Cogmaster – ENS / EHESS / Paris Descartes

Comment les cerveaux copient-collent-ils ? Dérive sémantique des citations dans la blogosphère

Rapport de Stage Long

Sébastien Lerique (sebastien.lerique@ens.fr)

Maître de stage : Camille Roth (roth@ehess.fr)

22 août 2012

Table des matières

1 Introduction

On assiste aujourd'hui à une convergence progressive entre des disciplines étudiant la cognition humaine, d'un côté, et des disciplines étudiant la culture et les *choses sociales*, de l'autre. Ces deux familles de disciplines ont en effet bien des frontières en commun ; celles-ci se rendent visibles par l'intermédiaire des thèmes pour lesquels il devient incontournable de s'intéresser à l'humain à la fois dans sa globalité et dans ses fonctionnements particuliers. L'anthropologie culturelle, qui cherche à comprendre l'émergence des similarités que représente la culture, étudie un de ces thèmes pour lesquels le dialogue avec les sciences cognitives est indispensable. Et ce dialogue est tout à fait prometteur : progressivement, on voit se développer des approches visant à étudier cognition individuelle, cognition sociale et culture de façon jointe.

L'une de ces approches, l'épidémiologie culturelle, met l'idée de *représentation* au centre de sa conceptualisation ; elle relie de cette façon la notion de *représentation mentale* développée par les sciences cognitives à celle de *représentation publique*, qui est le pendant de la première à l'extérieur du cerveau. Sperber fait la distinction entre les deux types de représentations dans son ouvrage fondateur du domaine : « Une représentation peut exister à l'intérieur même de l'utilisateur : il s'agit alors d'une *représentation mentale*. Un souvenir, une hypothèse, une intention sont des exemples de représentations mentales. L'utilisateur et le producteur d'une représentation mentale ne font qu'un. Une représentation peut aussi exister dans l'environnement de l'utilisateur comme par exemple le texte qui est sous vos yeux. Il s'agit alors d'une *représentation publique*. Une représentation publique est généralement un moyen de communication entre un producteur et un utilisateur distincts l'un de l'autre. » [1] (Sperber, 1996, p. 49.) Le point clé de cette notion est qu'une représentation n'est pas *diffusée*, elle est *interprétée* et, éventuellement, *reproduite* (non pas copiée). L'interprétation et la reproduction des représentations ont pour effet de faire se transformer et évoluer ces représentations, et l'épidémiologie culturelle cherche entre autres à décrire cette évolution grâce à la notion d'« attracteur culturel » (i.e. les domaines attracteurs du système dynamique ainsi formé, s'ils existent). Le domaine a ainsi apporté une alternative viable et nécessaire à d'autres théories de la culture peu satisfaisantes (telles que la mémétique ; voir Kuper, 2000, et Bloch, 2000, pour plus de détails sur les problèmes posés par cette théorie).

Bien qu'ayant connu des développements récents en modélisation (par exemple Claidière and Sperber, 2007, pour la notion d'attracteur culturel), la difficulté à effectuer des mesures quantitatives sur les notions clés du domaine a fait que les hypothèses à la base de l'édifice théorique développé n'ont pas encore pu être testées empiriquement.

Cependant, depuis plus d'une décennie le paysage des choses observables dans le domaine des représentations publiques est en train de changer radicalement. Nous vivons aujourd'hui non pas avec un manque de données concernant les représentations publiques, mais au contraire sous un véritable déluge de telles données. Quel est ce déluge ? Il est le produit d'une collecte qui est née avec l'utilisation commerciale et massive d'Internet. Cette collecte n'est pas un phénomène mineur, ni même un phénomène d'ampleur moyenne : les services sur Internet enregistrent aujourd'hui la quasi-totalité de ce

1. Les italiques sont de l'auteur.

qui est mesurable des interactions entre les personnes connectées ainsi que des interactions des personnes avec les machines. Vu l'importance que ce réseau a pris dans nos vies de tous les jours, on assiste bien à une explosion de ces données ; à tel point que certains observateurs d'Internet estiment que 90% des données stockées aujourd'hui par les services sur Internet ont étés générées il y a moins de deux ans.

Bien sûr, ces données ne sont pas des enregistrements des interactions *physiques* entre les personnes, au sens où on l'entend parfois en parlant de « vie réelle ». Mais elles sont bel et bien des observations sur le terrain nouveau que constituent l'action, la perception, et l'interaction connectées, au travers des machines et d'Internet. En un mot, elles sont en fait des observations en grandes quantités de *représentations publiques*. C'est-à-dire que depuis quelques années, on a accès à un outil permettant l'étude empirique des représentations dans de cadre de l'épidémiologie culturelle [2].

L'approche qu'on a esquissée ici n'est qu'une des façons d'aborder l'étude de la culture et de la cognition de façon empirique. En effet, les courants de recherche qui gravitent autour de ce thème sont nombreux et variés ; voici une rapide présentation des champs concernés et de la façon dont ils interagissent.

1.1 Littérature existante

L'idée d'étudier la culture en lien avec la cognition apparaît comme une extension naturelle pour plusieurs disciplines historiquement séparées qui, chacune de leur côté, s'en sont approchées en partant de leur propre espace de concepts et en étendant leurs outils. Les dialogues entre ces disciplines ne sont pas rares, mais les traditions et les épistémologies différentes ont rendu les unifications plus ardues. On peut distinguer au moins deux grands groupes de disciplines débattant autour de thématiques reliées à ce sujet [3].

Le premier groupe est issu des débats entre la biologie évolutionniste, la psychologie cognitive et les sciences sociales (et plus particulièrement l'anthropologie), desquels ont émergé les domaines de la psychologie évolutionniste et de l'anthropologie cognitive. Plusieurs théories ont été avancées et sont débattues en parallèle par ces disciplines. On se doit de mentionner ici en premier lieu le débat autour du programme mémétique initié par Dawkins (1976) et pour lequel l'ouvrage collectif de Aunger et al. (2000) est un bon exemple. Un deuxième terrain exploré par cet ensemble de disciplines est le développement de modèles évolutionnistes basés sur les normes (voir par exemple Ehrlich and Levin, 2005), avec comme travail fondamental celui de Boyd and Richerson (1985). Le troisième champ, particulièrement important aujourd'hui, est précisément celui de l'épidémiologie culturelle, initié par Sperber (1996) et défendu face aux autres courants dans des travaux comme celui de Atran (2003).

2. Bien sûr, une bonne partie des données collectées sur Internet appartient à des entreprises qui n'ont aucun intérêt à les partager ni à les exploiter autrement que financièrement. Mais les données publiquement accessibles forment tout de même une base considérable, qu'il est tout à fait possible d'exploiter comme telle.

3. On regroupe ici les disciplines par les thèmes qu'elles débattent, plutôt que par les similarités entre leurs façons de traiter ces thèmes.

Le deuxième groupe, plus diffus, provient des échanges entre le *data mining*, l'étude des systèmes complexes, et la sociologie quantitative (e.g. sociologie des réseaux). Un grand nombre de travaux apparaissent exploitant de grands corpus de données produits par des collectes sur Internet, allant de l'étude de l'impact des réseaux sociaux sur la diffusion des informations sur ces mêmes réseaux (Bakshy et al., 2012) à l'analyse du rôle de la presse en ligne dans la vie politique locale (Parasie and Cointet, 2012), en passant par le détail des interactions entre éditeurs sur Wikipedia (Jurgens and Lu, 2012) ou l'évolution du vocabulaire des langues au cours du temps (Lieberman et al., 2007). Une partie de ces travaux n'inscrivent leurs résultats que dans des cadres théoriques naïfs (autour de thèmes pour lesquels il existe pourtant des théories réfléchies développées par des champs des sciences sociales), et relèvent parfois plus des sciences de l'informatique que de l'étude de la culture (d'aucuns parlent d'ailleurs de *social computing*). Ce groupe de disciplines présente néanmoins la particularité de vouloir étudier la culture en exploitant les possibilités ouvertes par les larges corpus de données issus d'Internet.

État de l'art L'état de l'art en ce qui concerne l'épidémiologie culturelle peut être résumé ainsi : d'un côté, des études s'intégrant au deuxième groupe de disciplines permettent une bonne compréhension des phénomènes macroscopiques de *diffusion* des représentations publiques tels que le chemin emprunté par une information atomique donnée, les temps caractéristiques et les cycles de diffusion dans ces réseaux, la relation entre blogs et sites de médias, ou encore l'autorité et l'influence exercée par certains nœuds du réseau. On peut citer ici l'étude de la propagation et de la diffusion de l'information dans la blogosphère faite par Gruhl et al. (2004), l'étude du *news cycle* faite par Leskovec et al. (2009a) qui fait intervenir la plupart de ces aspects, ou encore l'analyse des influences réciproques entre un réseau social et le réseau sémantique formé par les contenus diffusés sur ce premier réseau, faite par Cointet and Roth (2009) ; on l'a déjà mentionné, ces études sont relativement indépendantes de l'anthropologie et de la cognition, ces thèmes étant explorés principalement par le premier groupe de disciplines. Celui-ci a, d'un autre côté, récemment fait avancer l'étude de la *transformation* des représentations publiques, qui en est encore à ses débuts ; la notion d'« attracteur culturel » a ainsi fait l'objet de modélisations permettant de rendre compte de l'évolution de certaines représentations et comportements (Claidière and Sperber, 2007). Dans les approches empiriques de la transformation des représentations, un des travaux les plus aboutis à ce jour est une étude des citations (au sens d'extraits cités, ou *quotation* en anglais) et de leurs transformations dans la blogosphère américaine, menée par Simmons et al. (2011). Ces derniers mettent en évidence plusieurs régularités dans les transformations de ces représentations publiques et proposent un modèle de diffusion-transformation des citations, relativement simpliste du point de vue cognitif.

On observe donc un fossé avec, d'un côté, des études empiriques macroscopiques bien avancées rendant compte de phénomènes relevant de la diffusion, et, de l'autre, des études encore balbutiantes et pour la plupart normatives de la transformation des représentations au niveau microscopique. On remarque de plus que le volet cognitif, central pour les phénomènes étudiés, est relativement peu développé dans les modèles et peu

exploré de façon empirique.

1.2 Problématique

On s'est proposé d'étudier de façon *empirique* la transformation des représentations publiques, c'est-à-dire d'explorer à l'aide de données obtenues in vivo l'effet du cerveau sur l'évolution des représentations ; une telle étude permettrait d'observer empiriquement l'effet d'un attracteur culturel tel que défini par Sperber. La question est donc la suivante : « peut-on observer un biais cognitif dans le traitement de certaines représentations culturelles par le cerveau, et si oui quel est-il ? ».

Il y a (au moins) trois travaux pertinents pour notre point de départ : le premier travail est celui de Simmons et al., mentionné plus haut, qui montre que même pour un type de représentation publique qui ne devrait pas changer, à savoir les citations, des transformations ont bien lieu et il est possible de les observer et de les mesurer. Cependant ces auteurs n'étudient les transformations observées qu'en fonction du type d'auteur qui reproduit les citations, et pas en fonction des citations elles-mêmes (en particulier les questions d'ordre cognitif concernant la raison pour laquelle ces transformations ont lieu, ou la raison pour laquelle elles ont lieu dans tel sens et pas dans tel autre, ne sont pas abordées). Les possibilités offertes par cette étude étant vastes, on a décidé de travailler sur les mêmes données que Simmons et al. pour traiter notre question.

Les deux autres travaux viennent de la psycholinguistique, et posent les fondations pour l'étude des propriétés des représentations (qui seront ici des citations). Le premier des deux est une étude de Griffiths et al. (2007), concernant une tâche où l'on présente aux sujets une lettre de l'alphabet en leur demandant de dire le premier mot commençant par cette lettre qui leur vient à l'esprit. Les auteurs montrent qu'une certaine propriété des mots (le PageRank calculé sur un réseau sémantique construit à partir de données d'associations de mots) est un bon facteur prédictif de quel mot sera rappelé parmi les mots commençant par la lettre demandée ; ils montrent ainsi un lien entre une caractéristique sémantique quantitative des mots et la facilité de rappel de ces mots dans cette tâche. La deuxième étude de psycholinguistique, de Chan and Vitevitch (2010), étudie une tâche où l'on demande aux sujets de nommer oralement une image. Les auteurs montrent qu'une autre propriété des mots, phonologique cette fois (le coefficient de regroupement des mots calculé sur le réseau phonologique formé par ces mots), a un effet sur la rapidité à nommer l'image présentée : c'est un autre cas de lien entre une caractéristique quantitative (phonologique) d'un mot et la facilité à produire ce mot.

Ces deux dernières études montrent donc que des propriétés des mots du vocabulaire, calculées sur un réseau formé par ces mêmes mots, sont des facteurs importants dans la façon dont on les traite cognitivement et dans la façon dont on les produit. C'est justement l'élément qui nous manquait : en se basant sur ces trois travaux, on va pouvoir utiliser des propriétés ainsi calculées sur les mots des citations pour mesurer quantitativement les transformations subies par ces citations lorsqu'elles sont reproduites. Pour rendre l'analyse possible, on se concentrera sur des cas de *substitution* d'un mot par un autre lorsqu'un auteur reproduit une citation : à chaque substitution observée, on assiste

à un échange entre deux mots pour lesquels on va pouvoir comparer les propriétés sémantiques.

On verra que par cette analyse de la reproduction des citations par les auteurs on observe un biais cognitif non trivial ; on observera ainsi un lien entre l'attractivité de certains mots (en termes d'épidémiologie culturelle) et les propriétés sémantiques de ces mêmes mots, calculées sur la base du réseau formé par ces mots.

Plan Le travail de recherche effectué étant essentiellement empirique, on le présente en deux grandes étapes : en premier lieu, on détaille le protocole d'expérimentation utilisé ainsi que les différentes tentatives et pistes explorées pendant son élaboration. Pour cela, on commencera par analyser la question ci-dessus plus à fond afin d'en distiller une question opérationnelle, puis on expliquera en détail la façon dont on a construit le protocole expérimental permettant de faire les mesures pour répondre à cette question [4]. La deuxième grande partie concerne les résultats : on montre tout d'abord la façon dont les mesures sont utilisées pour construire les observables qu'on cherche, ce qui permettra ensuite d'interpréter les résultats obtenus au regard de la question posée.

Une dernière partie concernera les perspectives que ce travail ouvre, et les façons dont il pourrait être poursuivi et étendu.

4. Le terme « protocole expérimental », utilisé ici, est bien le terme approprié pour ce dont il s'agit : on verra qu'on part d'un corpus de données pour lequel on construit une façon de faire les mesures qu'on considère pertinentes, et cette démarche correspond à la construction d'un protocole expérimental comme on peut le faire pour une expérience de psychologie cognitive ; c'est à ce moment qu'on définit ce qu'on va mesurer et comment cette mesure est faite.

2 Démarche et Méthode

Notons d'abord que l'analyse de la question posée, esquissée dans la problématisation ci-dessus, ne s'est pas faite indépendamment des données disponibles. En effet, une large partie de la réponse à la question dépend de la façon dont on définit la notion de représentation publique, notion pour laquelle les définitions possibles sont multiples et font l'objet de débats ouverts (et qui le resteront probablement encore un certain temps). Il a donc fallu définir cette notion pour qu'elle corresponde bien sûr à ce qu'on cherche, mais aussi à quelque chose dont il est possible de s'approcher par des mesures empiriques. Ceci explique le va-et-vient entre théorie et empirisme qu'on observe dans la suite de cette partie.

La première sous-partie détaille la façon dont la question a été approchée et affinée ; la deuxième sous-partie reprendra les problèmes ouverts en fin de première partie et exposera en détail les outils et la méthode utilisés pour répondre à la question.

2.1 Analyse de la question posée

2.1.1 Quelles représentations publiques ?

On a choisi de prendre une forme de représentation publique en apparence basique, mais qui entre bien dans la définition que Sperber fait des représentations publiques (Sperber, 1996, p. 49) : les citations. C'est une forme de représentation d'autant plus instructive à analyser que, si l'on s'en tient aux règles de l'écrit, les citations ne devraient pas changer : lorsqu'on cite quelqu'un il est attendu qu'on reproduise précisément les mots utilisés par l'auteur, éventuellement en n'en prenant qu'une partie pour mieux l'intégrer à son propos, mais certainement pas en en changeant les termes. On peut faire l'hypothèse que les auteurs qui citent *cherchent à le faire* en suivant cette règle. De ce fait, si l'on observe en effet une évolution des citations au-delà du « rognage » elle doit être l'effet d'un biais cognitif involontaire (de l'ordre de l'automatique). L'analyse de l'évolution des citations sur Internet est donc un terrain idéal pour observer un biais cognitif à l'origine de transformations des représentations publiques.

