Revue Sciences/Lettres

2 | 2014
Épistémologies digitales des sciences humaines et sociales

Modélisation multiniveau de la morphogenèse de familles de citations

Elisa Omodei et Jean-Philippe Cointet

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/rsl/510
DOI : 10.4000/rsl.510
ISSN : 2271-6246

Éditeur
Éditions Rue d'Ulm

Référence électronique
Elisa Omodei et Jean-Philippe Cointet, « Modélisation multiniveau de la morphogenèse de familles de citations », *Revue Sciences/Lettres* [En ligne], 2 | 2014, mis en ligne le 07 octobre 2013, consulté le 02 mai 2019. URL : http://journals.openedition.org/rsl/510 ; DOI : 10.4000/rsl.510

Modélisation multiniveau de la morphogenèse de familles de citations

Elisa Omodei et Jean-Philippe Cointet

Introduction

1 Rihard Dawkins, a introduit le concept de « mème » en 1976 pour défendre l'idée selon laquelle la culture serait soumise à des processus évolutifs au même titre que les être vivants soumis à « l'égoïsme » de leur gène (Dawkins, 2006). Cette hypothèse a suscité de fortes oppositions, notamment de la part d'ethnographes et d'anthropologues (Sperber, 1996), qui critiquent en particulier le fait d'assimiler la culture à des objets appelés « entités culturelles » (Aunger, 2003) et plus généralement l'absence de validation empirique permettant d'appuyer cette théorie (Edmonds, 2002). De ce point de vue, les médias sociaux offrent un terrain d'expérimentation propice pour étudier les dynamiques socioculturelles *in vivo*. Cet article emprunte cette orientation empirique en étudiant les transformations que subissent des « citations » dans la blogosphère. Une citation est simplement définie comme un extrait rapporté qui peut provenir d'une déclaration publique ou d'une autre source textuelle comme un article de presse. Pratiquement, on les définit comme tout bloc de texte entouré de guillemets. Même si les citations sont loin d'être représentatives de la complexité des objets culturels, l'étude de leur structure et de leur évolution est une opportunité pour s'aventurer sur les traces du programme de recherche imaginé par Dawkins.

2 Nous nous sommes donc intéressés à la diffusion de citations dans les médias sociaux (forums, blogs, presse) touchant au domaine politique sur le web. Ce type d'étude fait directement référence à la notion d'« intertextualité », terme fréquemment utilisé dans le cadre de l'analyse de discours. L'intertextualité fait référence à la reprise, à la répétition et à la modification de fragments de discours d'un texte à l'autre au cours du temps. En nous fondant sur les idées de Julia Kristeva (Kristeva, 1966), nous supposons que ces

distorsions ne sont pas neutres et reflètent la façon dont les idées se diffusent, dans différents médias et différentes communautés.

3 Plus généralement, les études sur les processus de diffusion ont suscité beaucoup d'intérêt chez les sociologues (Rogers, 1976), depuis les travaux fondateurs de Coleman sur la diffusion des médicaments chez les médecins (Coleman *et al.*, 1957), jusqu'à des études plus récentes sur les processus de diffusion dans les médias en ligne s'attachant, entre autres, aux recommandations de livres et de DVD, aux cascades de citations ou aux reprises d'URL d'un texte à l'autre (Leskovec *et al.*, 2007 ; Adar *et al.*, 2004).

4 Le processus en jeu dans chacun de ces exemples suppose que les objets sont parfaitement reproduits, leur stabilité étant une contrainte expérimentale pour les repérer et ainsi en étudier leur diffusion. Comme les éléments mentionnés ci-dessus, un ensemble de citations donné, même s'il peut subir des variations au cours du temps, réfère toujours à un même événement particulier externe. Cependant, et ceci est propre aux citations, les auteurs peuvent choisir de modifier plus ou moins sensiblement le texte de la citation, ce qui ouvre la voie à une analyse systématique des motifs qui sous-tendent ces modifications.

5 À ce stade, il peut être utile de revenir rapidement à notre métaphore. En biologie, une séquence de gènes peut être modifiée par des mutations qui peuvent affecter un seul nucléotide ou une longue séquence de nucléotides. Nous pensons que ce type d'analyse, prenant en compte des modifications locales et d'autres plus globales, est aussi valable dans le cas des citations. Nous reprenons donc ce type de distinction, à la différence de Simmons *et al.* (2011) qui parlent de recadrage *(reframing)* et de modification. Pour Simmons *et al.*, il y a recadrage quand une citation A est un fragment d'une citation B (une citation exacte mais partielle), tout autre changement étant qualifié de modification. Nous préférons quant à nous conserver la dichotomie employée en biologie distinguant les micromutations (transformations ponctuelles [concernant un seul nucléotide] pouvant prendre la forme d'une suppression, d'une insertion ou d'un remplacement) des micromutations (qui affectent toute une séquence de nucléotides).

