les trois premières lignes de l'article

Cette méthode fait partie de la configuration par défaut, mais son apport est en fait minime et plutôt négatif :

| | |
|---|---|
| Précision | -1% |
| Rappel | +0,01% |
| Fmesure | -0,01% |

Finalement, les patrons semblent plus intéressants que les méthodes basées sur la cooccurrence ; chacun d'entre eux permet d'améliorer à la fois le rappel et la précision.

### 3.2.3. Apport du choix de la mesure de liaison

Nous avons testé les dix mesures de liaison proposées à l'utilisateur. Mais quelque soit la mesure utilisée, les résultats varient très peu, avec une f-mesure allant de 66,57% à 67,14%. Cela est sans doute dû au fait que nous n'utilisons pas directement les scores d'association obtenus, mais plutôt le classement qu'ils permettent de faire entre les couples. Plusieurs mesures de liaison permettent d'obtenir ce même résultat de 67,14% de f-mesure (ce qui veut dire que le classement qu'elles renvoient est similaire) ; il s'agit des : coefficient de Kulczinsky, coefficient d'Ochiai, coefficient de McConnoughy, information mutuelle au cube et coefficient de vraissemblance (loglike).

Nous avons donc gardé comme mesure par défaut l'information mutuelle au cube, qui donne effectivement de meilleurs résultats que l'information mutuelle simple (qui obtient le moins bon résultat avec 66,57% de f-mesure).

### 3.2.4. Apport des indices contextuels

- Compatibilité du genre

Le critère de la compatibilité du genre permet de gagner assez sensiblement en précision :

| | |
|---|---|
| Précision | +1,10% |
| Rappel | 0 |
| F-mesure | +0,04% |

- Eloignement

On regroupe sous cette appellation le critère de saillance, qui diminue la saillance des antécédents les plus éloignés (ou de ceux qui se trouvent du « mauvais côté » du syntagme à résoudre), et le filtre, qui rejette tous les antécédents ne se trouvant pas dans la bonne fenêtre. Dans la mesure où il s'agit d'un filtre important, on s'attendrait à ce qu'il diminue le rappel, mais cela n'est pas le cas, au contraire ; vraisemblablement, ce filtre permet d'éviter bien des erreurs d'attribution, et donc augmente le rappel en même temps que la précision.

| | |
|---|---|
| Précision | +4,37% |
| Rappel | +0,72% |
| F-mesure | +0,69% |

- Priorité accordées à certaines fonctions syntaxiques

Nous basant sur des travaux de résolution d'anaphores, nous avons choisi d'accorder une légère priorité aux antécédents selon qu'ils sont sujet, COD, ou rien de cela. Les résultats ne sont pas très probants.

| | |
|---|---|
| Précision | 0 |
| Rappel | 0 |
| F-mesure | 0 |

- Parallélisme des fonctions syntaxiques

Nous avons également choisi d'augmenter la saillance d'un antécédent ayant la même fonction syntaxique que le syntagme à résoudre ; là encore, les résultats ne s'en ressentent pas beaucoup.

| | |
|---|---|
| Précision | 0 |
| Rappel | 0 |
| F-mesure | 0 |

Visiblement, ces deux derniers critères ne sont pas assez contraignants pour avoir une influence sur le choix de l'antécédent, c'est pourquoi ils ne changent rien aux résultats. Ils ont peu de poids par rapport aux critères de compatibilité en genre et d'éloignement, et il faudrait sans doute revoir ce poids à la hausse.

## 3.5. Autres paramètres

### 3.5.1. Taille du corpus

En étendant le corpus jusqu'au 20 avril 2007, on obtient les résultats suivants :

| | |
|---|---|
| Précision | 80,19% |
| Rappel | 58,72% |
| F-mesure | 67,80% |

La différence avec les résultats précédents est de :

| | |
|---|---|
| Précision | -2,42% |
| Rappel | 1,69% |
| F-mesure | 0,66% |

Le rappel augmente (plus de syntagme sont engrangés dans les listes créées par le module d'association), mais la précision baisse (plus grand taux d'erreurs dans ces listes ?), et finalement les résultats ne diffèrent pas beaucoup.