Revenons brièvement sur l'hypothèse faite ci-dessus car elle mérite un peu plus d'attention pour être valide. Elle est justifiée pour la simple raison qu'il est souvent facile d'aller vérifier les propos d'origine, et parce que cette transparence fait partie de la raison d'être et de la légitimité mêmes des citations. En effet, les auteurs qui en citent d'autres peuvent tout à fait intégrer une citation en en déformant l'intention d'origine, par exemple en la rognant à souhait ou en extrayant la citation de son contexte, ou encore en la juxtaposant à d'autres citations qui renforcent le propos de l'auteur citant, mais il est rarement à l'avantage de l'auteur de changer les mots utilisés. Ceci pour deux raisons : d'une part, parce que la légitimité et la force de l'information transmise ou de l'argumentaire déployé reposent en partie sur l'honnêteté intellectuelle qu'on peut prêter à l'auteur, d'autre part parce que l'exactitude des citations est justement le premier des indicateurs à utiliser lorsqu'on cherche à mesurer cette honnêteté intellectuelle.

Enfin, et pour terminer de justifier cette hypothèse, pour que ce garde-fou fonctionne il n'est pas nécessaire que les citations faites par un auteur soient vérifiées systématiquement, ni même qu'elles soient vérifiées souvent : il suffit que toute l'information nécessaire à la vérification soit publique, pour que le risque encouru par la déformation d'une citation soit bien réel. C'est bien souvent le cas sur Internet [5]. Il n'est pas non plus nécessaire que la vérification soit facile à faire pour que cette garantie tienne, mais il suffit que le risque de chute de légitimité soit assez grand pour que peu de monde s'y risque sérieusement [6].

2.1.2 Les données

La construction d'un corpus de données est une tâche longue et fastidieuse. Bien souvent, les informations récoltées ne peuvent pas être utilisées dans leur état brut, et plusieurs phases de filtrage et de traitement des informations sont nécessaires avant d'obtenir un corpus utilisable. On a donc choisi de s'appuyer sur des corpus déjà disponibles au lieu de construire le nôtre, ce qui est une des raisons pour lesquelles l'analyse de la question est si tributaire du choix des données. Les données disponibles étaient les suivantes :

- Le corpus utilisé par Leskovec et al. (2009a) et Simmons et al. (2011) (composé de citations extraites principalement de la blogosphère des États-Unis et d'articles de médias, allant d'août 2008 à avril 2009) ;
- Un corpus issu de Twitter (micro-blog) ;
- Un corpus issu de la blogosphère française ;
- Un corpus de dépêches AFP.

Données choisies Le jeu de données utilisé par Leskovec et al. (2009a) et Simmons et al. (2011) se prête bien à la notion de représentation publique qu'on a choisie ci-dessus. Construit et publié en 2009 par Leskovec et al., c'est un jeu de données reconnu et utilisé dans la littérature, largement pré-traité pour nos besoins comme on le verra ci-dessous. Il est d'une dimension bien suffisante pour atteindre des niveaux de significativité intéressants, tout en restant manipulable sur un serveur de calcul modeste ou même sur un bon ordinateur portable.

Origine Ce dataset a été constitué en scannant environ un million de sources en ligne pendant neuf mois (d'août 2008 à avril 2009), les sources allant du média de masse jus-

5. De la même manière qu'exiger que les comptes d'une entreprise soient publics garantit un certain niveau d'« honnêteté comptable » (sans pour autant que des vérifications soient faites systématiquement), le fait que les données nécessaires à la vérification d'une citation soient publiques est un bon garant du fait que les auteurs qui en citent d'autres cherchent à rester fidèles aux mots d'origine.

6. S'il n'est pas toujours facile de vérifier qu'une citation est exacte, le jour où quelqu'un montre qu'elle est inexacte il sera bien compliqué pour l'auteur citant de se défendre. Ce raisonnement s'applique moins à la déformation du contexte de la citation, cas dans lequel il est plus facile de se défendre en arguant de la divergence d'interprétation. La *citation inexacte*, en revanche, relève plus de la *donnée fausse* que de l'*interprétation divergente*.

qu'au blog personnel [7]. Les auteurs du dataset ont extrait de ces sources toutes les citations. Chacune de ces citations est un extrait, plus ou moins fidèle à l'origine, d'une phrase prononcée ou écrite par la personne citée. Les données contiennent donc des ensembles de citations qui sont toutes des sous-parties, potentiellement transformées, d'un énoncé parent.

Forme Un des travaux des auteurs a été de reconstituer ces ensembles de citations par une méthode de catégorisation (i.e. construction de *clusters*) [8], et c'est à ce stade de traitement qu'on obtient les données : elles sont constituées d'environs 70 000 ensembles de citations, à l'intérieur desquels toutes les citations proviennent d'un même énoncé parent (elles en sont des sous-parties transformées). Pour chaque citation à l'intérieur de chaque ensemble, on a une liste de triplets *(URL, date d'apparition*[9]*, nombre d'occurrences à cette URL)* décrivant les adresses et les dates auxquelles cette citation est apparue.

Inconvénients Ce dataset présente aussi quelques inconvénients avec lesquels il a fallu composer :

Premièrement, les données fournies sont toutes en lettres minuscules : il est donc impossible de faire de la reconnaissance d'entités nommées.

Deuxièmement, il présente un biais significatif dans la population qui l'a généré : les données ont été collectées sur la blogosphère et les sites de médias des États-Unis, et ont donc été créées par une population qu'on ne peut pas mesurer à partir des données mais qui n'est certainement pas représentative de l'ensemble de la société des États-Unis [10].

Enfin, si ce dataset est intéressant par sa taille, il comporte en revanche un bruit conséquent. En effet, le contrecoup de la constitution d'un corpus de données aussi massif est qu'il est extrait de sources non structurées (i.e. dont la catégorisation est lisible par un humain, pas par une machine), et cette extraction n'est jamais parfaite. On retrouve donc dans les données des éléments qu'on aurait voulu exclure, mais qu'il n'est pas possible

7. Tiré du site de la base de données : « MemeTracker builds maps of the daily news cycle by analyzing around 900,000 news stories and blog posts per day from 1 million online sources, ranging from mass media to personal blogs. » (Leskovec et al., 2009b.) L'article d'origine (2009a) n'offre pas plus de détails sur la sélection des sources.

8. Se référer à Leskovec et al. (2009a) pour plus de détails. Ce traitement des données comporte aussi quelques étapes de filtrage, décrites dans l'article.

9. On parlera de « date » alors qu'il s'agit d'une date et d'une heure, allant jusqu'à la seconde.

10. Ce point, fondamental pour un scientifique des sciences humaines et bien problématique dans le cas présent où il s'agit de représentations publiques, est souvent négligé dans les études de sciences cognitives ; en effet, on avance souvent l'idée que ce qui est cherché est un phénomène de complexité cognitive suffisamment faible pour être commun à tous les êtres humains. Mais cette idée est une hypothèse et non un argument, et est rarement testée empiriquement. Il faut noter cependant la difficulté à constituer un ensemble de sujets représentatif de la population qu'on cherche à explorer : tester effectivement cette hypothèse d'universalité des phénomènes qu'on observe nécessite de développer entièrement une nouvelle façon de recruter des sujets. Pour clore cet aparté, l'étude des différences de cognition entre différentes catégories de la population est un sujet à part entière qu'on n'a pas abordé pendant ce stage mais qu'il nous semblait important de ne pas laisser passer inaperçu. En guise de précaution, très grossière, on peut supposer que la population représentée dans les données fait majoritairement partie des catégories socioprofessionnelles éduquées.

de filtrer à la main vu la dimension du dataset. Comme on le verra plus bas (partie 2.2.3, page 26), le filtrage automatique des données est une tâche à part entière qui a fait l'objet de nombreux tâtonnements.

2.1.3 Affinement de la question

La question peut alors s'énoncer ainsi : étant donné un ensemble de citations provenant toutes d'un même énoncé parent, comment une citation se transforme-t-elle sous l'action du cerveau lorsqu'elle est reproduite par un auteur ? Cette question en amène une autre : que mesurer pour rendre compte de la transformation d'une citation lors de sa reproduction par un auteur ?

On a choisi de se concentrer sur les transformations des citations dans les cas où un seul des mots de la citation est changé. On observe donc des cas de *substitution* d'un mot par un autre au même endroit dans la citation, et l'on va chercher à décrire certaines propriétés de ces substitutions. À ce stade, il reste deux éléments clés à définir pour que l'analyse de la question puisse être implémentée : la façon dont on détecte les substitutions, et les mesures qu'on fait sur ces substitutions. On les mentionne ici pour montrer la pertinence de ces éléments, mais ils seront traités en détail dans la partie suivante (partie 2.2, page 15).

Détection des substitutions Il manque en effet certaines informations dans le corpus pour pouvoir lire ces substitutions de façon immédiate dans les données : les métadonnées pour chaque citation ne concernant que les URLs, les dates d'apparitions et les nombres d'occurrences à chaque URL, on n'a aucune information sur les liens entre les citations. Or, on a besoin de savoir sur quelles sources un auteur s'est appuyé pour reproduire une citation, c'est-à-dire qu'on a besoin de savoir quelle citation a été générée à partir de quelle autre. Cette information n'est pas présente dans les données et on va donc devoir l'inférer, à l'aide de plusieurs hypothèses. On a ainsi identifié six modèles possibles, qui ont tour à tour été implémentés dans la détection des substitutions. La partie suivante détaillera ce point (partie 2.2.1, page 17).

Mesures sur les substitutions Enfin, et c'est un point central, il reste à définir la façon dont on rend compte de l'évolution des citations, c'est-à-dire qu'il reste à définir les mesures qu'on effectue sur les substitutions détectées. Résoudre ce problème revient à définir les aspects ou traits caractéristiques des citations qu'on considère pertinents dans la transformation des citations [11]. Les aspects a priori les plus intéressants quant à l'évolution d'une citation étant ceux liés au sens véhiculé, on a choisi de mesurer des caractéristiques sémantiques de ces citations.

11. Alors qu'une analyse qualitative, au cas par cas, permettrait l'élaboration de caractéristiques spécifiques à chaque citation analysée, on est ici obligé de définir des caractéristiques d'une part suffisamment génériques pour être applicables à toutes les citations, d'autre part extractibles de façon automatisée, et ceci tout en les gardant pertinentes.

À notre connaissance, la seule façon d'analyser sémantiquement des phrases de façon automatique sans faire intervenir des techniques encore peu développées de sémantique formelle – telles que des inférences de modèles possibles du monde – est d'utiliser les caractéristiques des mots en tant que nœuds d'un réseau sémantique de concepts. On a donc restreint les mesures à des propriétés des deux mots échangés lors des substitutions, et on s'est attaché à calculer un certain nombre de caractéristiques sémantiques des mots concernés, sur la base du réseau formé par les concepts et les mots qui les représentent (on détaillera les caractéristiques utilisées dans la partie 2.2.2, page 21). Ceci nous permet de construire deux observables principales pour chaque caractéristique sémantique :

1. La variation de caractéristique sémantique lors du passage du mot remplacé au mot remplaçant dans une substitution. Ceci répond à la question « comment fait-on varier la caractéristique sémantique lorsqu'on substitue un mot ? ».
2. La susceptibilité que les mots ont à être remplacés, en fonction de leur caractéristique sémantique. Cette mesure répond à la question « quelles caractéristiques sémantiques a-t-on plus tendance à substituer ? » [12].

Enfin, on ajoute à ces mesures la catégorie grammaticale, dans les citations, des mots échangés lors des substitutions. On s'est restreint à quatre des catégories les plus fréquentes, en utilisant la catégorisation et la notation des *Part-of-Speech* (abrégé POS) [13] du Penn TreeBank Project [14] (Santorini, 1990) : verbes (POS dont le code commence par V), noms (POS dont le code commence par N), adjectifs (POS dont le code commence par J), et adverbes (POS dont le code commence par R). Cette mesure supplémentaire affine les observables définies ci-dessus, et permet d'observer un éventuel effet des catégories grammaticales sur les résultats.

2.2 Protocole expérimental pour répondre à la question

Maintenant que le problème est posé on peut passer à la construction proprement dite du dispositif qui va permettre de poser la question aux données. Cette partie, bien que plus technique, rend compte d'une très large partie du travail d'exploration et de recherche effectué. C'est l'équivalent de la description détaillée du protocole expérimental dans le cas d'une expérience de psychologie cognitive, mais sur un corpus de données mesurées *in vivo*.

12. Attention, cette mesure ne suffit pas en elle-même pour parler d'*influence* de la caractéristique sémantique d'un mot sur le fait qu'il soit substitué ou non, parce qu'on n'a pas fait d'analyse des causes impliquées dans le fait qu'un mot soit substitué ; en effet, d'autres facteurs pourraient être plus importants que ces caractéristiques sémantiques dans les causes d'une substitution. On ne répond donc *pas* à la question « quelle est l'influence de la caractéristique sémantique d'un mot sur sa propension à être substitué ? ».

13. Le terme *Part-of-Speech*, ou « partie du discours » en français, désigne les catégories grammaticales des mots (ou même des parties de mots) utilisées en linguistique.

14. Le Penn TreeBank Project définit une typologie très complète des POS, ainsi qu'une façon de les noter : à chaque POS correspond un code de deux ou trois lettres majuscules. Par exemple, tous les POS tenant du verbe commencent par la lettre V (e.g. VBN pour participe passé, VBG pour participe présent, etc.).

On commence par donner quelques définitions dont on aura besoin dans les sous-parties qui suivent ; on reprend ensuite les deux problèmes laissés incomplets dans la fin de la partie précédente : la détection des substitutions et les mesures faites sur ces substitutions ; on continue en expliquant les quelques « recettes » additionnelles mises en place pour améliorer la détection et le filtrage des substitutions ; enfin on termine en récapitulant l'ensemble des paramètres du protocole qui peuvent varier, paramètres pour lesquels on a testé chacune des différentes combinaisons.

Voici tout d'abord les quelques définitions dont on aura besoin dans la suite :

Définition 1. Distance de Hamming : On appelle *distance de Hamming entre deux listes* l_1 et l_2 de même longueur le nombre d'éléments différents entre l_1 et l_2. On la note $d_H(l_1, l_2)$.

Par exemple, la distance de Hamming entre les mots « chants » et « champs » pris comme listes de caractères est 2.

Définition 2. Sous-distance de Hamming : On appelle *sous-distance de Hamming entre deux listes* l_1 et l_2 (avec $|l_1| \geq |l_2|$) le minimum des distances de Hamming entre l_2 et toutes les sous-listes de l_1 de même longueur que l_2. On la note $\widetilde{d}_H(l_1, l_2)$. On a

$$\widetilde{d}_H(l_1, l_2) = \min_{\substack{l \text{ sous-liste de } l_1, \\ |l|=|l_2|}} d_H(l, l_2)$$

Par exemple, la sous-distance de Hamming entre les mots « branche » et « banc » pris comme des listes de caractères est 1 (en comparant « ranc » à « banc »).

Définition 3. Distance de Hamming-tokens : On appelle *distance de Hamming-tokens entre deux citations* s_1 et s_2 la distance de Hamming entre s_1 et s_2 considérés comme des listes de *tokens* (i.e. comme des listes de mots et de caractères de ponctuation). On la note $d_{H,tok}(s_1, s_2)$. Cette définition ne s'applique qu'à des citations. Si l'on note $\mathfrak{T}$ la fonction *tokenizer* qui à une liste de caractères associe la liste des tokens, on a $d_{H,tok}(s_1, s_2) = d_H(\mathfrak{T}(s_1), \mathfrak{T}(s_2))$.

Par exemple, la distance de Hamming-tokens entre les phrases « Elle a grandi dans les champs » et « Il a grandi dans les chants » est 2. Et la distance entre les phrases « Tu as faim. » et « Tu as faim ? » est 1 (la ponctuation change).

Définition 4. Sphère : On note $S_d(a, r)$ la *sphère* de centre a et de rayon r pour la distance d.

Par exemple, la sphère de centre « champs » et de rayon 2 pour la distance d_H contient les mots « chants » et « crampe » (en prenant les mots comme des listes de caractères). De la même manière, les phrases « Dessine-moi du mouton », « Dessine-moi un serpent », « Donne-moi un mouton » sont dans la sphère de centre « Dessine-moi un mouton » et de rayon 1 pour la distance $d_{H,tok}$.

Définition 5. Date-heure : On appelle *date-heure* un couple formé d'une date (du type `JJ-MM-AAAA`) et d'une heure précise à la seconde (du type `hh :mm :ss`). On note une date-heure `JJ-MM-AAAA hh :mm :ss`. La date et l'heure d'apparition d'une citation seront appelés ensemble date-heure. (Par exemple `22-12-2012 12:00:00`).

Définition 6. Cluster : On appelle *cluster* un ensemble de citations provenant toutes du même énoncé parent, associé à toutes les métadonnées sur ces citations (le corpus de données est formé d'un ensemble de clusters, avec pour chaque cluster les métadonnées sur les citations qu'il contient). On notera un cluster générique $\mathcal{C}$.

Définition 7. Notations pour les clusters : Pour un cluster donné $\mathcal{C}$, on définit :

- son *instant initial* t_{init} comme la première date-heure à laquelle une citation apparaît pour ce cluster,
- son *instant final* t_{fin} comme la dernière date-heure à laquelle une citation apparaît pour ce cluster,
- sa *durée de vie* $T \stackrel{\text{def}}{=} t_{fin} - t_{init}$.