6 Nous faisons l'hypothèse que les micromutations sont des altérations mineures de la citation originale qui peuvent être introduites volontairement ou plus probablement par erreur, sans intention particulière quant au sens original de la citation reproduite. Les macromutations sont quant à elles plus probablement dues à des changements volontaires de blogueurs ou de journalistes qui veulent par exemple attirer l'attention du lecteur vers une sous-partie du texte original cité.

7 Notre objectif est alors de décrire comment les changements d'ordre microscopiques ou macroscopiques transforment progressivement les citations au cours de leur diffusion. La première partie de l'article sera consacrée à une très courte description du jeu de données utilisé. Un algorithme original pour détecter les familles de citations est alors introduit. Sur la base de ces familles, nous mesurons dans la quatrième partie des indices de stabilité et de diversité à différents niveaux (au niveau des mots, des citations et des familles). Dans la dernière partie nous nous penchons sur la morphogenèse des familles de citations en introduisant un modèle dynamique qui reproduit en partie les phénomènes étudiés.

Description du jeu de données

8 Nous avons choisi d'analyser le corpus MemeTracker (construit par Leskovec *et al.*, 2009) qui contient un ensemble de citations automatiquement extraites de 90 millions d'articles de presse et de blog, et recueillies au cours des trois mois précédant et suivant l'élection présidentielle américaine de 2008 (soit 310 457 citations recueillies auprès de différents articles de presse et de blog, de début août 2008 jusqu'à la fin janvier 2009). Seules les citations mentionnées au moins cinq fois ont été conservées par les producteurs du corpus. Comme nous nous concentrons ici uniquement sur les transformations qu'ont subies ces citations en fonction de caractéristiques endogènes (aux phrases ou à la dynamique de diffusion), nous ne tenons pas compte du réseau social sous-jacent ni des liens hypertextes mentionnés, même s'il s'agit probablement également d'une dimension capitale pour comprendre la dynamique de diffusion.

Repérage des familles de citation

9 Afin d'analyser le corpus MemeTracker, il est nécessaire d'identifier les *familles de citations* , c'est-à-dire regrouper les différentes citations en relation avec une citation initiale jouant le rôle de modèle originel (ce modèle pouvant ensuite être tronqué, modifié ou reproduit tel quel).

10 Cette analyse se fait en trois étapes : (i) toutes les citations sont linguistiquement analysées et normalisées (par lemmatisation puis suppression des mots vides) ; (ii) la proximité entre citations est calculée et les citations dont la similarité est au-dessus d'un seuil donné sont liées de manière à obtenir un « graphe de citations » ; (iii) un algorithme de classification automatique *(clustering)* est utilisé pour détecter les communautés (c'est-à-dire les sous-graphes les plus fortement interconnectés dans le graphe), ce que nous appelons des « familles de citation ». Nous détaillons ce processus dans le paragraphe suivant ; nous proposons ensuite une évaluation et une discussion des résultats obtenus.

Une approche hybride (linguistique et structurelle) pour l'analyse des citations

11 Alors que la méthode définie par Leskovec et ses collègues pour identifier les familles de citations s'appuie avant tout sur une analyse des relations d'inclusion entre citations (Leskovec *et al.*, 2009), nous avons essayé de concevoir une mesure de proximité entre citations prenant davantage en compte le contenu textuel grâce à une analyse linguistique de surface.

12 Un domaine en relation direct avec celui qui nous intéresse ici est la détection de paraphrases, qui a été beaucoup exploré récemment en analyse du langage naturel pour des applications comme l'extraction d'information, le résumé automatique et la traduction automatique. La détection de paraphrases est une tâche difficile car elle implique théoriquement une analyse à la fois sémantique et syntaxique des phrases visées. Cependant, dans la pratique, la plupart des approches se fondent sur l'identification de mots similaires entre couples de phrases, ce qui permet de calculer assez simplement une valeur de similarité entre phrases (Mihalcea *et al.*, 2006). Diverses

améliorations ont été envisagées afin d'obtenir des résultats plus précis, comme l'introduction d'une notion de similarité entre mots (en utilisant par exemple une ressource comme Wordnet pour l'anglais) ou l'identification des relations entre mots. Par exemple, Qiu *et al.* (2006) utilisent l'analyseur syntaxique développé par Charniak pour obtenir une analyse syntaxique de la phrase et essayer de mettre au jour des ensembles prédicats-arguments (par exemple, un verbe et ses compléments) qui permettent de voir des équivalences sémantiques entre phrases au-delà des variations de surface.