### 3.5.2. Journaux

Comme nous l'avions déjà annoncé, les résultats varient quelque peu selon les journaux : Le Figaro (dont on avait noté qu'il se pliait bien à certaines méthodes d'extraction) obtient les meilleurs résultats, et Libération (qui se démarque le plus), les moins bons :

| | Le Monde | Le Figaro | Libération |
|---|---|---|---|
| Précision | 80% | 91,56% | 73,50% |
| Rappel | 58,11% | 67,79% | 45,87% |
| F-mesure | 67,32% | 77,90% | 56,49% |

### 3.5.3. Fréquence des noms propres/référents évalués

Dans la mesure où, comme nous l'avons plusieurs fois souligné, nous nous intéressons davantage à Nicolas Sarkozy et à Ségolène Royal qu'à Cécile de France, il nous a également paru intéressant de regarder les résultats que nous obtenions pour les 50 premiers noms propres (par la fréquence). Les résultats sont un peu supérieurs à ceux obtenus pour l'ensemble des noms propres, mais pas autant qu'on ne l'aurait souhaité :

| | |
|---|---|
| Précision | 85,49% |
| Rappel | 58,14% |
| F-mesure | 69,21 |

# 4. Limites et perspectives

## 4.1. Limites

Nous allons détailler ici quelques points qui montrent les limites de notre programme.

### 4.1.1. Traitement des syntagmes non-référentiels ou à référence indéterminée

Dans la mesure où c'est l'identification d'un référent qui distingue en dernier recours, parmi nos syntagmes nominaux potentiellement référentiels, entre ceux qui le sont effectivement et ceux qui ne le sont pas, aucune différence n'est faite entre éléments non-référentiels et ceux qui sont référentiels mais dont la référence est indéterminée. Comme nous l'avons souligné,

cela nuit beaucoup au rappel lors de la projection sur corpus : en effet, les syntagmes à référence indéterminée étant malgré tout référentiels, ils avaient été annotés manuellement.
En outre, malgré quelques efforts pour exclure les emplois génériques des syntagmes (en cherchant un antécédent dans une fenêtre réduite, et qui ait été associé au syntagme lors de l'application des patrons), ceux-ci sont très souvent étiquetés. En effet, même si l'emploi est générique lorsque l'on parle du « statut pénal du chef de l'Etat », « Jacques Chirac » n'est malgré tout jamais bien loin.
Ces problèmes n'en sont plus si l'on n'évalue non pas la projection sur corpus, mais les listes de couples obtenues à l'issue du module de calcul des scores d'association, telles qu'elles pourraient être utilisées par LexiMedia2007 (ce qui n'a pas été fait).

### 4.1.2. Association univoque nom propre/référent

L'association du nom propre au référent se fait de manière univoque : un nom propre (constitué d'un nom et éventuellement d'un prénom) = un référent. Cela pose problème, car un même individu peut être désigné par plusieurs « noms propres » : une abréviation (Sarko), des initiales (DSK), ou encore un nom propre tout différent comme avec le phénomène d'antonomase :

> Le docteur Folamour des finances publiques Nicolas Sarkozy (…)
>
> Le Zapata de Bordères [François Bayrou] (…)
>
> Le M. Internet de l'UMP (…)

ou avec l'appellation plaisante de François Hollande, « M. Royal ».

### 4.1.3. Traitement des prénoms isolés

Les prénoms isolés ne sont pas identifiés par notre programme comme référant à des personnes. Il est difficile de trouver un critère pour les extraire ; on ne peut en cela se baser sur les étiquettes de SYNTEX, qui n'annote comme prénoms que les noms propres appartenant à une liste de prénoms **et** suivis d'un autre nom propre (nom de famille). Par exemple, dans « Guillaume Bachelay », « Guillaume » est bien considéré comme un prénom, mais il ne l'est pas lorsqu'il apparaît seul :

> NomPrXXPrenom|Guillaume|Guillaume|1|NNPR;2|
> NomPrXXInc|Bachelay|Bachelay|2|CC;3|NNPR;1
>
> NomPr|Guillaume|Guillaume|43|OBJ;42|

Un effort pourrait toutefois être fait lorsqu'un prénom seul est employé après une occurrence, dans le même article, de ce même prénom accompagné de son nom de famille (cela arrive notamment pour Ségolène Royal).