Prenons par exemple un cluster ayant deux citations, q_a : « Dessine-moi un mouton », qui apparaît aux date-heures suivantes :

- `10-08-2012 12:45:00`,
- `10-08-2012 13:33:28`,
- `11-08-2012 06:25:54`,

et q_b : « Dessine-moi un patron », qui apparaît aux date-heures suivantes :

- `10-08-2012 21:12:32`,
- `11-08-2012 11:51:37`.

Alors :

- $t_{init} =$ `10-08-2012 12:45:00`,
- $t_{fin} =$ `11-08-2012 11:51:37`,
- $T =$ 23 heures, 6 minutes et 37 secondes.

Définition 8. Sac de citations, ou *timebag* : À partir d'un cluster $\mathcal{C}$, on construit un *sac de citations* en rognant le cluster entre deux instants t_1 et t_2 et en ne gardant comme informations que les chaînes de caractères des citations présentes entre t_1 et t_2 ainsi que la fréquence d'apparition de chaque citation dans cette fenêtre temporelle (en particulier, on ne garde plus les date-heures d'apparition des citations, ni leur ordre d'apparition). On note $\mathcal{B}_{\mathcal{C},[t_1,t_2]}$ un tel sac de citations. Enfin, on note $q_m\left(\mathcal{B}_{\mathcal{C},[t_1,t_2]}\right)$ la citation de $\mathcal{B}_{\mathcal{C},[t_1,t_2]}$ ayant la plus haute fréquence d'apparition à l'intérieur de ce sac de citations [15].

Par exemple en reprenant le cluster de l'exemple ci-dessus, on peut construire un sac de citations $\mathcal{B}$ défini entre les date-heures `10-08-2012 13:00:00` et `11-08-2012 07:00:00`. Il contient alors les informations suivantes :

- q_a : « Dessine-moi un mouton », apparaît 2 fois,
- q_b : « Dessine-moi un patron », apparaît 1 fois,

et rien de plus. On a $q_m(\mathcal{B}) = q_a$.

2.2.1 Détection des substitutions et modèles sous-jacents

L'étape de détection des substitutions est cruciale : comme le corpus de données ne contient pas d'informations sur le lien entre les citations (les informations présentes

15. C'est-à-dire que q_m est la fonction qui à un sac de citations donné associe la citation la plus fréquente dans ce sac.

sont exactement du type de celles données dans l'exemple de cluster de la définition 7, page 17), on ne connaît pas les substitutions qui ont réellement eu lieu lorsque les citations étaient reproduites par les auteurs. Il faut donc inférer ces substitutions sur la base des informations qu'on a et en ajoutant des hypothèses sur la façon dont les auteurs sélectionnent leurs sources lorsqu'ils reproduisent une citation. Si l'inférence est correcte, on aura en main les substitutions qui ont réellement eu lieu (et nos résultats correspondront à la réalité) ; dans le cas contraire, on aura en main des substitutions qui n'ont jamais eu lieu (et nos résultats seront basés sur des mesures erronées). Voici la procédure suivie :

- On commence par définir un modèle décrivant la façon dont les auteurs sélectionnent des sources lorsqu'ils créent une nouvelle citation.
- Alors, étant donné deux citations d'un même cluster à distance de Hamming-tokens 1 (i.e. ayant exactement un token de différence), le modèle permet de déterminer si l'une des deux citations a été générée sur la base de l'autre (en supposant ce modèle vérifié). Ce processus prend en compte la fréquence de chaque citation dans le cluster, ses date-heure d'apparition, et d'autres critères précisés dans les modèles ci-dessous. En pratique, l'implémentation du modèle permet d'extraire des données toutes les substitutions vérifiant ce modèle.
- Chaque modèle rend compte d'une des façons dont les sources peuvent être sélectionnées par un auteur écrivant une nouvelle citation, et il est probable que la réalité soit un mélange de chacun de ces modèles. Mais n'ayant pas de moyen de les discriminer, on a fait les mesures en utilisant tour-à-tour chacun des modèles définis, pour ensuite comparer les valeurs des observables obtenues à chaque fois et avoir ainsi une idée des erreurs relatives entre modèles.

On a défini six modèles pour la détection des substitutions, chacun rendant compte d'un processus possible dans la génération d'une nouvelle citation. On en décrit deux ici – le premier est le modèle de base duquel dérivent presque tous les autres, et le second est le modèle principal utilisé dans les résultats –, les autres étant détaillés dans l'annexe A.1 (page 43). Il faut noter cependant que dans aucun de ces modèles on n'envisage la possibilité d'un mélange de plusieurs sources dans la production d'une nouvelle citation : il s'agit d'un processus de sélection qui, étant donné le paysage de citations présentes à un instant t, en sélectionne une et une seule pour l'utiliser comme source dans la création d'une nouvelle citation avec une substitution.

Modèle 1. ***Sliding timebags*** **:** On considère que les citations les plus visibles à un instant donné sont celles qui ont eu des apparitions le plus récemment, dans une fenêtre de temps (d'une durée fixée) précédant l'instant présent. De plus, on considère que ces citations sont celles qui influent sur la génération de nouvelles citations. Formellement, le modèle est le suivant : lorsqu'un auteur crée une nouvelle citation à l'instant t, il le fait à partir de la citation la plus fréquente dans le sac de citations défini sur la période de temps $[t - \Delta t, t]$ où Δt dépend de la durée de vie du cluster [16]. Il peut modifier cette citation ou non, et s'il la modifie il peut le faire de plusieurs façons, par exemple par une substitution (mais il n'est

16. On utilise ici la durée de vie du cluster comme indicateur du temps caractéristique d'évolution de ce cluster

pas restreint à ce sous-ensemble de transformations). En pratique, $\Delta t \stackrel{\text{def}}{=} \frac{T}{n}$ avec n entier.

On peut alors détecter les substitutions de la façon suivante : lorsqu'une citation q apparaît à l'instant t, on la considère comme étant une substitution d'une autre citation si et seulement si la citation la plus fréquente dans le sac de citations défini sur $[t - \Delta t, t]$ est à distance de Hamming-tokens 1 de q, autrement dit ssi $q_m \left(\mathcal{B}_{\mathcal{C},[t-\Delta t,t]}\right) \in S_{d_{H,tok}}(q, 1)$.

Ce modèle souffre du problème suivant : supposons qu'un cluster $\mathcal{C}$ ne soit composé que de deux citations q_a et q_b à distance de Hamming-tokens 1 l'une de l'autre. La citation q_a apparaît aux instants $t_{a,k}, k \geq 1$, et la citation q_b apparaît aux instants $t_{b,l}, l \geq 1$, de telle sorte que q_a et q_b aient des fréquences du même ordre de grandeur mais que q_a soit quand même la plus fréquente dans n'importe quel sac de citations. Alors chaque apparition de q_b après le premier sac de citations sera compté comme une nouvelle instance de substitution $q_a \to q_b$, et non comme une répétition de q_b : le modèle considère en effet que q_b est trop peu fréquente pour être la source d'une nouvelle citation. Or il est légitime de penser qu'une bonne partie des instances de q_b ne sont pas des substitutions à partir de q_a, mais bien des copies des instances précédentes de q_b. La figure 1 illustre ce point.

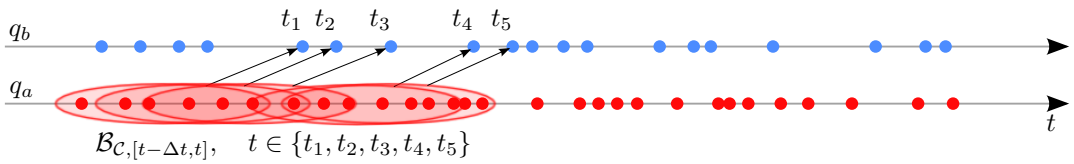


FIGURE 1 – Situation problématique pour le modèle *sliding timebags* (modèle 1)

En revanche, si q_a est plus fréquente pendant un premier tiers du cluster, puis q_b devient plus fréquente sur le deuxième tiers, les apparitions de q_b dans les deux premiers tiers seront comptées comme substitutions $q_a \to q_b$, mais les apparitions dans le troisième tiers ne seront pas comptées comme substitutions (puisque q_b sera devenue la citation la plus fréquente). Les apparitions de q_a dans le dernier tiers, par contre, seront comptées comme substitutions $q_b \to q_a$.

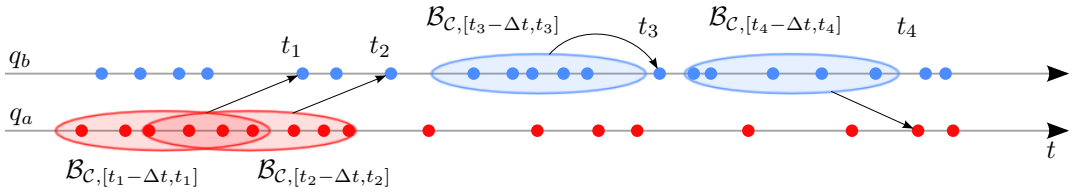


FIGURE 2 – Évolution des citations avec le modèle *sliding timebags* (modèle 1)

La figure 3 représente un cluster issu des données qui ressemble à cette situation :

Modèle 2. ***Fixed timebags*** **:** On reprend le modèle *sliding timebags* (modèle 1), avec deux approximations supplémentaires. Premièrement, on ne considère plus exactement

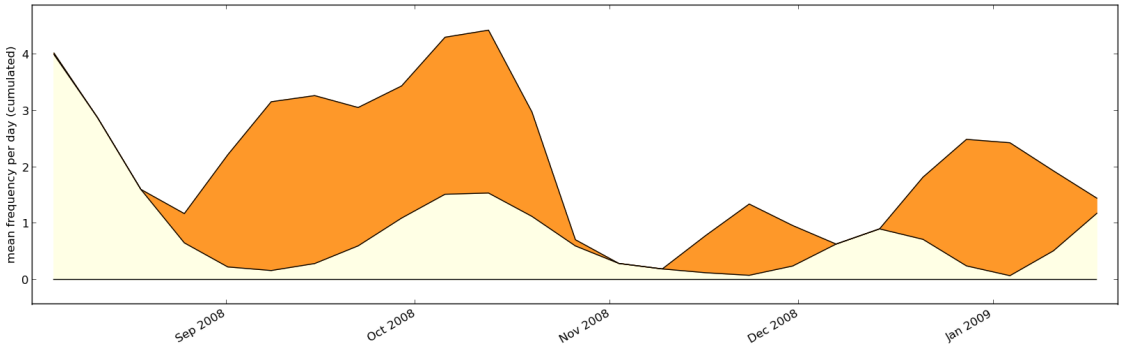


FIGURE 3 – Fréquence moyenne par jour des deux citations du cluster #2665. Chaque couleur correspond à une citation. Les fréquences sont cumulées : au pic de mi-octobre (le 13 octobre précisément), la citation jaune est autour de 1 et la citation orange autour de 3. Les courbes sont obtenues en lissant l'histogramme des fréquences par jour.

les durées $[t - \Delta t, t]$, mais on découpe chaque cluster en n sacs de citations définis sur $[t_{init} + i\Delta t, t_{init} + (i+1)\Delta t]$ (avec $\Delta t = \frac{T}{n}$ et $i \in [\![0, n-1]\!]$), qu'on note $\mathcal{B}_{\mathcal{C},i}$. Il y a donc n sacs de citations dont les positions sont fixées, et non plus glissantes. On considère alors qu'un auteur qui crée une nouvelle citation le fait à partir de la citation la plus fréquente dans le sac de citations qui le précède immédiatement dans le temps. Bien souvent, ce sac de citations ne contiendra pas toutes les citations qui sont apparues immédiatement avant l'instant présent : en effet, comme on prend le sac le plus récent qui termine avant l'instant présent, toutes les citations apparues entre la fin du sac et l'instant présent ne seront pas incluses. Cette approximation, qui paraît un peu alambiquée au premier abord, permet surtout d'introduire la deuxième approximation qui tempère le problème des deux premiers modèles : au lieu de regarder chacune des apparitions d'une citation et de toutes les considérer comme étant une nouvelle substitution (lorsque la distance de Hamming-tokens entre les deux citations est 1), on ne compte qu'une seule apparition par sac de citations. La figure 4 illustre ce point.

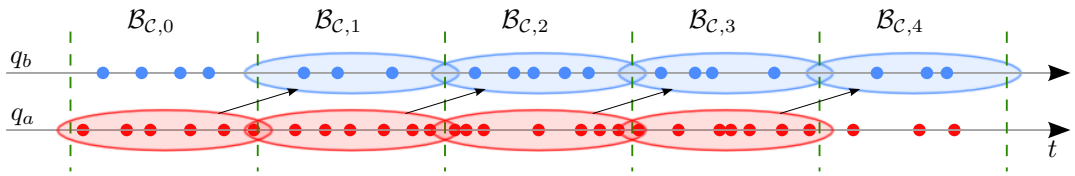


FIGURE 4 – Effet des approximations du modèle *fixed timebags* (modèle 2) sur le problème du modèle *sliding timebags* (modèle 1) ($n = 5$ ici)

Alors on peut détecter les substitutions de la façon suivante : lorsqu'une citation q apparaît à l'instant t, on la considère comme étant une substitution d'une autre citation si et seulement si la citation la plus fréquente dans le sac de citations la précédent directe-

ment est à distance de Hamming-tokens 1 de q, autrement dit si $q_m(\mathcal{B}_{\mathcal{C},i}) \in S_{d_{H,tok}}(q,1)$ où $i = \lfloor \frac{t-t_{init}}{T} \rfloor$. Si la citation q apparaît plusieurs fois dans le sac de citations dans lequel elle se trouve, on ne compte cette substitution qu'une seule fois.

Le modèle *fixed timebags* (modèle 2) a été implémenté avec $n \in [\![2,5]\!]$ (de même pour les modèles détaillés en annexe A.1, page 43), et tous les modèles ont aussi été implémentés en utilisant la sous-distance de Hamming-tokens au lieu de la distance classique : celle-ci permet de prendre en compte les transformations où un auteur sélectionne une citation et en copie une sous-partie, partie sur laquelle il opère une substitution (au lieu de ne prendre en compte que les copies avec substitution).

2.2.2 Mesures sur les substitutions

Revenons maintenant sur les caractéristiques sémantiques qu'on calcule sur les mots, basées sur le réseau sémantique formé par les concepts. On a utilisé la base de données WordNet (WordNet, 2010) [17], qui répertorie plus de 117 000 concepts et environ 147 000 mots rattachés à ces concepts, comme base pour le réseau sémantique.

Afin de bien clarifier les caractéristiques qu'on calcule et les graphes qu'on utilise, il faut avoir la structure de WordNet en tête : les deux éléments principaux dans cette base de données sont le *mot désambiguïsé*, ou « lemme », et le *concept*, ou « synset ». En effet, comme bon nombre de mots dans les langues naturelles sont polysémiques, WordNet utilise les lemmes comme brique de base en les désignant par l'orthographe du mot associée à une des significations de ce mot. Par exemple en français, le mot « glace » a au moins deux significations : l'une dans le champ sémantique des vitres et fenêtres, l'autre dans le champ sémantique de l'eau. Un WordNet français répertorierait donc au moins deux lemmes pour le mot « glace » : (« glace », signification « vitre »), et (« glace », signification « eau »). Un lemme est donc l'orthographe d'un mot indicée par la signification revêtue par ce mot (ce à quoi est ajouté une courte description de ce lemme, ainsi que quelques autres propriétés générales).

Pour WordNet, les significations sont en fait des concepts, et sont appelés « synsets » : en effet, tous les lemmes rattachés à la même signification sont des synonymes, et leur ensemble forme un synset (d'où le nom, qui vient de *synonym set* en anglais). Par exemple, le mot « glace » a une signification qui se rapporte au concept de surface réfléchissante, et ce nouveau lemme (« glace », signification « surface réfléchissante ») est synonyme d'un autre lemme : (« miroir », signification « surface réfléchissante »). Ces deux lemmes appartiennent au même synset, « surface réfléchissante ».

Enfin, il faut noter que le choix du synset pour un mot détermine la catégorie grammaticale de ce mot. En reprenant l'exemple de « glace », on peut distinguer de façon bien plus détaillée les sens suivants qui déterminent chacun une catégorie grammaticale (la liste n'est pas exhaustive) :

- Le sens « eau à l'état solide », pour lequel « glace » est un nom ;

17. WordNet est accessible dans le langage Python grâce au *Natural Language Processing Toolkit*, ou NLTK en abrégé (NLTK).

- La forme conjuguée du verbe « glacer » au sens propre dans le champ lexical de l'eau (e.g. « il glace de l'eau »), signification pour laquelle « glace » est un verbe conjugué ;
- La forme conjuguée du verbe « glacer » au sens figuré (e.g. au sens de « son regard me glace »), signification pour laquelle « glace » est encore un verbe conjugué.