13 En ce qui nous concerne, les traitements que nous avons appliqués au texte sont relativement simples par rapport à l'état de l'art mais ils sont *a priori* suffisants pour analyser efficacement de grandes quantités de texte en un temps raisonnable. Ils visent à repérer les éléments de sens essentiels afin de servir de base à une comparaison entre phrases qui soit à la fois robuste sur le plan technique et précise sur le plan sémantique. Nous procédons en trois grandes étapes :

1. Nous avons d'abord remplacé chaque mot par son lemme en utilisant le logiciel TreeTagger (Schmid, 1994), puis nous avons éliminé les mots sémantiquement vides comme les articles et les prépositions. Cette étape permet de ne conserver que les éléments sémantiquement signifiant pour le calcul de proximité entre phrases.
2. Nous bâtissons ensuite un graphe en reliant entre elles les citations les plus proches, ce qui implique de définir un moyen de mesurer la proximité entre citations. Une stratégie classique consisterait par exemple de procéder à une comparaison du nombre de mots communs entre deux phrases, mais cette méthode ne tiendrait pas compte du tout de l'ordre des mots, ce qui semble un peu sommaire.
   Afin de remédier à cet écueil, nous utilisons une version modifiée de la distance de Levenshtein, qui est normalement destinée à calculer la distance entre deux chaînes de caractères. D'une manière générale, deux phrases peuvent bien évidemment avoir un sens proche même si l'ordre des mots employé est différent (par exemple si l'on compare une phrase active avec une phrase passive) mais ceci n'est pas vraiment pertinent dans le cas des citations, dans la mesure où les modifications possibles affectent généralement peu l'ordre des mots. Enfin, dans le calcul de proximité entre phrases, les mots sont pondérés en fonction de leur pouvoir discriminant (en utilisant la mesure classique du tf.idf [Salton et McGill, 1983], qui prend en compte la fréquence d'un terme et sa répartition dans le corpus : un mot est d'autant plus discriminant qu'il est fréquent et concentré dans un petit nombre de documents). Une fois les calculs de proximité terminés, les phrases dont la proximité est au-dessus d'un seuil fixé *a priori* sont liées entre elles pour former le graphe visé.
3. La dernière étape de l'analyse consiste à appliquer un algorithme de classification *(clustering)* pour la détection des communautés dans le graphe, c'est-à-dire identifier les différentes familles de citations. À cet effet, nous avons choisi l'algorithme Infomap, élaboré par Rosvall et Bergstrom (2008). Il s'agit d'un algorithme basé sur la théorie de l'information, qui utilise la probabilité du flux des marches aléatoires sur un réseau comme modèle du flux d'information transitant dans le système réel et décompose le graphe en communautés en proposant la description la plus parcimonieuse de ces flux. Nous avons choisi cet algorithme en nous basant sur l'étude de Lancichinetti et Fortunato (2009), qui ont évalué différents algorithmes de détection de communautés et ont constaté qu'Infomap obtient de bonnes performances tout en ayant une faible complexité de calcul, ce qui rend cet algorithme particulièrement pertinent pour analyser de grandes masses de données comme c'est le cas ici.

Évaluation des résultats et comparaison avec la méthode de Leskovec

14 Les méthodes de classification sont largement utilisées en traitement automatique du langage naturel quand il s'agit d'opérer des regroupements de quelque nature que ce soit. Toutefois, l'évaluation de ces méthodes est difficile, dans la mesure où l'on dispose rarement de jeux de données de référence et que, de surcroît, différentes classifications peuvent avoir un sens différent en fonction du domaine et de la tâche.

15 Dans notre cas, aucune classification de référence n'est disponible mais on peut malgré tout se baser sur les résultats du projet MemeTracker pour obtenir un point de départ. Nous avons alors choisi d'évaluer manuellement un ensemble de familles (ou regroupements ou clusters) choisis au hasard parmi les familles produites par les deux méthodes (la nôtre et celle de Leskovec *et al.*). L'évaluation a porté sur trente familles de citations. On procède en deux étapes :

16 Soit une famille F_1 produite par la méthode M_1 et composée de *n* phrases. La qualité de la famille est évaluée en estimant manuellement, pour chaque phrase de F_1, si elle doit faire partie de la famille ou non (*i.e.* si elle est en situation de paraphrase avec les autres). Ceci permet d'avoir une idée de la précision de la famille (la *precision* mesure la proportion de citations pertinentes parmi celles qui ont été trouvées par la méthode M_1).

17 Ensuite toutes les phrases contenues dans les familles produites par la méthode M_2 contenant au moins une phrase de F_1 sont ajoutées à la famille et évaluées. Ceci permet d'avoir une idée du rappel de la méthode (proportion des citations trouvées parmi celles identifiées par la méthode alternative).

18 Nous procédons d'abord avec trente familles produites par notre système puis, en retour, avec trente familles produites par la méthode de Leskovec *et al.*

19 Avant de détailler le résultat de cette évaluation, il est nécessaire d'examiner rapidement certaines questions de méthode. Tout d'abord, un certain nombre de fragments de texte ne sont pas de réelles citations, mais des titres *(High School Musical)*, des expressions courtes sans signification claire *(a bit)* et des mots étrangers *(La Vida ne vale nada)*. Les familles correspondant ont été exclues de l'évaluation. Il fallait ensuite que l'évaluateur détermine le sujet *(topic)*, c'est-à-dire la phrase de référence pour la recherche de paraphrase au sein de la famille. Enfin, les évaluateurs avaient pour instruction de marquer comme équivalentes des phrases qui partagent des informations identiques essentielles mais il est évidemment difficile de déterminer ce qui est une information essentielle par opposition à ce qui est une information secondaire.