## 4.2. Perspectives

### 4.2.1. Améliorations du programme

#### *4.2.1.1. Plus grande utilisation des informations offertes par SYNTEX ?*

Pour améliorer notre programme, on première voie serait d'exploiter davantage les informations offertes par SYNTEX. On l'a vu, le peu d'utilisation qu'on en a fait n'a pas toujours été utile (notamment avec les critères contextuels basés sur les fonctions syntaxiques).

#### *4.2.1.2. Plus de patrons ?*

Cette plus grande exploitation pourrait se faire à travers la programmation de nouveaux patrons ; en effet, chacun de nos patrons permet à la fois d'augmenter la précision et le rappel de notre programme ; en en appliquant davantage (basés sur certaines constructions syntaxiques), on pourrait progressivement améliorer nos résultats.

### 4.2.2. Elargissement du champ applicatif

#### *Traitement de la référence évolutive*

Afin que notre programme puisse continuer à analyser le corpus de LexiMedia2007 malgré les chamboulements politiques qui affectent les expressions référentielles, il serait intéressant de parvenir à traiter la référence évolutive. Nous avons prêté attention à la manière dont notre programme a annoté, à partir du 26 mars 2007, les occurrences de « ministre de l'Intérieur. Les premières occurrences post-26 mars de ce syntagme sont toutes bien annotées, dans la mesure où elles sont systématiquement accompagnée d'un adjectif : « ancien/ex- » lorsque la référence est à Nicolas Sarkozy, et « nouveau » quand il s'agit de désigner François Baroin (il ne s'agit alors plus du *même* syntagme). Par contre, par après, les occurrences de « ministre de l'Intérieur » (sans adjectif) ne sont bien annotées (comme désignant François Baroin) que lorsque le nom de Nicolas Sarkozy est absent de l'article (et qu'il ne peut donc être choisi comme antécédent). Il serait sans doute possible de se baser sur les marques qui accompagnent un changement (les adjectifs sus-cités) pour mettre à jour les associations entre une fonction et un nom propre.

D’autres phénomènes seraient intéressants à traiter, comme notamment la référence plurielle : il ne paraît pas impossible, lorsqu’un syntagme référentiel pluriel est accompagné d’une énumération de ses coréférents partiels (dont la référence est incluse dans la sienne), comme dans :

> (…) les députés Christian Blanc et Pierre Christophe Baguet (…)

de résoudre ses liens coréférentiels.

# CONCLUSION

Le travail de master que nous avons présenté dans ce mémoire avait pour but la réalisation d’un outil de résolution de coréférences dédié à un système de veille terminologique sur internet, LexiMédia2007. Dans une première partie de ce mémoire, nous avons posé les balises théoriques de notre travail. Nous avons discuté de l’application LexiMédia2007 en la positionnant d’un point de vue diachronique dans la lignée d’autres applications de lexicométrie et de veille terminologique, et d’un point de vue synchronique par rapport à d’autres outils statistiques dédiés aux élections présidentielles de 2007. L’analyse syntaxique des données en amont de tout traitement statistique offre une plus-value par rapport aux autres systèmes considérés ; dans LexiMédia2007, l’unité d’analyse n’est pas le mot graphique ou le lemme, mais le syntagme. C’est ce qui nous a amenée à nous pencher sur le problème de la résolution de coréférences ; en effet, c’est le syntagme nominal dans son entier (par exemple « la candidate socialiste ») qui est porteur de la référence, et non le nom isolé (« candidate »). Nous avons ensuite dressé un panorama des notions de référence et de coréférence, et de leur traitement en linguistique et en TAL. Cela nous a permis de mieux définir nos objectifs et notre méthode, en prenant en considération les spécificités des données que nous avions à traiter (un corpus important et très homogène). Nous avons ainsi choisi de faire appel à des mesures de liaison, ce qui consitue une originalité de notre travail. Une troisième point de notre état de l’art est donc consacré à ces mesures.