Graphe Grâce à cette base de données on peut construire un graphe de synonymes de la façon suivante :

1. On fait l'hypothèse que le niveau de détail pertinent sur les mots lorsqu'une substitution s'opère est le *mot non-désambiguïsé*, et non le mot déjà désambiguïsé. En effet, un mot dans une citation peut avoir plusieurs significations relativement proches mais distinguées dans WordNet (participant ainsi à la possibilité de plusieurs interprétations de la citation), et on suppose que les substitutions s'opèrent sur ces ensembles de significations proches, i.e. au niveau du mot lui-même et pas au niveau du mot désambiguïsé [18].
2. On convertit les mots avec des majuscules à leur version tout en minuscules, en convertissant tous les lemmes rattachés à la version avec des majuscules en des lemmes n'ayant que des minuscules. Par exemple en français, on convertirait « Anglais » en « anglais », transformant ainsi le lemme (« Anglais », signification « personne anglaise ») en (« anglais », signification « personne anglaise ») [19]. On verra à l'étape suivante que ceci implique que le mot « anglais » dans le graphe sera connecté à la fois à tous les mots qui ont un lemme synonyme de (« Anglais », signification « personne anglaise ») et à tous les mots qui ont un lemme synonyme de (« anglais », signification « adjectif ») (ainsi qu'à tous les autres mots qui ont des lemmes synonymes des autres significations de « anglais »).
3. On construit alors le graphe non-dirigé représentant la relation $\sim$:

$$w_1 \sim w_2 \iff w_1 \text{ et } w_2 \text{ sont les orthographes de deux lemmes synonymes}$$

4. On élimine les nœuds isolés ; ces nœuds correspondent aux mots dont aucun des lemmes n'a de synonymes.
5. On remarque enfin que deux mots peuvent être synonymes pour plusieurs significations différentes : par exemple, pour le mot « yell » en anglais, WordNet distingue les significations « a loud utterance ; often in protest or opposition », pour laquelle « yell » est un nom (on notera cette signification A), et « utter a sudden loud cry »,

18. C'est en fait une hypothèse qui mérite toute une série d'expériences pour être vérifiée. Cependant, on peut séparer les deux aspects : une substitution peut avoir une composante *mot* et une composante *concept* (ou mot désambiguïsé), et on peut faire des mesures sur les deux aspects. On choisit ici de se restreindre à la composante *mot*, en partie à cause du fait qu'il est très difficile de désambiguïser les mots dans les données de façon automatique.

19. C'est bien (« anglais », signification « personne anglaise ») ici, et non (« anglais », signification « adjectif »).

pour laquelle « yell » est un verbe (on notera cette signification B). Le mot « shout » a lui aussi les deux significations A et B, et on voit alors que « yell » et « shout » ont deux significations en commun. Le mot « cry » est encore plus lié à « yell » : en plus d'avoir les significations A et B en commun, ces deux mots ont aussi la signification « a loud utterance of emotion (especially when inarticulate) » en commun. Ils sont donc liés par trois significations.

On décide donc de pondérer chaque lien du graphe par le nombre de significations pour lesquelles les deux extrémités du lien sont synonymes (l'exemple des mots « yell », « cry », et « shout » est illustré dans la figure 5b ci-dessous).

On obtient ainsi un graphe pondéré non-dirigé, dont on peut voir deux extraits sur la figure 5 (page 24).

Caractéristiques sémantiques On a calculé deux caractéristiques sémantiques sur la base de ce graphe. Les voici succinctement expliquées [20] :

- **PageRank :** le PageRank est une mesure de la centralité des nœuds dans un graphe, et dans le cas présent c'est un bon indicateur de la polysémie d'un mot [21]. De ce fait cette mesure couple le nombre de significations d'un mot avec le nombre de synonymes rattachés à chacune des significations du mot : on peut visualiser ce couplage sur la figure 6. Enfin, on rappelle que l'utilisation du PageRank s'appuie sur Griffiths et al. (2007), dont on a décrit l'étude dans l'introduction (partie 1.2, page 8).
- **Coefficient de regroupement :** le coefficient de regroupement [22] d'un mot w mesure à quel point ce mot fait partie d'une clique. Il est défini par le ratio du nombre de liens *existants* entre les voisins de w et le nombre de liens *possibles* entre ces voisins (et la définition est adaptée pour le cas des graphes pondérés, comme ici). On peut montrer qu'un mot aura un haut coefficient de regroupement si les significations auxquelles il appartient sont elles-mêmes liées sémantiquement [23]. L'utilisation du coefficient de regroupement s'appuie cette fois-ci sur Chan and Vitevitch (2010), dont on a aussi décrit l'étude dans l'introduction (partie 1.2, page 8).

Enfin, le tableau 1 (page 25) montre l'exemple des 10 lemmes à plus haut scores pour le PageRank (on a omis le coefficient de regroupement car les 10 premiers lemmes on tous un coefficient égal à 1.0, ce qui ne permet pas de les classer).

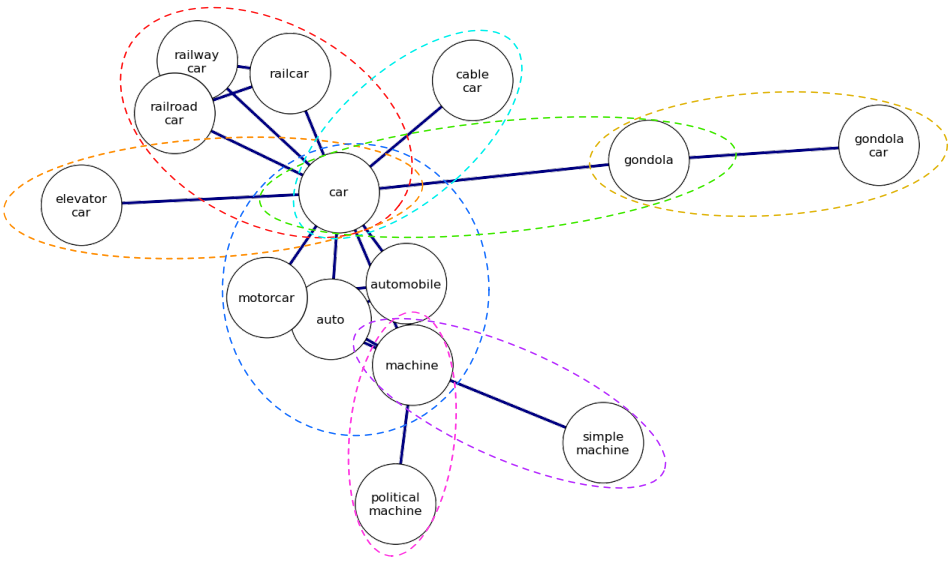
Catégories grammaticales Pour affiner l'étude au niveau des catégories grammaticales des mots, il a fallu inclure un module d'étiquetage grammatical des citations. Pour cela on a utilisé TreeTagger (Schmid, 1994, 1995), un étiqueteur probabiliste indépendant de la langue et atteignant une précision de plus de 96% sur la base de données

20. L'annexe A.2 (page 45) détaille plus précisément comment ces caractéristiques sont calculées, et l'implémentation qu'on en a fait.

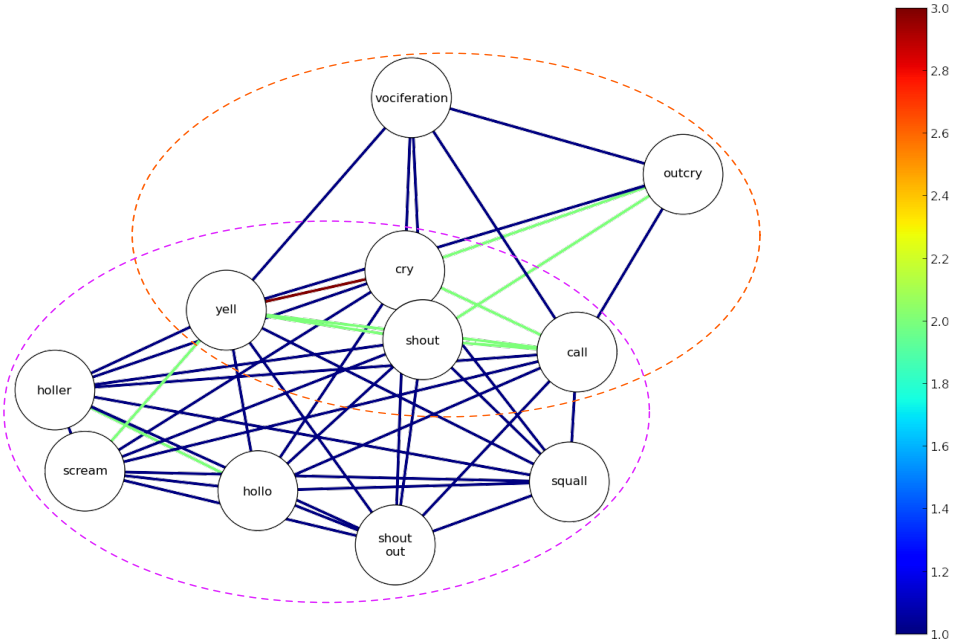
21. Voir l'annexe A.2 sur les caractéristiques sémantiques pour une explication de pourquoi le PageRank est un bon indicateur de la polysémie.

22. Appelé *clustering coefficient* en anglais.

23. Voir l'annexe A.2 pour une explication de cette interprétation, qui est non triviale.



(a) Voisinage à distance 3 de « car »



(b) Voisinage à distance 1 de « yell ». Le poids des liens est représenté en couleur : par exemple le lien « yell »–« shout » a un poids 2, et le lien « yell »–« cry » a un poids 3. Cet extrait du graphe total n'inclut pas tous les synsets à l'origine du poids des liens ; par exemple, le troisième synset liant « yell » et « cry » n'est pas représenté ici.

FIGURE 5 – Extraits du graphe de synonymes de WordNet. Chaque synset dans WordNet génère dans le graphe un ensemble de nœuds totalement connectés (une *clique*) ; on a entouré ces ensembles en pointillés.

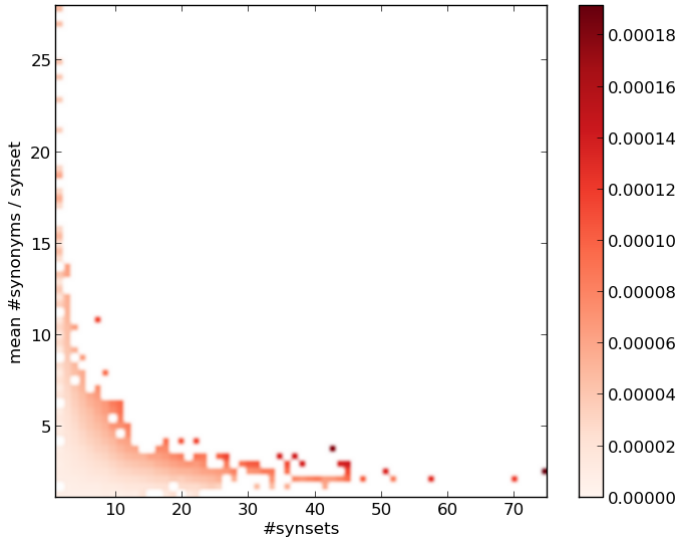


FIGURE 6 – PageRank en fonction du nombre de significations des mots (en abscisse) et du nombre moyen de synonymes pour chaque signification des mots (en ordonnée) ; la couleur en un point indique le PageRank moyen des mots placés autour du point.

#	Lemme	PageRank $\times 10^{-4}$
1	break	1.916
2	pass	1.800
3	hold	1.404
4	get	1.387
5	check	1.354
6	take	1.354
7	go	1.354
8	make	1.321
9	run	1.272
10	cut	1.173

TABLE 1 – 10 premiers lemmes pour le PageRank, normalisé pour que la somme de tous les rangs fasse 1. On note qu'une large majorité des mots ont des significations de verbes, mais pas seulement (e.g. « break » a de nombreuses significations en tant que nom, et encore plus en tant que verbe).

Penn Treebank, entraîné sur cette même base de données [24]. Mais une question se pose quant aux caractéristiques sémantiques restreintes à une catégorie grammaticale donnée : étant donné que, dans cette condition, on veut étudier l'effet du cerveau pour chaque catégorie, il faut un modèle des caractéristiques sémantiques des mots restreintes à cette catégorie. Autrement dit, les caractéristiques devraient-elles être calculées sur la base du graphe complet en ne gardant ensuite que les mots de la catégorie choisie, ou devraient-elles être calculées sur le sous-graphe, bien plus petit, des mots de la catégorie choisie ? On a implémenté les deux possibilités, et en pratique les résultats ne changeaient pas qualitativement. Cependant on a choisi de présenter les résultats obtenus pour le calcul sur le graphe complet, car la même hypothèse que celle pour l'utilisation des mots non-désambiguïsés s'applique (partie 2.2.2 et note 18, page 22) : on fait l'hypothèse que le niveau de détail pertinent lors du choix du nouveau mot dans une substitution est le mot non-désambiguïsé et hors catégorie grammaticale ; en d'autres termes, si le mot d'arrivée vérifie nécessairement la contrainte de s'insérer correctement dans la phrase de destination (et donc d'être probablement de la même catégorie grammaticale que le mot d'origine), le processus qui génère ce mot n'est pas nécessairement restreint à cette catégorie grammaticale. On pense donc que les caractéristiques sémantiques pertinentes sont celles calculées sur le graphe complet.

2.2.3 Recettes de filtrage

Comme on l'a expliqué précédemment, la contrepartie de la constitution d'un corpus de données aussi large est le bruit qu'on y retrouve ; celui-ci implique un travail fouillé de filtrage pour extraire les informations qu'on cherche et seulement celles qu'on cherche. Pour cela on a testé plusieurs façons de filtrer les clusters eux-mêmes ainsi que les substitutions détectées à l'aide des modèles ci-dessus. Le terme « recette » peut faire sourire, mais c'est bien ce dont il s'agit ici : les méthodes de filtrage utilisées ne découlent pas toujours d'une théorie en amont, et la plupart sont des techniques ad-hoc qui ont été adaptées par un processus d'essai-erreur.

Filtrage des clusters On a repris le filtrage utilisé par Simmons et al. dans leur analyse des mêmes données. On a les deux méthodes suivantes pour chaque cluster :

- ***Framing*** **:** on estime le pic d'activité du cluster en extrayant la fenêtre de 24 heures pendant laquelle la fréquence d'occurrence de citations est maximale (cette estimation est faite à la demi-heure près), et on rogne le cluster temporellement en le faisant commencer deux jours avant le début de ces 24 heures et terminer deux jours après la fin de ces mêmes 24 heures. On obtient donc un cluster d'une durée de cinq jours. Cette procédure réduit la taille des clusters, mais n'en élimine aucun. Ceci permet a priori d'éviter de garder des citations rattachées au cluster par exemple à cause de leur faible nombre de mots sans qu'elles aient un réel lien avec les citations principales du cluster ; ces citations génériques formeraient un « bruit de fond », toujours

24. TreeTagger est encapsulé en Python grâce à Pointal.

présent (parce que rattaché à aucun évènement en particulier), qui pourraient dégrader les résultats.

- ***Token-filtering* :** on supprime les citations du cluster ayant moins de m tokens, avec en pratique $m = 5$. On ajoute à cela un module de détection de la langue des citations, et on supprime les citations qui ne sont pas en anglais. Si le cluster résultant est vide (i.e. s'il ne contient plus aucune citation), on le supprime. Cette procédure fait passer la base de données de 71 568 à 55 483 clusters : on élimine d'une part les quelques citations dans des langues autres que l'anglais (pour lesquelles on n'a ni outils linguistiques ni caractéristiques sémantiques), d'autre part les clusters formés de citations très courtes et très variables du type « I love you » (décliné en « I love him/her » et autres variations qui sont probablement indépendantes et ne proviennent pas les unes des autres).

On a testé les quatre combinaisons possibles de ces deux méthodes (aucun filtrage, framing seul, token-filtering seul, et framing et token-filtering ensemble) pour enfin observer que les résultats ne changeaient pas qualitativement mais seulement en clarté et, parfois, en significativité. Dans la suite les résultats montrés sont ceux issus du token-filtering seul.

Filtrage des substitutions Les substitutions elles-mêmes ont fait l'objet d'un long travail de tâtonnement autour du filtrage. En effet, un certain nombre de substitutions détectées étaient des faux positifs, par exemple parce qu'elles correspondaient à deux sous-parties de citations différentes, ou parce qu'elles consistaient en des changements mineurs qui ne devraient pas être pris en compte (e.g. changement entre pluriel et singulier, ou entre orthographes anglaise et américaine). On a donc implémenté toute une série de méthodes pour améliorer le filtrage des substitutions, qui sont détaillées dans l'annexe A.3 (page 48). Le résultat, vérifié manuellement sur une partie des substitutions, a été une nette amélioration de la détection de celles-ci.

2.2.4 Paramètres du protocole

L'analyse des données de cette façon produit des résultats pour toutes les combinaisons des paramètres suivants :

- Le type de filtrage des clusters : aucun filtrage, *framing*, *token-filtering*, ou les deux ;
- La catégorie grammaticale désirée : verbe, nom, adjectif, adverbe, ou toutes (auquel cas la règle de filtrage requérant la même catégorie grammaticale pour les deux mots de la substitution s'applique) ;
- Le modèle de détection des substitutions : les six modèles implémentés sont possibles ici, avec leurs variantes prenant en compte les sous-chaînes de caractères ;
- Le coefficient utilisé pour la taille des sacs de citations dans les modèles qui utilisent ce paramètre (en particulier le modèle 2 décrit page 19) : $n \in [\![2, 5]\!]$;
- Le type de caractéristique sémantique utilisée : PageRank ou coefficient de regroupement.