20 Malgré le peu d'instructions données à nos évaluateurs (les informations données étaient volontairement minimales pour ne pas biaiser leurs jugements), nous avons obtenu des résultats intéressants et fiables. L'accord entre annotateurs (mesuré au moyen du kappa de Cohen) est élevé (0,69), surtout quand on prend en compte la subjectivité relative de la tâche et le peu d'instructions données. Nous avons obtenu les résultats suivants en termes de précision et de rappel relatif (le rappel est dit relatif car il est fondé sur la comparaison entre deux méthodes automatiques et non sur une référence établie manuellement).

Méthode de *clustering*	**Précision**	**Rappel relatif**	**F-mesure**

Notre méthode	.58	.90	.70
Leskovec *et al.*	.47	.78	.58

21 Notre méthode obtient de meilleurs résultats que Leskovec *et al.*, pour la précision comme pour le rappel relatif. L'amélioration de la précision est probablement due à l'étape de prétraitement linguistique qui rend l'ensemble du processus plus précis (notre analyse se concentre sur les mots sémantiquement pleins, qui sont eux-mêmes pondérés en fonction de leur pouvoir discriminant). L'amélioration du rappel que nous constatons avec la méthode que nous avons développée est probablement due au fait que les familles générées par MemeTracker contiennent beaucoup de fragments de très petite taille, que les juges ont du coup écartés car ne rendant pas compte du contenu fondamental de la citation.

22 Nous avons enfin fait une mesure de significativité des résultats grâce au test SIGF V2 (Padó, 2006), qui se base sur un cadre de répartition aléatoire sans hypothèses. Le résultat obtenu est 0,049, ce qui signifie que notre résultat est consistant et peut être considéré comme fiable.

Filtrage

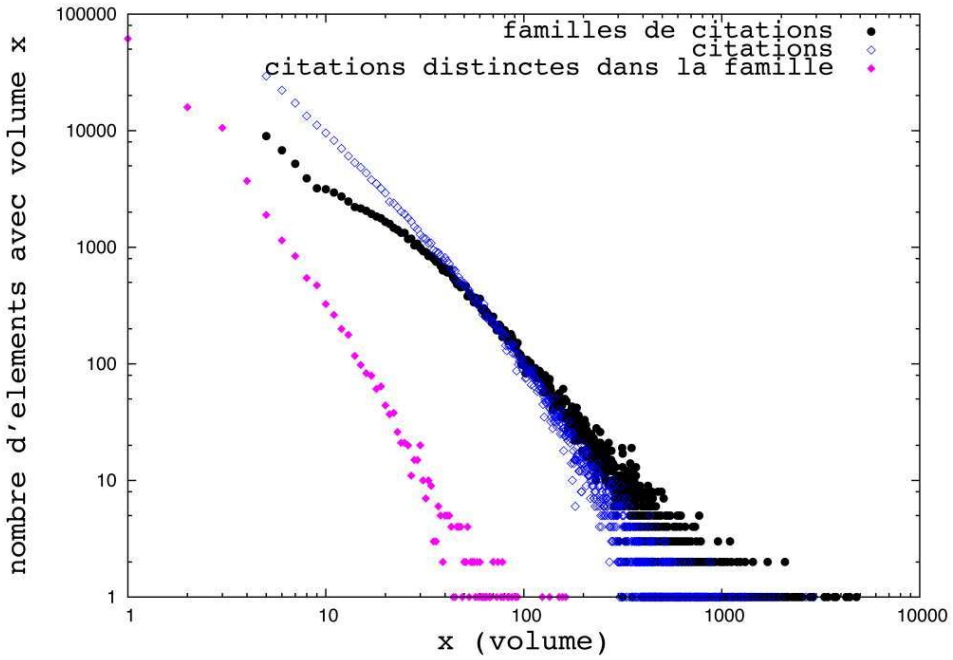
23 La base de données contient de nombreuses citations qui ne sont pas en anglais ou sont trop courtes pour pouvoir être singularisées. Nous avons alors décidé de « filtrer » nos résultats en considérant seulement les citations en anglais contenant au moins cinq mots.

Description des familles de citations

24 Nous représentons la distribution de la taille des familles (c'est-à-dire le nombre total des citations dans une famille de citations données), la distribution du nombre de citations distinctes dans une famille et la distribution du nombre de mentions par citation (fig. 1).

25 La structure des trois distributions ressemble à une loi de puissance et est comparable aux distributions trouvées dans Leskovec *et al.* (2009), bien que les familles aient été définies différemment.