La seconde partie de ce mémoire est consacrée au travail informatique effectué. Nous décrivons l’outil informatique que nous avons réalisé, dont le but est de résoudre les liens coréférentiels entre syntagmes nominaux et noms propres dans un corpus donné, et d’extraire des listes de d’association *nom propre / syntagme coréférentiel.* Les résultats obtenus sont extrêmement encourageants : les expressions référentielles sont détectées avec 88.5% de précision et 84.4% de rappel ; leur attribution est bonne à 86.1%. Il s’agit de résultats très

compétitifs pour cette tâche. De plus, plusieurs voies sont encore ouvertes pour les améliorer, notamment par l'implémentation de nouveaux patrons, faisant un plus large usage de l'analyse syntaxique fournie par SYNTEX.

Beaucoup de choses restent en suspens dans ce travail ; l'évaluation présentée à l'issue de ce mémoire n'a pas vocation à en marquer la fin, mais plutôt une nouvelle étape. Le fait de disposer d'un corpus annoté et de pouvoir chiffrer les résultats de notre outil doit être la base d'améliorations futures.

# BIBLIOGRAPHIE

## 1. Analyse du discours, lexicométrie et politique


Asnes M. (2004). *Référence nominale et verbale : analogies et interactions.* Presses de Paris-Sorbonne.

Beaudouin, V. (2000). « Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle ». *Texto !* septembre 2000 [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html>.

Bourigault D. & Fabre C. (2000). « Approche linguistique pour l'analyse syntaxique de corpus ». *Cahiers de Grammaires*, n° 25, Université Toulouse - Le Mirail, pp. 131-151

Bourigault D. (2002). « Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus ». *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues.*

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. (2005). Syntex, analyseur syntaxique de corpus, in *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France.

Brunet É. (2000). « Qui lemmatise, dilemme attise ». *Lexicometrica*, n°2.

Chetouani L. (1998). « L'effet de serre entre les incertitudes scientifiques, la rhétorique politique et les registres de la lexicométrie ». *Séminaire ADEST du 10 mars 1998.* Disponible sur : <http://webu2.upmf-grenoble.fr/adest/seminaires/Chetouani.htm>

Collectif Saint-Cloud (A. Geffroy, P. Lafon, M. Tournier) (1974). « L'Indexation minimale. Plaidoyer pour une non-lemmatisation », ENS de Saint-Cloud. Communication au « Colloque sur l'analyse des corpus linguistiques : Problèmes et méthodes de l'indexation minimale », Strasbourg, 21-23 mai 1973.

Crouzet T. (2007). Le Cinquième pouvoir. *Bourin éditeur. Paru le 18 janvier 2007.*

Desmarchelier D. (2005). « Le sens du texte : entre opacité statistique et transparence énonciative ». *Corpus, Numéro 4 Les corpus politiques : objet, méthode et contenu*. Disponible sur : <http://corpus.revues.org/document.html?id=372&format=print>.

Dubois J. & Sumpf J. (1969). « Problèmes de l'analyse du discours », *Langages* 13.

Fouetillou G. « Le Web et le débat sur la constitution européenne en France ». *Disponible sur : http://www.observatoire-presidentielle.fr/?pageid=20*.

Gorcy G., Martin R., Maucourt J., et Vienney R., (1970). « Le traitement des groupes binaires », Cahiers de Lexicologie, n° 2, p. 15-46

Guilhaumou J. (1997). « L'analyse du discours et la lexicométrie ». *Lexicometrica*.

Guilhaumou J. (2002). « Le corpus en analyse de discours : perspective historique ». *Corpus, Numéro 1 Corpus et recherches linguistiques.* Disponible sur : <http://corpus.revues.org/document8.html>

Guiraud P. (1955-1958). *Index du vocabulaire du théâtre classique.* Tome I-V. Paris, C. Klincksieck.

Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris, P.U.F.