Le calcul pour tous ces jeux de paramètres a permis d'observer les erreurs relatives entre les différents paramètres.

3 Résultats

Dans la suite on notera de façon générale $\mathfrak{C}(w)$ une caractéristique sémantique (quelconque) calculée sur le mot w. On désignera le mot disparaissant lors d'une substitution indifféremment par « mot de départ », « mot 1 », ou « mot substitué », et le nouveau mot apparaissant lors d'une substitution par « mot d'arrivée », « mot 2 », ou « mot nouveau ».

La première sous-partie, plutôt formelle, explique la façon dont on construit les observables décrites dans la partie 2.1.3 (page 14) ; la deuxième sous-partie expose les résultats proprement dits, et en discute les interprétations.

3.1 Construction des observables

Les données produites par le protocole sont, pour chaque jeu de paramètres, la liste des couples $(\mathfrak{C}(w_1), \mathfrak{C}(w_2))$ représentant les caractéristiques sémantiques du mot substitué (w_1) et du nouveau mot (w_2) pour toutes les substitutions détectées, et une liste de métadonnées pour chaque tel couple représentant une substitution (ces métadonnées incluent principalement des données sur l'origine de chaque substitution, comme le cluster et les citations d'origine).

Les deux observables qu'on veut construire sont la variation moyenne de caractéristique sémantique lors d'une substitution, et la susceptibilité pour un mot à être substitué, en fonction de ses caractéristiques sémantiques. On commence par modéliser les observations qu'on effectue : on pose ainsi un cadre clair qui permettra ensuite de définir ces observables, d'une part, et de calculer un intervalle de confiance ainsi qu'un niveau de significativité pour les variations de caractéristiques sémantiques, d'autre part.

3.1.1 Modélisation des observations

Modèle continu Un auteur peut, à chaque instant, créer une nouvelle citation sur la base de sources externes, ou de sources déjà présentes sur Internet. On ne se préoccupe pas du cas où les sources sont externes, puisque cela ne donne pas lieu à une substitution observable.

On note $\mathcal{P}$ (pour *pool*) la famille de tous les clusters, d'où les auteurs tirent leurs sources. Cette famille de clusters évolue avec l'évolution des clusters donc avec le temps, et on note $\mathcal{P}(t)$ son état à l'instant t. On décrit l'état $\mathcal{C}(t)$ d'un cluster $\mathcal{C}$ à l'instant t par l'ensemble des triplets *(chaîne de caractères, date-heure d'apparition, nombre d'occurrences lors de cette apparition)* représentant les apparitions jusqu'à cet instant des citations rattachées à ce cluster (un cluster peut être vide si aucune citation n'est apparue jusqu'à l'instant t). On a donc $\mathcal{P}(t) = (\mathcal{C}_i(t))_{i\in I}$, où I est l'ensemble des indices des clusters existants dans le modèle, dont on note N le cardinal. On modélise alors la création d'une nouvelle citation en trois étapes :

1. Choix du thème sur lequel l'auteur va écrire : ceci correspond au choix d'un cluster $\mathcal{C}_i$ d'où l'auteur va tirer sa source ;

2. Choix de la citation d'origine s_1 que l'auteur va prendre comme source ;
3. Création d'une nouvelle citation s_2 sur la base de s_1, et choix d'un nombre d'occurrences de s_2 à cet instant (i.e. le nombre de fois où la citation s_2 est répétée) ; cette nouvelle citation s'intègre au cluster et fait passer celui-ci de l'état $\mathcal{C}_i(t)$ à l'état $\mathcal{C}_i(t+dt)$ (l'instant t étant l'instant *précédant* la création de la citation) [25]. L'évolution du cluster $\mathcal{C}_i$ fait passer $\mathcal{P}$ de l'état $\mathcal{P}(t)$ à l'état $\mathcal{P}(t+dt)$.

Si $d_{H,tok}(s_1, s_2) = 1$ cette transformation est une substitution. Alors on note j l'index de cette substitution dans le cluster $\mathcal{C}_i$ (i.e. cette substitution est la j-ème dans le cluster), et on associe à la substitution la variable aléatoire $R_{i,j} \stackrel{\text{def}}{=} \frac{\mathfrak{C}(w_2)}{\mathfrak{C}(w_1)}$ à valeurs dans $\mathbb{R}_+^*$ ($R_{i,j}$ représente la variation de caractéristique sémantique observée lors de cette substitution). Enfin, on note n_i le nombre de substitutions qui ont lieu dans ce cluster (c'est une variable aléatoire à valeurs dans $\mathbb{N}$).

Prenons un exemple : le cluster est le numéro 5, et à l'instant t il y a déjà eu 7 substitutions dans ce cluster. Un auteur veut créer une nouvelle citation à partir de ce cluster, alors il sélectionne sa citation de départ : « Dessine-moi un mouton ». Il choisit sa citation d'arrivée : « Dessine-moi un patron ». Comme la distance de Hamming-tokens entre les deux citations est 1, cette transformation est une substitution, la 8ème dans ce cluster. On y associe donc la variable $R_{5,8} = \frac{\mathfrak{C}(\text{patron})}{\mathfrak{C}(\text{mouton})}$, variable qu'on stocke pour être analysée plus tard. À la fin de l'itération à travers toutes les substitutions ce cluster aura au moins 8 substitutions, donc $n_5 \geq 8$.

Discrétisation du modèle Pour simplifier la représentation du modèle on discrétise le temps. Ceci n'a pas d'impact sur le calcul des résultats puisqu'on ne modélise que le processus lui-même de création d'une citation, et non le *déclenchement* de ce processus ; or c'est seulement le processus de déclenchement qui dépend de la forme qu'on prend pour le temps. En discrétisant le temps on ne change donc que la distribution du nombre n_i de variables aléatoires $R_{i,j}$ (pour i fixé) et non la distribution des $R_{i,j}$ ni leurs propriétés d'indépendance. On permet surtout la représentation graphique du modèle faite sur la figure 7.

Approximation sur $\mathcal{P}$ Enfin, on fait l'approximation que l'effet de l'évolution des clusters sur $\mathcal{P}$ est négligeable, c'est-à-dire que $\mathcal{P}$ est stationnaire et que les choix d'un cluster dans $\mathcal{P}$ sont des évènements indépendants. En effet, le pool réel de clusters disponibles est probablement bien plus grand que celui qu'on observe puisque les données ne contiennent que des clusters pour lesquels de nouvelles citations apparaissent pendant la période d'observation (sans même contenir tous ces clusters-là), et non tous les autres clusters pour lesquels aucune citation n'apparaît pendant cette période ; on considère donc que le pool de clusters disponibles se comporte comme un thermostat, ce qui correspond à négliger l'effet des flèches $\mathcal{P}(t) \to \mathcal{P}(t+1)$ et $\mathcal{C}_i(t+1) \to \mathcal{P}(t+1)$ dans la figure 7. Ceci

25. Le cas bien réel où plusieurs citations, rattachées à des clusters différents ou non, sont créées au même instant dans le même billet de blog est modélisé par plusieurs instances des trois étapes décrites.

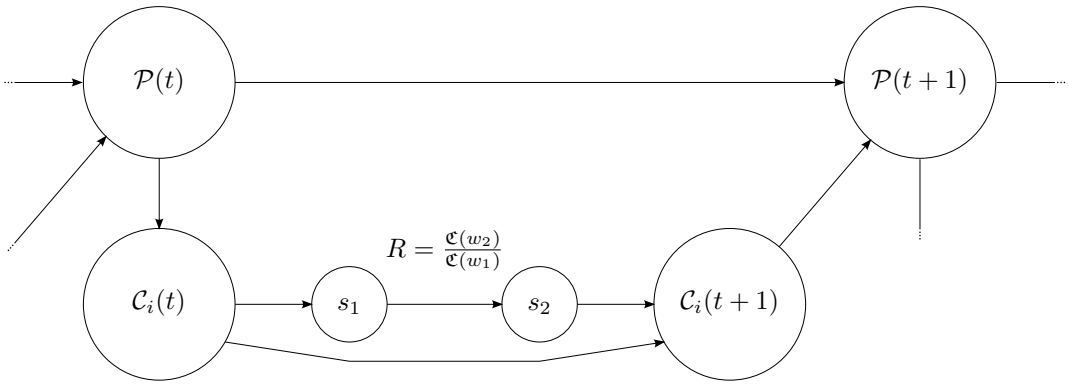


FIGURE 7 – Représentation graphique du modèle discret

implique que les variables aléatoires $R_{i,j}$ *pour des clusters différents* sont indépendantes, bien que les variables *à l'intérieur d'un même cluster* ne le soient pas.

Alors, pour $i \in I$, soit la suite $U_i \stackrel{\text{def}}{=} (R_{i,0}, R_{i,1}, \ldots, R_{i,n_i-1}, 0, \ldots)$. On peut voir nos observations comme les réalisations des variables aléatoires $U_i, i \in I$ qui sont à valeurs dans l'ensemble des suites de réels à support fini. On obtient une suite aléatoire par cluster, et l'approximation précédente revient à considérer que les $U_i, i \in I$ sont des variables aléatoires indépendantes et identiquement distribuées. Alors on peut définir les variables

$$\overline{R}_i \stackrel{\text{def}}{=} \frac{1}{n_i} \sum_{j=0}^{n_i-1} R_{i,j}, \quad i \in I$$

qui sont elles aussi indépendantes et identiquement distribuées, et qu'on va pouvoir utiliser pour obtenir les intervalles de confiance et les niveaux de significativité des résultats.

3.1.2 Observables

Le modèle ci-dessus nous permet de définir les deux observables qu'on cherche à mettre en évidence : la variation moyenne de caractéristique sémantique lors d'une substitution, et la susceptibilité des mots à être substitués.

Variation de caractéristique sémantique On reprend les variables aléatoires $\overline{R}_i, i \in I$ qu'on a défini au paragraphe précédent : ce sont les moyennes, par clusters, des ratios $\frac{\mathfrak{C}(w_2)}{\mathfrak{C}(w_1)}$ calculés pour chaque substitution détectée. À partir de ces variables, on estime la variation moyenne de caractéristique sémantique par la moyenne $\widehat{R}_N$:

$$\widehat{R}_N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i \in I} \overline{R}_i$$

Intervalle de confiance On voudrait maintenant avoir une idée de la précision des observations qu'on aura ; on sait que les mesures $\overline{R}_i, i \in I$ sont indépendantes et identiquement distribuées et, sans connaître leur distribution ou faire d'hypothèse supplémentaire, on peut construire un intervalle de confiance et un test asymptotiques qui se basent sur le grand nombre de mesures qu'on va obtenir. En effet, si l'on note $\mu \overset{\text{def}}{=} \mathbb{E}\left(\overline{R}_0\right)$ et $\sigma^2 \overset{\text{def}}{=} \mathbb{V}\left(\overline{R}_0\right)$ respectivement l'espérance et la variance de $\overline{R}_0$, et $\hat{\sigma}_N^{*2} \overset{\text{def}}{=} \frac{1}{N-1}\sum_{i\in I}\left(\widehat{R}_N - \overline{R}_i\right)^2$ l'estimateur non biaisé de σ^2, on a :

$$\begin{cases} \sqrt{N}\left(\widehat{R}_N - \mu\right) \xrightarrow[N\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right) & \text{(Théorème Central Limite)} \\ \hat{\sigma}_N^{*2} \xrightarrow[N\to+\infty]{p.s.} \sigma^2 & \text{(Loi Forte des Grands Nombres)} \end{cases}$$
$$\implies \frac{\sqrt{N}\left(\widehat{R}_N - \mu\right)}{\sqrt{\hat{\sigma}_N^{*2}}} \xrightarrow[N\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \quad \text{(Lemme de Slutsky)} \tag{1}$$

Cette dernière propriété nous permet de construire un intervalle de confiance asymptotique pour $\widehat{R}_N$: si l'on note $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale (tel que $\Phi\left(z_{1-\alpha/2}\right) = 1 - \alpha/2$ si Φ est la fonction de répartition de la loi normale), on a l'intervalle de confiance asymptotique de degré $1 - \alpha$ suivant :

$$\begin{aligned} IC_{1-\alpha} &= \left[\widehat{R}_N - \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}_N^{*2}}}{\sqrt{N}}, \widehat{R}_N + \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}_N^{*2}}}{\sqrt{N}}\right] \\ &= \left[\widehat{R}_N - \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}_N^{2}}}{\sqrt{N-1}}, \widehat{R}_N + \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}_N^{2}}}{\sqrt{N-1}}\right] \end{aligned} \tag{2}$$

où l'on a noté $\hat{\sigma}_N^2 \overset{\text{def}}{=} \frac{N-1}{N}\hat{\sigma}_N^{*2}$ l'estimateur biaisé de la variance. C'est la formule (2) qu'on a utilisé pour le calcul des intervalles de confiance montrés dans la partie suivante [26] (partie 3.2, page 34).

Significativité Pour savoir si la valeur mesurée de l'observable « variation de caractéristique sémantique » est significative, il y a un point important à ne pas négliger : comme cette observable est essentiellement une moyenne de ratios, il ne suffit pas de comparer la valeur mesurée à 1. En effet, on peut calculer la valeur de référence en supposant que les mots de départ et d'arrivée sont tirés aléatoirement et de façon indépendante, et cette

26. On note bien que cet intervalle de confiance est *asymptotique* : il n'est valable que dans la limite où $N \to +\infty$. En pratique, ceci est bien justifié car le nombre de mesures indépendantes qu'on obtient varie de la centaine à plus de 3 000.

référence est *nécessairement plus grande* que 1 (quelle que soit la distribution de tirage des mots, tant que c'est la même pour les mots de départ et d'arrivée).

Pour mieux le voir, on considère justement l'hypothèse nulle $\mathcal{H}_0$ selon laquelle les mots de départ et d'arrivée sont tirés de façon indépendante dans le vocabulaire répertorié par le graphe qu'on a construit à partir de WordNet ; on compare cette hypothèse à l'hypothèse $\mathcal{H}_1$ où ce n'est pas le cas, c'est-à-dire où il y a un effet du cerveau dans le tirage des mots[27]. Notons W_1 et W_2 les variables aléatoires, indépendantes et identiquement distribuées, correspondant respectivement au tirage d'un mot de départ et au tirage d'un mot d'arrivée de façon uniforme dans le vocabulaire répertorié par notre graphe. Alors, sous $\mathcal{H}_0$, on a pour tout i, j :

$$\begin{aligned}\mathbb{E}_0\left(R_{i,j}\right) &= \mathbb{E}_0\left(\frac{\mathfrak{C}(W_2)}{\mathfrak{C}(W_1)}\right) \quad (\mathbb{E}_0 \text{ désigne l'espérance sous } \mathcal{H}_0)\\ &= \mathbb{E}_0\left(\mathfrak{C}(W_2)\right)\cdot\mathbb{E}_0\left(\frac{1}{\mathfrak{C}(W_1)}\right) \quad \text{(parce que les tirages sont indépendants)}\end{aligned}$$

et donc la valeur de référence est $\mu_0 = \mathbb{E}_0\left(\mathfrak{C}(W_2)\right)\cdot\mathbb{E}_0\left(\frac{1}{\mathfrak{C}(W_1)}\right)$, et *cette valeur est toujours supérieure ou égale à 1*.[28] On remarque aussi que cette valeur de référence va varier en fonction de la catégorie grammaticale (puisque la distribution sur W_1 et W_2 n'est plus la même quand la catégorie grammaticale change).

Pour compléter le test, on définit la statistique $T_N \overset{\text{def}}{=} \frac{\sqrt{N}\left(\widehat{R}_N - \mu_0\right)}{\sqrt{\widehat{\sigma}^{*2}_N}}$ qui suit une loi asymp-

27. Il faut noter qu'il est difficile de différencier le fait que les mots tirés soient indépendants, d'un côté, du fait qu'ils soient tirés selon une autre distribution que la distribution uniforme dans le vocabulaire répertorié par notre graphe, de l'autre. L'alternative au test qu'on présente ici, qui serait de prendre l'hypothèse $\mathcal{H}'_0$ où les mots sont tirés de façon indépendante mais suivant les distributions empiriques qu'on trouve dans les résultats pour les mots de départ et d'arrivée, ne testerait en effet que l'indépendance des mots mais passerait à côté d'un effet venant de la forme des distributions de départ et d'arrivée : en effet, imaginons un cas (dégénéré, mais dont le principe reste vrai) où la distribution de départ ne sélectionne que des mots tels que $\mathfrak{C}(w) \leq 1$, et où la distribution d'arrivée ne sélectionne que des mots tels que $\mathfrak{C}(w) \geq 2$, mais où les substitutions ne se font que depuis des mots de caractéristique sémantique *égale* à 1 vers des mots de caractéristique *égale* à 2. Dans ce cas il y a clairement un effet du cerveau dans la sélection et la substitution des mots, et cet effet va dans le sens de l'augmentation de la caractéristique sémantique des mots. Le test qu'on présente dans le corps du texte verrait cet effet, parce que l'augmentation de caractéristique serait supérieure à celle dans le cas $\mathcal{H}_0$. Mais le test alternatif passerait à côté de cet effet et conclurait d'une part que les mots tirés ne sont pas indépendants (ce qui est bien le cas), d'autre part que le cerveau fait augmenter la caractéristique sémantique *moins* que si les mots étaient indépendants et tirés selon les deux distributions particulières décrites au-dessus ; ceci parce que la variation de caractéristique observée serait effectivement inférieure à celle dans le cas $\mathcal{H}'_0$. On penserait donc que la tendance du cerveau est à freiner l'augmentation de caractéristique, alors que dans ce cas elle favorise clairement l'augmentation de caractéristique. (Le principe de cet exemple est que l'effet du cerveau peut se retrouver dans les distributions de départ et d'arrivée, et peut ne pas se retrouver dans la dépendance entre les mots.) La partie 4.2 (page 41) mentionnera une solution envisageable à ce problème, qu'on a pas pu utiliser ici.