Figure 1



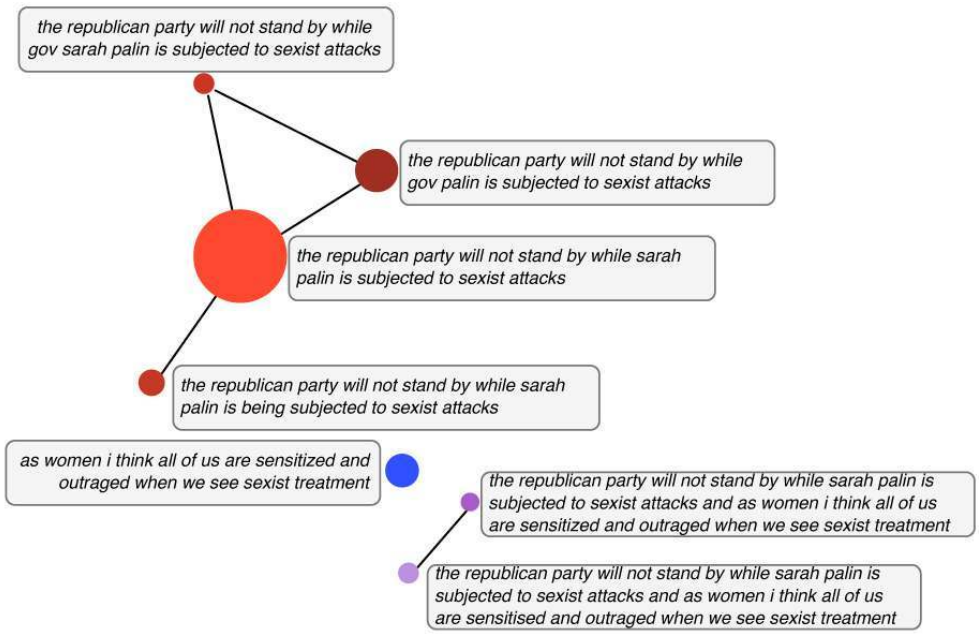
Analyse multiniveau des transformations

26 Observons d'abord une famille composée de sept citations, d'après MemeTracker. Le 3 septembre 2008, Carly Fiorina, ex-patronne du géant de l'informatique Hewlett-Packard, a déclaré à une conférence de presse : « Le parti républicain ne va pas rester à ne rien faire alors que Sarah Palin est l'objet d'attaques sexistes, et en tant que femme, je pense que chacun devrait être indigné face à des traitements sexistes. » La citation est reproduite seize fois dans le jeu de données de référence. Mais on trouve simultanément au moins six formes modifiées de cette citation, la plus fréquente étant aussi beaucoup plus courte que l'originale : « Le parti républicain ne va pas rester à ne rien faire alors que Sarah Palin est l'objet d'attaques sexistes » (56 occurrences).

27 Si on regarde de plus près la répartition des citations au sein de la famille en question, trois sous-familles se dégagent clairement, reprenant seulement la première partie de la citation, seulement la seconde ou bien la citation dans son entier.

28 Au sein de chaque sous-famille, on observe également de petites variations, du fait de variantes orthographiques *(sensitised/sensitized)* et de mots coupés ou ajoutés *(Sarah Palin/gov Palin/gov Sarah Palin, is being subjected/is subjected)*. Les variations au niveau des mots sont appelées des microtransformations, tandis que les changements au niveau de séquences plus importantes de la citation sont appelées macrotransformations.

Figure 2



Variation au niveau des mots

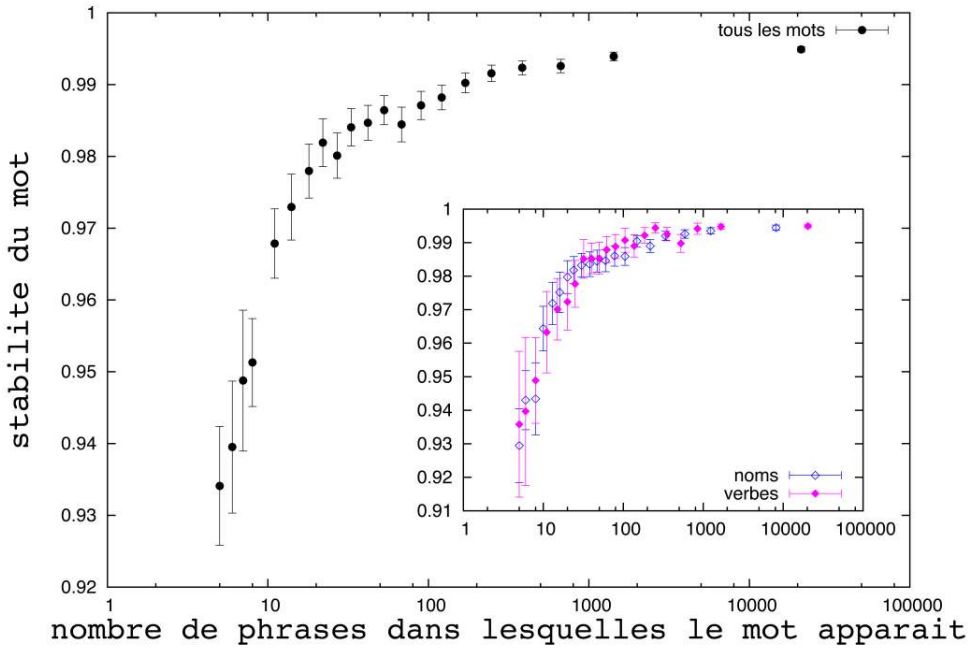
29 La variation au niveau des mots laisse entrevoir des alternances orthographiques classiques en anglais. On observe ainsi les alternances suivantes : *defense/defence, program/programme* ou encore *behaviour/behavior*. D'autres variations concernent les mots avec un tiret *(cease-fire/ceasefire)*, les abréviations *(gov / gouverneur)* et des mots étrangers *(al-qaeda/al-qaida)*. Enfin, les mots d'argot sont souvent omis (on trouve par exemple *fuck* parmi les 20 mots les plus instables du corpus ; le mot est souvent supprimé ou remplacé par un simple « f » entre guillemets).