Heiden S., Tournier M. (1998). « Lexicométrie textuelle, sens et stratégie discursive ». In *Actes I Simposio Internacional de Análisis del Discurso*, Madrid.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

Labbé D. (1990). *Le vocabulaire de François Mitterrand*, Paris, Presses de la fondation nationale des sciences politiques.

Labbé, D. (2003). Corneille dans l'Ombre de Molière : Histoire d'une dècourverte. *Paris, Impressions Nouvelles.*

Maingueneau D. (1991). *L'analyse du discours. Introduction aux lectures de l'archive.* Paris, Hachette.

Mayaffre D. (2005-a). « De la lexicométrie à la logométrie. » *Astrolabe.* Disponible sur : <http://www.uottawa.ca/academic/arts/astrolabe>.

Mayaffre D. (2005-b). « Les corpus politiques : objet, méthode et contenu. Introduction », *Corpus, Numéro 4 Les corpus politiques : objet, méthode et contenu.*

Mayaffre D. (2006). « Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la Vème République ». *8es Journées internationales d'Analyse statistique des Données Textuelles*.

Mellet S. (2003). « Lemmatisation et encodage grammatical : un luxe inutile ? » *Lexicometrica*, numéro Spécial. Disponible sur : <http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1/spec1-texte2.htm)/>.

Mill J. S. (1724/1988). *Système de logique*. Mardaga.

Muller C. (1964). *Essai de statistique lexicale : L'illusion comique de P. Corneille.* Klincksieck, Paris.

Muller Ch. (1968). *Initiation à la statistique linguistique*. Paris, Larousse.

Muller Ch. (1984). «De la lemmatisation», préface à Lafon P., *Dépouillements et statistiques en lexicométrie*, Paris-Genève, Slatkine-Champion

Pincemin B. (2004). « Lexicométrie sur corpus étiquetés ». *7es Journées internationales d'Analyse statistique des Données Textuelles*.

Rouveyrol L. (2005). « Vers une logométrie intégrative des corpus politiques médiatisés. L'exemple de la subjectivité dans les débats-panel britanniques », *Corpus, Numéro 4 Les corpus politiques : objet, méthode et contenu*. Disponible sur : http://corpus.revues.org/document.html?id=293&format=print

Salem A. (1986). « Segments répétés et analyse statistique des données textuelles », *Histoire et mesure, I - N° 2*.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris, INaLF, Klincksiek.

Tournier M. (1985). « Sur quoi pouvons-nous compter ? Réponse à Charles Muller », *Etudes de philologie et de linguistique offertes à Hélène Nais*, *Verbum* (numéro spécial), Presses universitaires de Nancy.

## 2. Référence, coréférence

Baldwin B (1997). "CogNIAC : high precision coreference with limited knowledge and linguistic resources" In *Proceedings of ACL/EACL workshop on Operational factors in pratical, robust anaphora resolution, Madrid.*

Benveniste E. (1939/1966). *Problèmes de linguistique générale, 1*. Ch. IV "Nature du signe linguistique", p 49-55, Gallimard.

Bergler S., Witte R., Khalife M., Li Z., Rudzicz F. (2003) "Using knowledge-poor Coreference Resolution for Text Summarization". *Concordia University.*

Boudreau S. (2004). « Résolution d'anaphores et identification de chaînes de coréférence selon le contexte ». *Université de Montréal.*

Boudreau S., Kittredge R. (2006). « Résolution d'anaphore et identification des chaînes de coréférence : une approche minimaliste ». *JADT 2006, 8e journées internationales d'analyse des données textuelles.*

Chastain C. (1975). « Reference and Context ». in Language Ming and Knowledge, Minneapolis, University of Minnesota Press, p. 194-269.

Chinchor N. A. (1998). « Overview of MUC-7/MET-2 ». in Proceedings of the Seventh Message Understanding conference.

Corblin F. (1987). *Indéfini, défini et démonstratif : constructions linguistiques de la référence*. Librairie Droz, Paris.

Ducrot O., Schaeffer J.-M. *et al.* (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*. Paris, Seuil, p. 360-372, entrée « référence ».