28. La démonstration de ce point est très simple : puisque la fonction $x \mapsto \frac{1}{x}$ est convexe, on applique l'inégalité de Jensen pour trouver que $\mathbb{E}_0\left(\frac{1}{\mathfrak{C}(W)}\right) \geq \frac{1}{\mathbb{E}_0(\mathfrak{C}(W))}$ quelle que soit la distribution de W. On aura donc toujours $\mu_0 \geq 1$ sous l'hypothèse nulle.

totiquement normale (d'après la propriété (1), page 32), et permet de définir le test asymptotique bilatère de niveau α suivant :

$$\begin{cases} \mathcal{H}_0 : & T_N \in \left[z_{\alpha/2}, z_{1-\alpha/2}\right] \\ \mathcal{H}_1 : & T_N < z_{\alpha/2} \text{ ou } T_N > z_{1-\alpha/2} \end{cases}$$

qui donne une p-valeur $p = 2 \cdot (1 - \Phi(|T_N|))$ (en notant encore Φ la fonction de répartition de la loi normale). Autrement dit, si on observe T_N en dehors de $\left[z_{\alpha/2}, z_{1-\alpha/2}\right]$, le résultat est significatif de niveau α ; il l'est même jusqu'au niveau p. C'est ce test qu'on a utilisé dans la partie suivante (partie 3.2, ci-dessous) pour évaluer la significativité des résultats.

Susceptibilité Enfin, on définit pour chaque mot w une notion de susceptibilité en comparant le nombre de fois où w est apparu en position substituable au nombre de fois où il a effectivement été substitué ; on la note $\mathfrak{S}(w)$. On considère qu'un mot est en position substituable lorsqu'il est dans une citation où un mot est substitué (celui-là ou un autre). Pour construire cette observable on utilise deux compteurs pour chaque mot : un pour le nombre d'apparitions en position substituable, qu'on note $c_p(w)$ (p pour *possible*), et l'autre pour le nombre de substitutions effectivement subies, qu'on note $c_r(w)$ (r pour *réalisée*). À chaque occurrence d'une citation comme dans la figure 7, on ajoute 1 au compteur de chacun des mots de s_1 et on ajoute 1 au compteur du mot substitué lors du passage de s_1 à s_2. Une fois terminée l'itération à travers toutes les substitutions, il suffit de prendre le rapport des deux compteurs de w pour obtenir $\mathfrak{S}(w) = \frac{c_r(w)}{c_p(w)}$, qui prend ses valeurs entre 0 et 1.

3.2 Résultats des mesures sur les observables

On peut enfin montrer les résultats sur les observables, obtenus après itération à travers toutes les substitutions avec les différents jeux de paramètres (détaillés en partie 2.2.4, page 27).

Le premier point important à noter est que les résultats ne varient essentiellement pas en fonction du jeu de paramètres : la seule chose qui varie est la significativité de certains résultats, et ce d'une façon qui n'a pas paru interprétable. On présente ici les résultats pour le modèle *fixed timebags* (modèle 2, détaillé page 19) avec $n = 5$, c'est-à-dire 5 sacs de citations sur chaque cluster.

On regarde successivement les deux caractéristiques sémantiques (PageRank et coefficient de regroupement), pour terminer avec quelques remarques sur les résultats quant aux catégories grammaticales des mots échangés lors d'une substitution.

3.2.1 PageRank

La figure 8 montre les résultats pour le PageRank, en fonction des catégories grammaticales. Sur cette figure, le pointillé noir représente la valeur de référence sous l'hypothèse $\mathcal{H}_0$ décrite pour le test statistique (on remarque qu'elle est supérieure à 1, et qu'elle varie

en fonction de la catégorie grammaticale ; voir le paragraphe sur la significativité, partie 3.1.2, page 32, pour plus de détails) ; les points bleus sont les $\widehat{R}$ mesurés, et les points magenta représentent les bornes de l'intervalle de confiance à 5% autour de $\widehat{R}$. Concrètement, le test implique que les valeurs bleues sont significatives à 5% si et seulement si le pointillé noir est en dehors de l'intervalle formé par les points magenta.

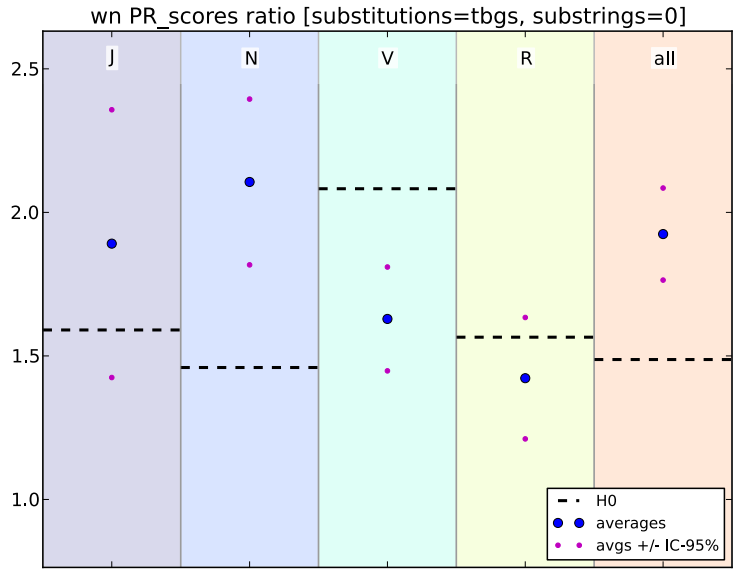


FIGURE 8 – Variations moyennes de PageRank lors d'une substitution. La lettre J veut dire adjectif, N veut dire nom, V veut dire verbe, et R veut dire adverbe. La dernière colonne, all, est le résultat en intégrant toutes les catégories.

On observe tout d'abord que de façon générale, les auteurs de nouvelles citations utilisent en moyenne des mots de polysémie plus importante lorsqu'ils font une substitution, et ce de façon significative ($p < 10^{-7}$ pour le PageRank). Mais cette variation de polysémie est en fait tributaire de la catégorie grammaticale ; en effet les noms et les verbes ne sont pas traités de la même façon : les noms augmentent (significativement) en polysémie, alors que les verbes diminuent (significativement aussi). Le cas des adjectifs et des adverbes est moins clair, mais on observe tout de même une tendance à la hausse pour les premiers et une tendance à la baisse pour les seconds (mais ces résultats ne sont pas significatifs).

Pour observer la susceptibilité des mots à être substitués, on forme d'abord l'histogramme des caractéristiques sémantiques sur lequel on représente la susceptibilité en couleurs. On voit ainsi la propension qu'ont les mots à être substitués, en fonction de leur caractéristique sémantique et en relation avec le nombre de mots qui ont cette caractéristique-là. La figure 9 montre le résultat pour le PageRank, toute catégories grammaticales confon-

dues. La figure 10 (page 37) montre cette même susceptibilité différenciée selon les catégories grammaticales.

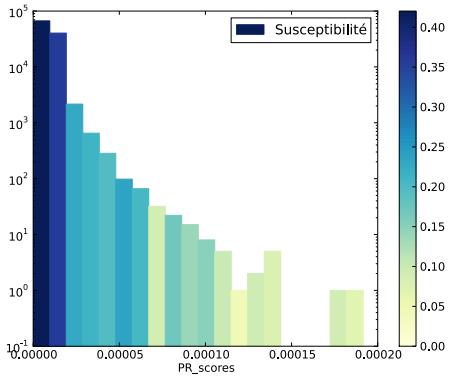
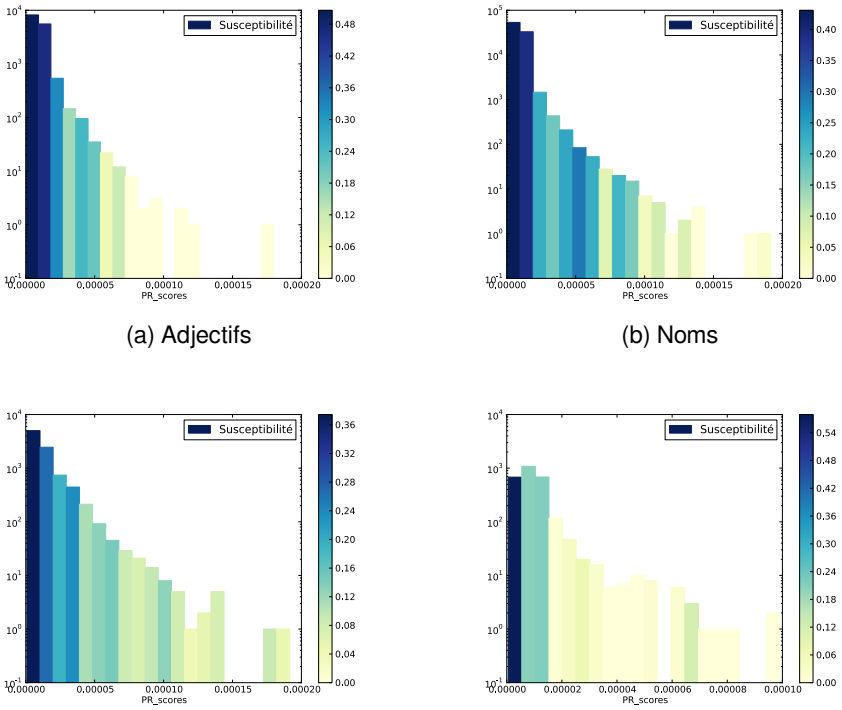


FIGURE 9 – Susceptibilité des mots à être substitués en fonction du PageRank, en couleurs sur l'histogramme des valeurs de celui-ci, pour toutes les catégories grammaticales confondues

Il apparaît clairement sur ces figures que les mots de plus faible PageRank ont la plus forte propension à être substitués : les mots peu polysémiques ont une plus forte tendance à être substitués, indépendamment de la catégorie grammaticale des mots en question



(a) Adjectifs

(b) Noms

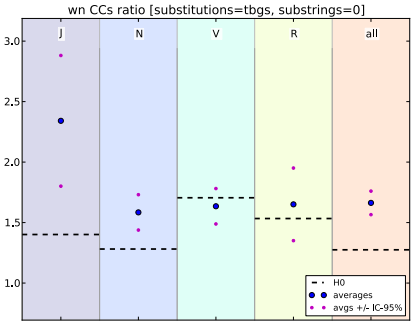
(c) Verbes

(d) Adverbes

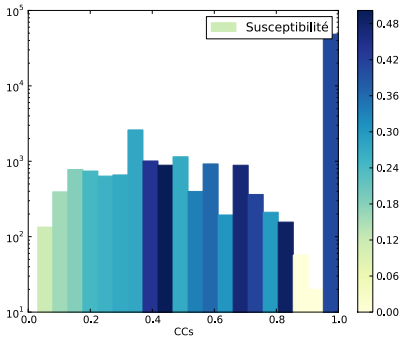
FIGURE 10 – Susceptibilité en fonction du PageRank, différenciée selon les catégories grammaticales (l'histogramme est différent selon la catégorie, car on le restreint aux mots de cette catégorie-là)

3.2.2 Coefficient de regroupement

Le coefficient de regroupement (voir partie 2.2.2, page 23) joue un autre rôle, et ses résultats, bien que similaires, sont plus compliqués à interpréter. La figure 11 montre les résultats pour cette caractéristique sémantique (avec les mêmes conventions que pour les figures à propos de PageRank).



(a) Variations de coefficient de regroupement. La lettre J veut dire adjectif, N veut dire nom, V veut dire verbe, et R veut dire adverbe. La dernière colonne, all, est le résultat en intégrant toutes les catégories.



(b) Susceptibilité en fonction du coefficient de regroupement, toutes catégories grammaticales confondues

FIGURE 11 – Résultats pour le coefficient de regroupement

On observe que le coefficient de regroupement augmente lui aussi lors d'une substitution (figure 11a) : l'augmentation est significative pour l'ensemble des catégories grammaticales confondues et individuellement pour les adjectifs et les noms. Une substitution remplace donc les mots par d'autres mots à plus fort coefficient de regroupement, c'est-à-dire par des mots ayant des significations plus liées sémantiquement entre elles [29]. On remarque que cette augmentation est moins tributaire des catégories grammaticales que dans le cas du PageRank.

Enfin, on note que les mots les plus susceptibles d'être substitués sont ceux à plus plus fort coefficient de regroupement. On peut relier cette observation au résultat montré par Chan and Vitevitch (2010) : les mots à forts coefficient de regroupement ont tendance à être plus difficiles à retrouver en mémoire (dans le cas de l'étude de Chan and Vitevitch). Dans notre cas, les mots à fort coefficient de regroupement ont tendance à ne pas être reproduits (i.e. ils sont substitués) lors de la reproduction d'une citation.

29. Voir l'annexe A.2 pour la justification de cette interprétation.

3.2.3 Catégories grammaticales

Un dernier point à remarquer est la variabilité des résultats en fonction de la catégorie grammaticale, notamment en ce qui concerne le PageRank. Il est clair sur la figure 8 (page 35) que les verbes et les adverbes, d'un côté, et les adjectifs et les noms, de l'autre, ne sont pas traités de la même manière. Ceci est vrai tout particulièrement pour les noms et les verbes : les uns augmentent significativement plus que sous $\mathcal{H}_0$, les autres augmentent significativement moins que sous $\mathcal{H}_0$. D'autre part, on voit que la catégorie des adverbes a une signature en termes de susceptibilité différente de celle des autres catégories, et le contraste est particulièrement fort avec les noms (figure 10, page 37). (Ces observations ne sont que des tendances et il faudrait en vérifier la significativité.)

3.3 Interprétation

Les résultats obtenus sont bien cohérents avec les travaux sur lesquels on s'est basé : d'une part, le PageRank sémantique a tendance à augmenter lorsqu'un mot est substitué par un autre. Ceci est cohérent avec l'idée selon laquelle le retrait d'un mot en mémoire suivrait un processus d'exploration du réseau sémantique formé par les mots (comme le suggèrent Griffiths et al., 2007), situation dans laquelle les mots à plus haut PageRank sémantique sont des attracteurs de l'exploration du réseau. D'autre part, les mots à haut coefficient de regroupement ont une forte tendance à être substitués : cette observation est à mettre en lien avec le travail de Chan and Vitevitch (2010), montrant que les mots à haut coefficient de regroupement sur un réseau *phonologique* (non pas sémantique) sont plus difficiles à retrouver en mémoire que les mots à faible coefficient de regroupement.

En revanche, il serait hasardeux d'interpréter ces résultats dans un cadre plus large que les seules substitutions [30] : en effet, le rôle que joue la topologie des réseaux aussi bien sémantiques que phonologiques dans le traitement du langage par le cerveau est une question totalement ouverte. L'étude de ces réseaux en lien avec la cognition est une question qui n'est apparue que très récemment. Les études existantes lient des propriétés de ces réseaux à des observations psychologiques (par exemple Borge-Holthoefer and Arenas, 2010, mettent en relation certaines propriétés de connectivité du réseau sémantique des enfants avec la richesse du vocabulaire de ces enfants lors de leur développement ; ou encore Troyer et al., 1997, suggèrent que certaines capacités de navigation mentale sur les réseaux sémantiques et phonologiques jouent un rôle central dans des tâches de fluidité de langage), mais elles restent souvent à un niveau conceptuel très élevé ; à notre connaissance, il n'y a pas de travaux expliquant, au niveau neurologique, pourquoi et comment des propriétés des réseaux sémantiques et phonologiques peuvent avoir un rôle dans le traitement cognitif des mots du langage (bien qu'il soit clair que ce rôle existe). Notamment, il n'existe pas de modèle prédictif de la relation entre ces réseaux et des phénomènes cognitifs.

30. On pourrait par exemple interpréter en termes de polysémie le fait que les mots à faible PageRank sont plus susceptibles d'être substitués, en avançant l'hypothèse selon laquelle le cerveau aurait plus de difficultés à retenir exactement les mots peu polysémiques dans une phrase, et plus de facilités à retenir les mots plus polysémiques.

On peut enfin ajouter à cette description de l'état actuel de ce champ le fait que l'étude des réseaux complexes eux-mêmes en est encore à ses balbutiements. On a ici utilisé deux propriétés des nœuds, à savoir le PageRank et le coefficient de regroupement, qui ne sont que deux parmi des dizaines voir même des centaines de propriétés qu'il est possible de calculer sur ces nœuds ; en d'autres termes il n'existe pas à ce jour de façon de catégoriser les réseaux, parce qu'on ne sait pas ce qui les caractérise en leur essence.