30 Outre ces observations plutôt anecdotiques, des schémas de variation plus systématiques apparaissent. Ainsi, la figure 3 montre que la fréquence des mots est corrélée avec leur stabilité. Les mots les plus fréquents sont aussi les plus stables : les termes apparaissant plus de cent fois dans le corpus ont une stabilité supérieure à 99 %, cette valeur pouvant même atteindre 99,5 % pour les plus fréquents. À l'inverse, les mots rares ont une stabilité beaucoup plus faible. Ce constat se vérifie même en faisant abstraction des mots vides et quelle que soit la catégorie grammaticale observée. Cette observation est en accord avec le fait qu'un événement fréquent est plus facilement mémorisé qu'un événement rare (mais à l'inverse, un événement très rare peut marquer l'esprit, ce qui ne semble pas être le cas ici). Notons enfin que les mots rares sont parfois aussi ceux qui peuvent subir des variantes orthographiques (comme certains noms propres étrangers), ce qui peut contribuer à leur instabilité.

31 Le même type de résultat est observé dans les études sur l'évolution des langues. Par exemple, dans Lieberman (2007), les auteurs ont montré que le taux de régularisation des verbes irréguliers anglais diminue rapidement avec leur fréquence, ce qui indique que les verbes irréguliers rares sont soumis à plus d'erreurs, conduisant à leur "rapide" régularisation. En outre, la spécificité n'est pas synonyme de stabilité. Une étude récente

analysant la même base de données a essayé de classer les termes en fonction de leur généricité, sur la base du réseau sémantique Wordnet (où les mots sont classés hiérarchiquement, ce qui peut indiquer leur degré de généricité). Les auteurs de cette étude ont montré que les termes les plus spécifiques sont les plus susceptibles d'être remplacé, surtout par des termes plus génériques (Lerique et Roth, 2012). Cette « préférence naturelle » pour des termes génériques peut expliquer la forme de la courbe, car il est probable que les termes les plus spécifiques sont moins fréquents que d'autres plus polysémiques.

Figure 3 – Stabilité en fonction de la fréquence des termes. En médaillon : la stabilité des verbes et des noms seulement.

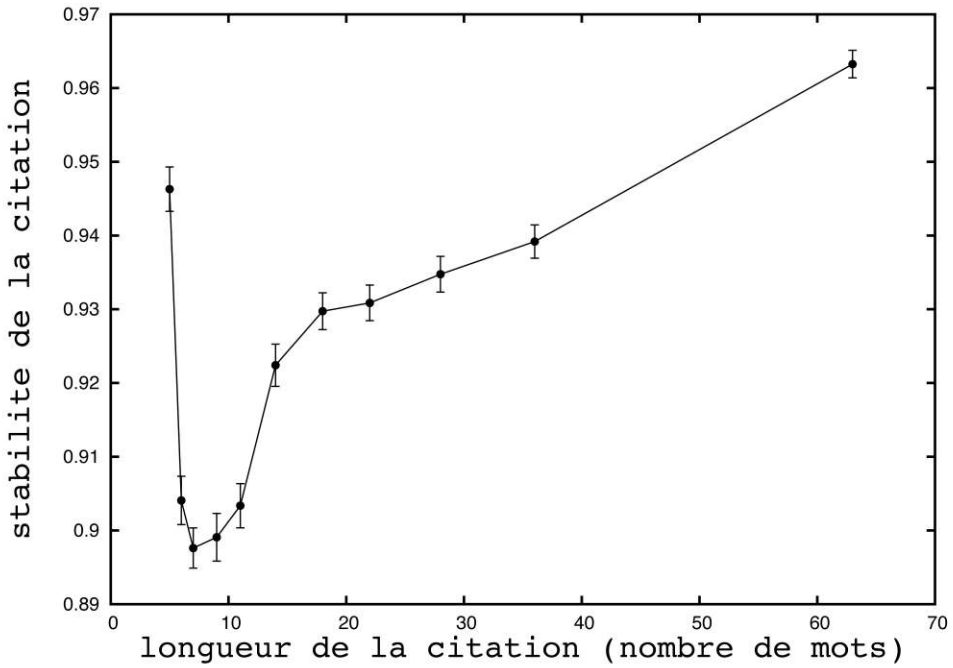


Variation au niveau des citations

32 Nous calculons un indice de stabilité des citations et faisons varier cet indice en fonction de la longueur des citations, c'est-à-dire le nombre de mots qu'elle contient. Les citations les moins stables sont celles qui comportent environ huit mots : les citations moins longues comme celles qui sont plus longues sont plus stables.