Dupont M. (1998). « Le calcul de la référence dans un système de compréhension automatique limitée de corpus homogènes ». *La Référence, statut et processus, Travaux linguistiques du CERLICO.*

Engler R. (1968-1974). Edition critique des Cours de linguistique générale de Saussure. Wiesbaden, Otto Harrassowitz.

Frege G. (1892/1994). « Sens et dénotation ». in Ecrits logiques et philosophiques, trad. C. Imbert, Paris, Seuil.

Frege G. (1905). « On denoting ». Disponible sur : http://www.cscs.umich.edu/~crshalizi/Russell/denoting/

Hobbs J. R. (1978). "Resolving Pronoun References". *Lingua n°44.*

Hobbs, J. R. (1996). « Pronoun Resolution ». *Research report 76-1. City University of New York.*

Karttunen L. (1976). "Discourse referents". in Syntax and Semantics 7 : Notes from the Linguistics Underground, Academic Press, New York.

Kehler A. (1997). "Probabilistic Coreference in Information Extraction".

Kleiber G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Centre d'analyse syntaxique, Metz.

Kleiber G. (1997). « Sens, référence et existence : que faire de l'extra-linguistique ? » Langages n°127.

Kripke S. (1980). *Naming and Necessity*. Harvard University Press, Cambridge.

Kripke S. (1982). *La logique des noms propres*. trad. Jacob P. et Recanati F., Paris, Editions de Minuit, coll. "Propositions".

Lappin Sh., Leass H. J. (1994). "An Algorithm for Pronominal Anaphora Resolution". *Computational Linguistics n°20.*

Linsky L. (1967). *Le problème de la référence*. Paris, Seuil.

Milner J.-C. (1978). *De la syntaxe à l'interprétation*. Paris, Seuil.

Milner J.-C. (1982). *Ordres et raisons de langue*. Paris, Seuil.

Mitkov R. (1996). « Anaphora and Machine Translation. » *Machine Translation Review.*

Mira A. (1990). « Accessing noun-phrase antecedents ». *Routledge, London.*

Pichon E. (1937). « La linguistique en France : problèmes et méthodes ». *Journal de psychologie normale et pathologique.*

Russell (1905). « On Denoting ». *Disponible sur : http://www.cscs.umich.edu/~crshalizi/Russell/denoting/*

Saussure (1916/1972). *Cours de linguistique générale*. Payot.

Schneidecker C. (1997). *Nom propre et chaînes de référence*. Centre d'études linguistiques des textes et des discours de l'Université de Metz.

Searle J. (1972). *Les Actes de langage*. Paris, Hermann.

Siblot P. (1994). « Une Linguistique qui n'a plus peur du réel ». *Cahiers de praxématique, n°15.*

Siblot P. (1997). Présentation du numéro de Langages intitulé « Langue, praxis et production de sens ». Langages n°127.

Siddharthan A. (2003). “Resolving Pronouns Robustly : Plumbing the depths of shallowness”. *In Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), pages 7-14, Budapest.*

Sundheim B. M. (1995). « Overview of results of the MUC-6 evaluation ». in *Proceedings of the Sixth Message Understanding Conference, p. 13-31.*

### 3. Statistiques et mesures de liaisons

Daille B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.

## WEBOGRAPHIE

Site du Monde :

http://www.lemonde.fr/

Site de Libération :

http://www.liberation.fr/

Site du Figaro :

http://www.lefigaro.fr/

LexiMédia2007 :

http://erss.irit.fr/LexiMedia2007/

Blog de Jean Véronis :

http://aixtal.blogspot.com/

Presse 2007 :

http://www.up.univ-mrs.fr/veronis/Presse2007/

Observatoire-présidentielles :

http://www.observatoire-presidentielle.fr/

BonVote, moteur de recherche consacré aux sites et blogs politiques :

http://www.bonvote.com/

Répertoire des blogs politiques :
http://www.placedelademocratie.net/mediawiki/index.php?title=Accueil

Médiamétrie :
http://www.mediametrie.fr/

Moteur de recherche consacré aux blogs :
http://www.technorati.com/