Tout ceci justifie qu'on ait limité ici l'interprétation des résultats obtenus au champ dans lequel ils ont été mesurés, bien qu'il serait souhaitable d'élargir ce champ d'interprétation.

4 Conclusions et perspectives

4.1 Retour à la question de départ

On a cherché s'il était possible d'observer un biais cognitif dans les transformations subies par les représentations publiques, à l'aide de données issues d'Internet. Au travers de l'élaboration de plusieurs modèles sous-jacents, ainsi que de plusieurs mesures sur les évolutions détectées, on a pu construire des observables nous renseignant sur la façon dont les citations sont transformées lorsqu'elles sont reproduites par les auteurs. On a ainsi observé un biais clair sélectionnant tout particulièrement les mots les moins polysémiques pour les faire évoluer vers des mots plus polysémiques (c'est-à-dire plus polysémiques que s'ils étaient tirés au hasard), et un biais analogue faisant augmenter les liens entre les synonymes des mots qu'on substitue (c'est l'augmentation de coefficient de regroupement).

Il est donc bien possible de donner une réalité empirique aux notions de représentation publique et d'attracteur culturel. Les biais qu'on a mis en évidence sont précisément le genre de phénomènes dont il faut pouvoir rendre compte et qu'il faut pouvoir tester pour faire avancer le programme de l'épidémiologie culturelle au-delà des modèles normatifs. Il faut noter cependant que même dans un cas en apparence aussi simple que celui des citations, l'extraction et le traitement des informations est particulièrement ardu.

Dans un autre registre, on peut enfin mentionner l'observation du traitement différent subi par les différentes catégories grammaticales : l'affinage des observables à ce niveau a permis de voir que certaines catégories, notamment les verbes, se comportent individuellement d'une façon presque opposée au comportement moyen décrit ci-dessus. Ce résultat laisse à penser qu'on pourrait utiliser l'analyse des substitutions (ou plus largement l'analyse des transformations des phrases sur Internet) pour extraire de nouvelles propriétés sur les catégories grammaticales, voire même délimiter de façon intrinsèque ces catégories.

4.2 Perspectives

Questions ouvertes Comme souvent, l'étude qu'on a faite ici soulève plus de nouvelles questions qu'elle n'en résout. L'analyse présentée peut être approfondie d'au moins deux façons pour répondre aux questions laissées ouvertes dans les parties précédentes.

Premièrement, la question des caractéristiques sémantiques qu'on utilise, pour laquelle on s'est basé sur des avancées récentes en psycholinguistique (Griffiths et al., 2007 ; Chan and Vitevitch, 2010), peut être explorée plus à fond. En effet, les caractéristiques pourraient être approfondies aussi bien du point de vue sémantique, en allant au-delà du PageRank et du coefficient de regroupement, que du point de vue syntaxique, en dépassant le stade du simple mot. On pourrait même aller plus loin dans la dimension cognitive en étudiant un support plus riche que les simples citations et un mécanisme plus complexe que le copier-coller qu'on a choisi d'observer ici. En développant un grand nombre de telles caractéristiques il serait possible d'extraire les composantes principales

sur lesquelles les transformations s'opèrent, affinant ainsi la compréhension des mécanismes impliqués dans les transformations. Les deux barrières à ces avancées sont le développement des caractéristiques, d'un côté, et le temps de calcul pour les résultats, de l'autre.

Deuxièmement, on n'a pas pu séparer dans les substitutions l'effet de la dépendance entre les mots échangés, d'un côté, de l'effet des distributions d'où sont tirés ces mots, de l'autre. Une façon de faire serait d'estimer la distribution conditionnelle de la caractéristique sémantique du mot d'arrivée *sachant* la caractéristique sémantique du mot de départ. Cette distribution nous informerait sur le lien entre les mots échangés dans les substitutions sans avoir à faire de ratios ou de différences de caractéristiques sémantiques, nous affranchissant ainsi des problèmes rencontrés ci-dessus. Malheureusement, bien que le corpus de données utilisé soit de grande taille, on n'a pas assez de mesures pour construire une telle estimation (en effet, cette analyse nécessite de mesurer le même type d'observable que celles qu'on a mesurées, mais ces nouvelles observables se répartissent dans un espace de dimension deux ; il faut donc multiplier le nombre de mesures nécessaires par le nombre de points qu'on veut estimer dans cet espace). Mais ici aussi, les seules barrières sont la quantité des données récoltées et le temps de calcul : il « suffit » d'augmenter la taille du corpus utilisé pour obtenir plus de mesures et répondre à cette question.

Possibilités futures On termine en mentionnant les possibilités ouvertes non pas par les résultats, mais par l'outil développé pour extraire ces résultats : le code de l'analyse est publié et documenté [31], et a été écrit dans un souci de modularité qui rend l'ajout de nouvelles caractéristiques sémantiques et de nouvelles façons de détecter des substitutions très rapide (quelques minutes si les caractéristiques sont déjà calculées). Il serait par exemple très facile d'appliquer cet outil pour relier plus rigoureusement les expériences étudiant l'effet d'autres réseaux de mots (sémantiques, phonologiques, syntaxiques) sur la production orale ou écrite, d'un côté, et l'effet de ces réseaux sur les substitutions, de l'autre côté. Plus largement, l'analyse rigoureuse de données provenant d'Internet, en tandem avec la publication du code qui permet ces analyses, ouvre aux sciences cognitives et sociales de très nombreuses possibilités qui semblent avoir été rarement envisagées jusqu'à présent.

31. Le code source est disponible à l'adresse suivante : https://code.launchpad.net/~seblerique/socio-semantic-representations/main-seb.

A Détails du protocole expérimental

Cette annexe reprend les points de détail qu'on a pas intégrés au corps du rapport pour ne pas en alourdir la lecture.

A.1 Détection des substitutions

On complète ici l'explication de la façon dont on a détecté les substitutions dans le corpus de données. Le corps du rapport mentionne en effet qu'on a défini six modèles pour détecter ces substitutions, dont seulement deux ont été détaillés. Voici les quatre autres modèles qu'on a développés et implémentés.

Modèle 3. ***Growing timebags*** **:** On reprend le modèle *sliding timebags* (modèle 1, présenté page 18), mais on néglige le déclin de l'influence d'une citation dans le temps : au lieu de définir les sacs de citations sur $[t - \Delta t, t]$ on les définit sur $[t_{init}, t]$, de façon à prendre en compte l'influence de chaque citation *depuis le début du cluster*. La détection des substitutions se fait de la même manière que dans le modèle *sliding timebags* (modèle 1).

Ce modèle est illustré dans la figure 12. Il souffre du même problème que le modèle *sliding timebags* (modèle 1).

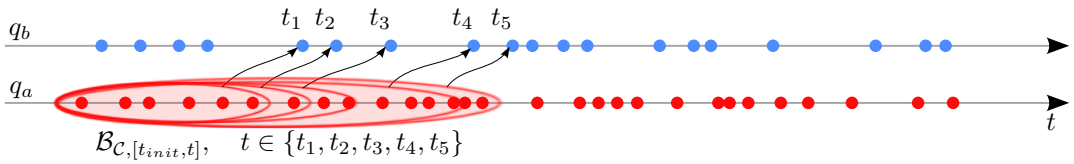


FIGURE 12 – Problème du modèle *sliding timebags* (modèle 1) reproduit dans le modèle *growing timebags* (modèle 3)

Modèle 4. ***Cumulative fixed timebags*** **:** Ce modèle-ci est au modèle *fixed timebags* (modèle 2, présenté page 19) ce que le modèle *growing timebags* (défini ci-dessus) est au modèle *sliding timebags* (modèle 1) : lorsqu'une nouvelle citation apparaît, on prend en compte les citations précédentes *depuis le début du cluster*. Pour cela, au lieu de découper chaque cluster en sacs de citations définis sur $[t_{init} + i\Delta t, t_{init} + (i+1)\Delta t]$, on définit ces sacs sur les périodes $[t_{init}, t_{init} + (i+1)\Delta t]$. On note ces sacs « cumulés » $\mathcal{B}'_{\mathcal{C},i}$.

Lorsqu'un auteur crée une nouvelle citation, il le fait à partir de la citation la plus fréquente dans le sac de citations cumulé ainsi défini depuis l'instant initial du cluster. La détection d'une substitution se fait comme précédemment : lorsqu'une citation q apparaît à l'instant t, on la considère comme une substitution d'une autre citation si et seulement si $q_m(\mathcal{B}'_{\mathcal{C},i}) \in S_{d_{H,tok}}(q, 1)$ où $i = \lfloor \frac{t - t_{init}}{T} \rfloor$ (remarquer que c'est $\mathcal{B}'$ et non $\mathcal{B}$ qui intervient). Si la citation q apparaît plusieurs fois dans le sac de citations *non cumulé* dans lequel elle se trouve, on ne compte cette substitution qu'une seule fois.

Ce modèle utilise les mêmes approximations que le modèle *fixed timebags* (modèle 2). La figure 13 (page 44) illustre ce modèle.

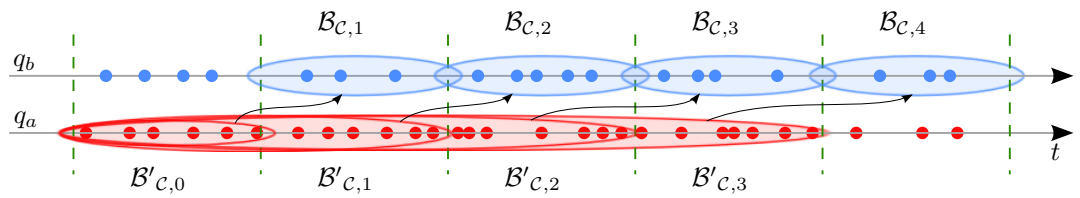


FIGURE 13 – Illustration du modèle 4 pour la situation problématique du modèle 1

Modèle 5. ***Root*** **:** On utilise un des sous-produits de l'analyse faite par Leskovec et al. (2009a) sur le corpus de données utilisé : leur algorithme de groupement des citations en clusters produit une citation racine pour chaque cluster, de laquelle les membres du cluster sont des sous-chaînes de caractères (ou presque, parce que leur algorithme permet certaines approximations). Le modèle est le suivant : toute nouvelle citation est faite à partir de la citation racine du cluster auquel elle se rattache. Pour éviter le même biais que dans le modèle *sliding timebags* (modèle 1), on découpe ici aussi les clusters en sacs de citations et on ne compte qu'une seule substitution par sac de citations.

On détecte une substitution de la façon suivante : lorsque la citation q apparaît à l'instant t, on considère que c'est une substitution si et seulement si elle est à distance de Hamming-tokens 1 de la citation racine du cluster.

La figure 14 illustre ce modèle.

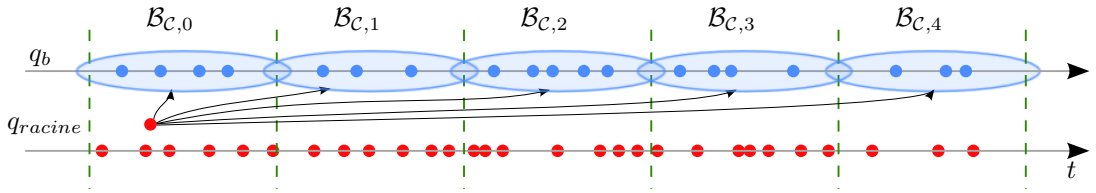


FIGURE 14 – Illustration du modèle 5 pour la situation problématique du modèle 1

Modèle 6. ***Time*** **:** Les modèles précédents n'autorisent comme citation mère d'une substitution qu'une citation considérée comme « racine », ou une citation particulièrement visible par sa fréquence d'apparition. Ce modèle-ci vise à prendre en compte toutes les substitutions extractibles et cohérentes temporellement, c'est-à-dire pour lesquelles la citation mère est apparue avant la citation fille.

Pour cela, on considère que lorsqu'un auteur crée une nouvelle citation il peut le faire à partir de n'importe quelle citation déjà présente à cet instant quelles que soient les fréquences d'apparition. S'il transforme cette citation source, on suppose qu'il est *le premier*

à le faire de cette façon, ce qui implique que la citation qu'il crée soit nouvelle (elle n'est jamais apparue avant ce moment). On exclut le cas contraire, où l'auteur transformerait sa source d'une façon qui a déjà été faite, produisant ainsi une citation qui serait déjà apparue avant ce moment. Ce choix de modèle implique que les substitutions détectées sont exactement celles allant d'une citation q_a à *toutes* les citations à distance de Hamming-tokens 1 de q_a qui sont apparues *pour la première fois après* q_a, en ne comptant ces substitutions qu'une seule fois (et ce pour toutes les citations q_a). En particulier, aucune des occurrences d'une citation après sa première apparition ne peut être comptée comme le résultat d'une substitution (puisque sinon l'auteur aurait reproduit une transformation ayant déjà été faite avant) ; d'autre part, la *première* occurrence d'une citation q est comptée comme le résultat d'autant de substitutions qu'il existe de citations à distance de Hamming-tokens 1 de q et qui sont apparues *avant* q (en comptant une substitution pour chaque citation à distance 1 précédant q).

On exclut donc des situations qu'on voit dans la figure 4, comme le passage de q_a dans le sac $\mathcal{B}_{\mathcal{C},1}$ à q_b dans le sac $\mathcal{B}_{\mathcal{C},2}$. En revanche, ce mode de détection va englober toutes les substitutions envisageables dans les modèles précédents, mais probablement en comptant chaque substitution moins de fois que dans ces modèles (puisqu'elles ne sont comptées qu'une fois ici).

La figure 15 illustre ce modèle.

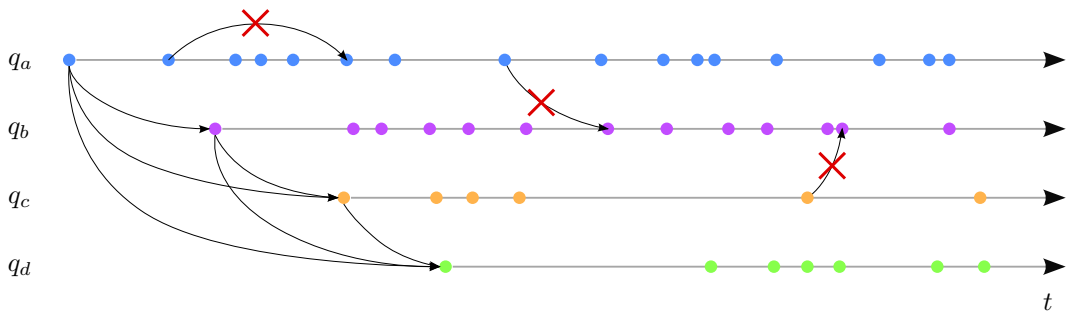


FIGURE 15 – Illustration du modèle *Time* (modèle 6)

A.2 Caractéristiques sémantiques

Le calcul et l'interprétation des caractéristiques sémantiques (présentées dans la partie 2.2.2, page 23) ne sont pas triviaux. On en explique ici le détail, d'abord pour le PageRank puis pour le coefficient de regroupement.

A.2.1 PageRank

Le PageRank est un algorithme de calcul de la *centralité* des nœuds d'un graphe. Il a été introduit par Page et al. (1999) et fait notoirement partie des algorithmes utilisés

par l'indexation Web de Google, mais ses applications vont bien au-delà de la recherche sur Internet ; c'est en fait un algorithme de centralité utilisé couramment dans l'étude des graphes.

Définition et calcul L'idée derrière le PageRank est que les nœuds centraux dans un graphe sont ceux qui sont liés à beaucoup de nœuds qui eux-mêmes sont liés à beaucoup de nœuds qui eux-mêmes sont liés à beaucoup de nœuds qui eux-mêmes... et ainsi de suite récursivement. La façon intuitive de comprendre comment le PageRank est défini est d'imaginer une personne se déplaçant aléatoirement de nœud en nœud sur le graphe. Si on laisse cette personne se déplacer sur le graphe pendant un temps très long (infini en fait), alors le PageRank de chaque nœud est défini comme la probabilité que ce marcheur aléatoire soit sur ce nœud après ce temps de marche très long. Mathématiquement, le PageRank correspond aux valeurs (normalisées pour que la somme des PageRanks fasse 1) du vecteur propre associé à la valeur propre 1 de la matrice d'adjacence du graphe normalisée sur les colonnes. On trouvera une explication complète du sens de ce qui précède dans Page et al. (1999).

Le calcul du PageRank revient à déterminer le vecteur propre associé à la plus grande valeur propre d'une certaine matrice (la matrice d'adjacence du graphe, normalisée sur les colonnes) ; cette matrice a autant de lignes qu'il y a de nœuds dans le graphe. Donc si le graphe est grand, comme c'est le cas pour notre graphe construit à partir de WordNet (plus de 100 000 nœuds), ce calcul peut poser certains problèmes techniques (même si la matrice en question contient beaucoup de 0, comme c'est encore le cas ici). On a testé deux bibliothèques en Python pour calculer le PageRank sur notre graphe : la bibliothèque d'algèbre linéaire de SciPy [32], et la bibliothèque SLEPc [33]. Les résultats calculés avec ces deux bibliothèques ne concordant pas, on a finalement implémenté le calcul nous-mêmes en utilisant l'algorithme des puissances [34], ce qui nous a permis d'obtenir des valeurs cohérentes. Pour une précision relative de 10^{-15}, le calcul dure de l'ordre d'une heure sur un ordinateur portable puissant.