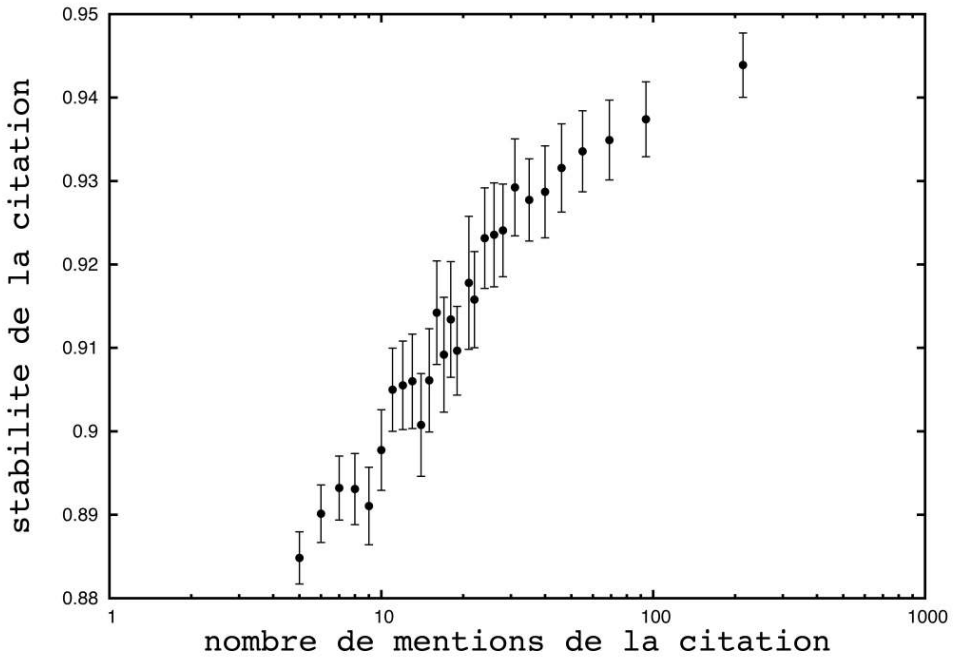
33 Deux processus sont en concurrence pour expliquer la « dissémination » des citations : les blogueurs peuvent soit rapporter de mémoire une citation vue ou entendue auparavant, soit ils procèdent par simple copier/coller. Le premier processus est davantage employé pour des citations courtes (au-delà de dix mots, il semble plus difficile de se souvenir d'une citation) mais il est aussi susceptible d'introduire des erreurs de recopie, ces erreurs étant d'autant plus probables que la citation est longue, ce qui peut expliquer le taux croissant d'instabilité jusqu'à environ huit mots. Au-delà d'une dizaine de mots, le second processus (copier/coller direct) s'applique de façon privilégiée, ce qui explique pour nous la grande stabilité des citations les plus longues (fig. 4).

Figure 4 – Stabilité des citations en fonction de leur longueur (nombre de mots).



34 Nous avons également cherché à savoir si la stabilité d'une citation était liée à son nombre d'occurrences. La figure 5 montre que la stabilité des citations croît rapidement en fonction de leur fréquence (l'abscisse est logarithmique). Deux processus peuvent expliquer cette stabilité accrue en fonction de la fréquence. D'un côté, les citations plus fréquentes sont peut-être plus stables car elles se dupliquent facilement. Par ailleurs, un autre facteur de stabilité peut être tout simplement le fait que les formes les plus fréquentes finissent par supplanter les autres. Les deux phénomènes peuvent agir simultanément : certaines citations sont intrinsèquement plus faciles à copier que d'autres, ce qui les rend plus fréquentes, donc dynamiquement plus populaires et plus stables, vu qu'une citation fréquente est peu susceptible de subir des variations.

Figure 5 – Stabilité des citations en fonction de leur nombre de mentions.