Interprétation Vu sa définition, on remarque que le PageRank est une sorte de degré augmenté (ou récursif) ; on observe d'ailleurs empiriquement que, dans le cas du graphe qu'on a construit sur la base de WordNet, le degré et le PageRank diffèrent peu [35]. Prenons donc d'abord le cas du degré pour l'interprétation du PageRank : dans le graphe qu'on a construit, le degré d'un mot est la somme, pour chacune des significations de ce mot, du nombre de synonymes que ce mot a pour cette signification (par exemple sur la figure 5a, page 24, « car » a 5 significations et un degré de 10). Donc un mot ayant beaucoup de significations aura un degré élevé, tout comme un mot ayant beaucoup de synonymes. Un mot ayant beaucoup de significations et beaucoup de synonymes dans chacune de ses

32. Voir http://docs.scipy.org/doc/scipy/reference/linalg.html pour plus de détails.
33. Voir http://www.grycap.upv.es/slepc/ pour plus de détails.
34. Voir http://en.wikipedia.org/wiki/Power_iteration pour plus de détails.
35. Ceci n'est pas toujours vrai, et est dû à la structure du graphe qu'on utilise.

significations aura un degré très élevé. On voit que le degré est une bonne indication de la polysémie des mots (au sens où un mot est polysémique s'il a non seulement plusieurs significations, mais aussi plusieurs synonymes pour chacune de ses significations). Alors, reprenant l'interprétation du PageRank faite au début du paragraphe précédent et en prenant en compte que le degré indique une certaine notion de polysémie, le PageRank sera un bon indicateur des mots qui sont polysémiques en étant liés à des mots polysémiques qui sont liés à des mots polysémiques qui... et ainsi de suite. Le PageRank apparaît donc comme un indicateur de la polysémie encore plus fin que le degré.

A.2.2 Coefficient de regroupement

On a déjà défini le coefficient de regroupement dans le corps du rapport (partie 2.2.2, page 23) : cette mesure estime à quel point les voisins d'un nœud sont connectés entre eux, c'est-à-dire à quel point les voisins de ce nœud forment une clique [36]. Formellement, pour calculer le coefficient de regroupement d'un mot w : on compte le nombre de liens n_r (r pour « réel ») présents entre les voisins de w, ainsi que le nombre n_p (p pour « possible ») de liens *possibles* entre les voisins de w ; alors, par définition, le coefficient de regroupement de w est $\frac{n_r}{n_p}$. Un nœud peut donc avoir un très haut degré (i.e. beaucoup de voisins), et un coefficient de regroupement très faible. À l'inverse, un nœud à faible degré peut tout à fait avoir un fort coefficient de regroupement.

Si ce coefficient joue un rôle central dans des études comme celle de Chan and Vitevitch (2010), l'interprétation de ce qu'il représente dans le cas qui nous intéresse n'est pas pour autant triviale. Il faut pour cela distinguer les facteurs qui influent dessus.

Interprétation Dans notre graphe, les voisins d'un mot sont les synonymes de ce mot pour les différentes significations qu'il peut revêtir. Si un mot w n'a qu'une seule signification, tous ses voisins font partie de cette unique signification et sont donc tous connectés entre eux ; ce mot aura donc un coefficient de regroupement égal à 1.

Supposons maintenant que w ait deux significations, notées s_1 et s_2. Il se peut que les mots appartenant à ces deux significations se recoupent beaucoup (par exemple les synonymes de « yell » pour sa signification « le fait de crier », signification pour laquelle « yell » est un verbe, et ses synonymes pour sa signification « cri », signification pour laquelle « yell » est un nom, ont beaucoup de mots en commun). Dans ce cas les deux significations se recouvrent, et w a un coefficient de regroupement proche de 1.

À l'inverse, il se peut aussi que les mots appartenant aux deux significations s_1 et s_2 soient différents la plupart du temps et ne soient pas connectés entre eux (par exemple les synonymes de « glace » pour sa signification « verre » ont très peu de mots en commun avec ses synonymes pour la signification « confiserie froide et sucrée ») ; c'est le cas si les synonymes de w pour la signification s_1 ne sont *pas* synonymes des synonymes de w pour la signification s_2. Alors w aura un coefficient de regroupement plutôt proche de 0.

36. Une *clique* est un ensemble de nœuds tous connectés entre eux.

Enfin, il se peut que les mots appartenant aux deux significations s_1 et s_2 soient différents la plupart du temps mais soient *tout de même connectés entre eux*, parce que ces mots sont synonymes pour des significations tierces (autres que s_1 et s_2). Dans ce cas, les voisins de w seront très connectés entre eux et w aura un coefficient de regroupement proche de 1.

Ces trois cas stylisés rendent compte des deux facteurs qui influent sur le coefficient de regroupement d'un mot w : le niveau de recouvrement des différentes significations de w, et le fait que les synonymes de w pour ses différentes significations aient d'autres significations, tierces, en commun (auquel cas ils sont connectés entre eux, mais pas à cause des significations qu'ils ont en commun avec w). C'est ici que l'interprétation du coefficient de regroupement qu'on a donnée dans le corps du rapport prend sens : en combinant ces deux facteurs on remarque qu'un mot aura un fort coefficient de regroupement lorsque les significations auxquelles ce mot appartient sont sémantiquement liées (en effet, l'idée qu'on peut se faire de la notion de lien sémantique entre deux significations correspond bien à avoir soit des lemmes en commun, soit des lemmes qui sont synonymes pour des significations tierces).

A.3 Recettes de filtrage

La partie 2.2.3 (page 26) dans le corps du rapport mentionne l'utilisation de plusieurs « recettes » de filtrage pour améliorer la détection des substitutions ; on va les détailler ici. On définit d'abord une notion supplémentaire dont on aura besoin dans la suite :

Définition 9. Distance de Levenshtein : On appelle *distance de Levenshtein entre deux listes* l_1 et l_2 de longueurs quelconques le nombre nombre minimum d'opérations du type *insertion, suppression, substitution* qui permet de transformer l_1 en l_2. Cette définition définit bien de manière unique une distance au sens mathématique du terme. On la note $d_L(l_1, l_2)$.

Par exemple, la distance de Levenshtein entre les mots « palme » et « charme » (pris comme des listes de caractères) est 3 : on substitue « p » en « c », on insère « h », et on substitue « l » en « r ».

Méthodes de filtrage Rappelons les étapes suivies avant d'ajouter les méthodes de filtrage. Jusque là la procédure d'extraction des informations qu'on cherche est comme suit :

1. Construction du graphe de synonymes WordNet et calcul des caractéristiques sémantiques pour les nœuds
2. Filtrage des clusters de la base de données
3. Itération à travers toutes les substitutions détectées par les différents modèles, pour enregistrer les caractéristiques sémantiques des mots échangés dans les substitutions, ainsi que leurs catégories grammaticales

On a inséré dans ce protocole plusieurs méthodes de filtrage supplémentaires, testées en examinant manuellement l'effet de chacune sur quelques dizaines de détections de substitutions. Une difficulté supplémentaire est que WordNet ne répertorie pas chaque mot sous toutes ses formes ou toutes ses orthographes, et pour chaque mot observé en substitution il faut pouvoir passer de la forme observée à la forme que WordNet connaît pour pouvoir l'inclure dans les résultats. Une deuxième série de méthodes a été testée pour pallier ce problème, en examinant les effets manuellement. Voici une synthèse des principales tentatives, indiquant celles qui ont été finalement retenues :

- Tous les mots dans les dictionnaires de caractéristiques sémantiques sont en minuscules ; on commence donc par convertir les mots d'une substitution en minuscules si ce n'est pas déjà le cas.
- On a testé plusieurs façons d'extraire à partir d'un mot la forme que WordNet connaît :
  - Lemmatisation en utilisant le lemmatiseur WordNet intégré dans la bibliothèque NLTK ; celui-ci ne donnait pas de résultats très probants, et de nombreuses substitutions détectées n'étaient pas intégrées aux résultats parce que les mots impliqués n'étaient pas retrouvés dans les dictionnaires de caractéristiques sémantiques.
  - Lemmatisation en utilisant la fonction `morphy` de WordNet, qui est la fonction sous-tendant la recherche de synsets à partir d'un mot ; cette méthode a donné de bien meilleurs résultats.
  - Pour les verbes conjugués on a utilisé un sous-produit de l'étiquetage des citations par TreeTagger, qui est la forme infinitive des formes conjuguées ; ceci a permis d'inclure un nombre bien plus large de substitutions impliquant des verbes, étant donné que ceux-ci sont à l'infinitif dans WordNet.
- Un certain nombre de substitutions détectées correspondaient en réalité à deux sous-parties de citations différentes, dans lesquelles le mot substitué n'a pas du tout la même fonction que le nouveau mot ; on a tenté de filtrer ces faux positifs en ne gardant que les substitutions pour lesquelles le mot substitué et le nouveau mot étaient des synonymes. Mais ceci a éliminé plus de la moitié des substitutions détectées (c'est-à-dire bien plus que ces faux positifs), et on s'est rendu compte a posteriori que bon nombre des substitutions pour lesquelles les mots impliqués ne sont pas des synonymes stricts (i.e. à distance 1 dans le graphe des synonymes) sont parfaitement légitimes. Enfin, se restreindre à ces substitutions synonymes aurait largement affaibli les résultats puisque la plupart des synonymes ont des caractéristiques sémantiques très proches [37]. On a donc exclu cette méthode.
- Il se trouve que la catégorie grammaticale des mots échangés lors d'une substitution est un meilleur indicateur pour filtrer les faux-positifs mentionnés au point précédent :

37. Étant donné que le graphe des synonymes qu'on a construit est composé essentiellement de cliques (chaque synset génère une clique), les mots synonymes ont souvent beaucoup de voisins en commun et ont donc des caractéristiques sémantiques proches, sauf si un des deux mots a beaucoup plus de significations que l'autre : celui-ci est alors lié à tous les autres synonymes de ses autres significations, ce qui lui permet d'être plus indépendant de son premier synonyme, en termes de caractéristiques sémantiques. Se restreindre aux substitutions synonymes implique de perdre bon nombre de substitutions légitimes pour lesquelles ces caractéristiques varient beaucoup, alors que ce sont justement les substitutions cruciales à ne pas manquer.

après examen des effets, on a décidé de ne garder que les substitutions où les deux mots impliqués avaient la même catégorie grammaticale. (Dans le cas où l'analyse était faite pour une catégorie grammaticale particulière cette étape n'était pas nécessaire, étant donné qu'on ne garde que les substitutions pour lesquelles les deux mots appartiennent à la catégorie qu'on cherche.)

- Enfin, après toutes ces étapes il reste un nombre significatif de substitutions qui ne sont qu'un changement entre orthographes anglaise et américaine [38], ou un changement entre singulier et pluriel. On a donc éliminé les substitutions qui, après les filtrages précédents, étaient composées de mots à distance de Levenshtein inférieure ou égale à 1. Ceci permet d'éliminer les substitutions pour lesquelles le filtrage précédent a déterminé que les mots racines étaient les mêmes (distance nulle), ainsi que la plupart des changements d'orthographe U.K. / U.S. et des changements singulier / pluriel (distance 1).

On sait que malgré ce filtrage, certains types de substitutions détectées seront des faux positifs : par exemple les changements d'entités nommées (indétectable puisque les données de base sont en minuscules), ou les apparitions ou disparitions de contractions. Par exemple « I will » → « I'll » est compté comme substitution puisque « I'll » est découpé en « I » et « 'll » (donc la distance de Hamming-tokens entre deux citations qui n'ont que cette différence est 1, et la distance de Levenshtein entre « will » et « 'll » est 2) ; cette substitution ne sera pas incluse car « 'll » ne figure pas dans WordNet, mais d'autres substitutions du même type pourraient se faufiler entre les mailles du filet [39]. Malheureusement, il n'y a pas de moyen autre que manuel pour vérifier la qualité de la détection et du filtrage des substitutions.

38. On rappelle que de nombreux mots ont, aux États-Unis, une orthographe différente de celle utilisée dans la plupart des pays du Commonwealth : par exemple « neighbour » (U.K.) / « neighbor » (U.S.), « centre » (U.K.) / « center » (U.S.), ou encore tous les mots en « -ise » (et leurs dérivés) s'écrivent « -ize » aux États-Unis. Voir http://en.wikipedia.org/wiki/American_spelling pour une liste complète des différences.

39. On peut noter cependant que la transformation « does not » → « doesn't » est filtrée : « doesn't » est découpé en « does » et « n't » ce qui donne une distance de Hamming-tokens 1 entre les deux citations, mais la distance de Levenshtein entre « not » et « n't » est 1, ce qui exclut la substitution.

Références

Scott Atran. Théorie cognitive de la culture. *L'Homme*, 166(2) :107–143, 2003.

Robert Aunger, Susan Blackmore, Maurice Bloch, Robert Boyd, Rosaria Conte, David L. Hull, Adam Kuper, Kevin Laland, John Odling-Smee, Henry Plotkin, Peter J. Richerson, and Dan Sperber. *Darwinizing Culture : The Status of Memetics as a Science*. Oxford University Press, 2000.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The Role of Social Networks in Information Diffusion. In *Proceedings of ACM WWW 2012*, 2012.

Maurice Bloch. A well-disposed social anthropologist's problems with memes. In Robert Aunger, editor, *Darwinizing Culture : The Status of Memetics as a Science*, chapter 10, pages 189–203. Oxford University Press, 2000.

Javier Borge-Holthoefer and Alex Arenas. Semantic Networks : Structure and Dynamics. *Entropy*, 12(5) :1264–1302, 2010. doi : {10.3390/e12051264}.

Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, 1985.

Kit Ying Chan and Michael S Vitevitch. Network structure influences speech production. *Cogn Sci*, 34(4) :685–97, 2010. doi : {10.1111/j.1551-6709.2010.01100.x}.

Nicolas Claidière and Dan Sperber. The role of cultural attraction in cultural evolution. *Journal of Cognition and Culture*, (7) :89–111, 2007. doi : {10.1163/156853707X171829}.

Jean-Philippe Cointet and Camille Roth. Socio-semantic Dynamics in a Blog Network. In *2009 International Conference on Computational Science and Engineering*, pages 114–121, 2009. doi : {10.1109/CSE.2009.105}.

Richard Dawkins. *The Selfish Gene*, chapter 11, pages 189–201. Oxford University Press, 1976. « Memes : The New Replicator ».

Paul R Ehrlich and Simon A Levin. The evolution of norms. *PLoS Biol.*, 3(6) :e194, 2005. doi : {10.1371/journal.pbio.0030194}.

Thomas L Griffiths, Mark Steyvers, and Alana Firl. Google and the mind : predicting fluency with PageRank. *Psychol Sci*, 18(12) :1069–76, 2007. doi : {10.1111/j.1467-9280.2007.02027.x}.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information Diffusion Through Blogspace. In *Proceedings of the 13th International World Wide Web Conference (WWW'04)*, pages 491–501, 2004.

David Jurgens and Tsai-Ching Lu. Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2012.

Adam Kuper. If memes are the answer, what is the question ? In Robert Aunger, editor, *Darwinizing Culture : The Status of Memetics as a Science*, pages 180–193. Oxford University Press, 2000.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the Dynamics of the News Cycle. *KKD'09*, (June 28-July 1) :497–505, 2009a.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. MemeTracker : tracking news phrase over the web. http://memetracker.org/, 2009b. Visité le 19.08.2012.

Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163) :713–6, 2007. doi : {10.1038/nature06137}.

NLTK. Natural Language Processing Toolkit. http://nltk.org/. Visité le 19.08.2012.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking : Bringing Order to the Web. Technical Report 1999-66, November 1999. URL http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120.

Sylvain Parasie and Jean-Philippe Cointet. La presse en ligne au service de la démocratie locale. *Revue française de science politique*, 62(1) :45–70, 2012. doi : {10.3917/rfsp.621.0045}.

Laurent Pointal. TreeTaggerWrapper. http://perso.limsi.fr/Individu/pointal/doku.php?id=dev:treetaggerwrapper. Visité le 19.08.2012 ; révision 20.

Beatrice Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*, 3rd revision, 2nd printing edition, 1990.

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland, 1995.

Matthew Simmons, Lada Adamic, and Eytan Adar. Memes Online : Extracted, Subtracted, Injected, and Recollected. In Nicolas Nicolov and James G. Shanahan, editors, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2011.

Dan Sperber. *La Contagion des Idées*. Paris : Odile Jacob, 1996.

Angela K. Troyer, Morris Moscovitch, and Gordon Winocur. Clustering and switching as two components of verbal fluency : Evidence from younger and older healthy adults. *Neuropsychology*, 11(1) :138–146, 1997. doi : {10.1037/0894-4105.11.1.138}.

WordNet. Princeton University « About WordNet. ». http://wordnet.princeton.edu, 2010. Visité le 19.08.2012.