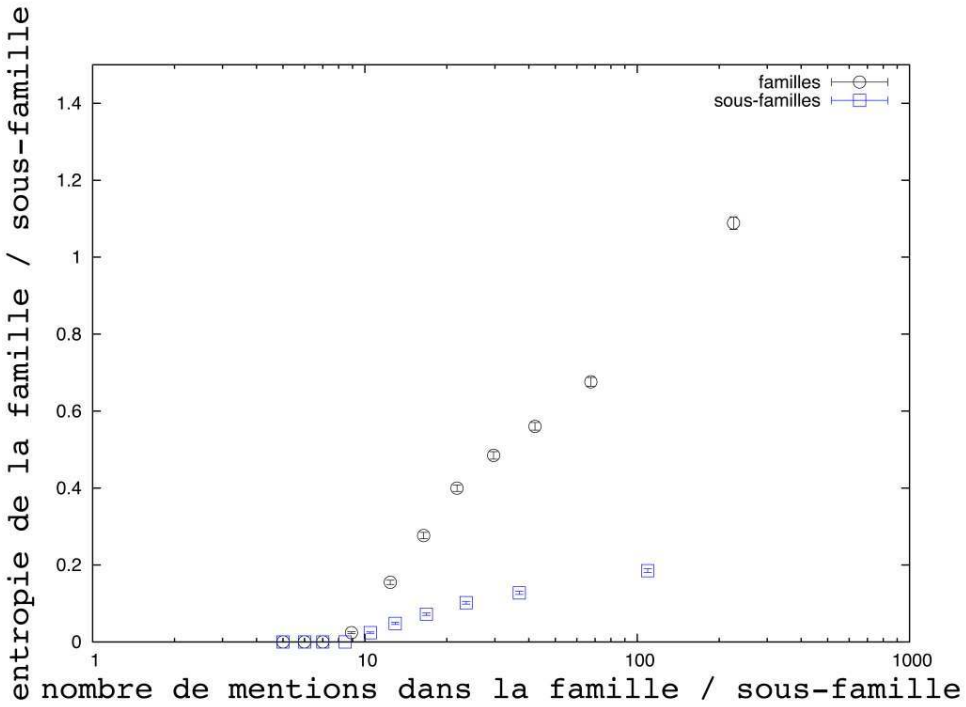


Variations au niveau des familles et des sous-familles

35 Pour mieux comprendre la composition des familles de citations et mieux caractériser leur diversité, nous avons mesuré leur entropie (plus exactement, nous avons mesuré l'entropie de Shannon des familles et des sous-familles). Nous rappelons que les familles regroupent toutes les citations liées à une citation d'origine, quelle que soit l'échelle des transformations que la citation d'origine peut avoir subies, tandis que les sous-familles regroupent les citations qui peuvent être liées par le biais de microvariations seulement. L'entropie de Shannon (1956), qui a été initialement appliquée à des séquences de lettres, est souvent utilisée comme un indice de diversité. Appliquée aux citations, l'entropie mesure la diversité des citations qui composent une famille ou une sous-famille.

36 L'entropie présente des profils très différents suivant que l'on s'intéresse aux sous-familles ou aux familles. La figure 6 montre comment l'entropie est corrélée avec la taille des familles et des sous-familles. Nous pouvons en particulier observer que la valeur de l'entropie pour les familles avec un certain nombre de mentions est toujours beaucoup plus élevée que la valeur de l'entropie des sous-familles du corpus regroupant le même nombre de mentions. En d'autres termes les sous-familles présentent moins de diversité que les familles, ce qui suggère qu'au « niveau micro » la concurrence entre les différentes versions de la même phrase conduit finalement à une situation dans laquelle une version « domine » en fréquence les versions concurrentes, alors qu'au « niveau macro », la coexistence de différentes sous-familles relativement indépendantes garantit une certaine diversité.

Figure 6 – Entropie des familles et sous-familles en fonction du nombre total de mentions des citations qui les composent.



Modéliser la morphogenèse des familles de citations

37 Nous introduisons maintenant un modèle de morphogenèse de familles et de leurs sous-familles qui permette aussi bien de reconstruire les distributions de tailles empiriques que les mesures de diversité réalisées à chaque échelle. Pour construire un modèle réaliste, nous devons d'abord analyser plus finement les transformations possibles qui altèrent les citations lors de leur diffusion. Dans cette section, nous examinerons donc plus précisément l'évolution des familles dans le temps, pour apprécier le volume et la composition des mutations.

Taux de mutation

38 Deux types de mutations peuvent « altérer » la copie d'une citation. Les mutations que l'on appellera « macro » consistent à ne sélectionner qu'une partie de la citation originale, en la privant de certains de ses éléments originaux. Par ailleurs, nous appellerons mutation « micro » les altérations qui se limitent à quelques mots (et dans la suite un seul). Dans le premier cas (mutation macro), on considère qu'une nouvelle sous-famille est créée, dans l'autre cas, que la sous-famille dont la phrase originale a été tirée est simplement enrichie d'une nouvelle version. Nous faisons l'hypothèse que les événements de mutations étant « rares », l'occurrence d'une citation déjà observée dans l'écosystème relève nécessairement d'une opération de copie sans altération plutôt que d'une mutation depuis une citation tierce.

39 Chaque nouvelle version résultant par hypothèse d'un événement de mutation (et inversement), nous pouvons très facilement mesurer les taux de mutation empiriques. Précisément, le taux de mutation est donné par le ratio entre le nombre d'événements de mutation (soit le nombre de « versions » différentes après avoir exclu la citation d'origine dans la famille [macromutation] ou dans la sous-famille [micromutation]) et le nombre total d'événements de réplication (soit le nombre total de mentions dans la famille [macro]/sous-famille [micro]) :

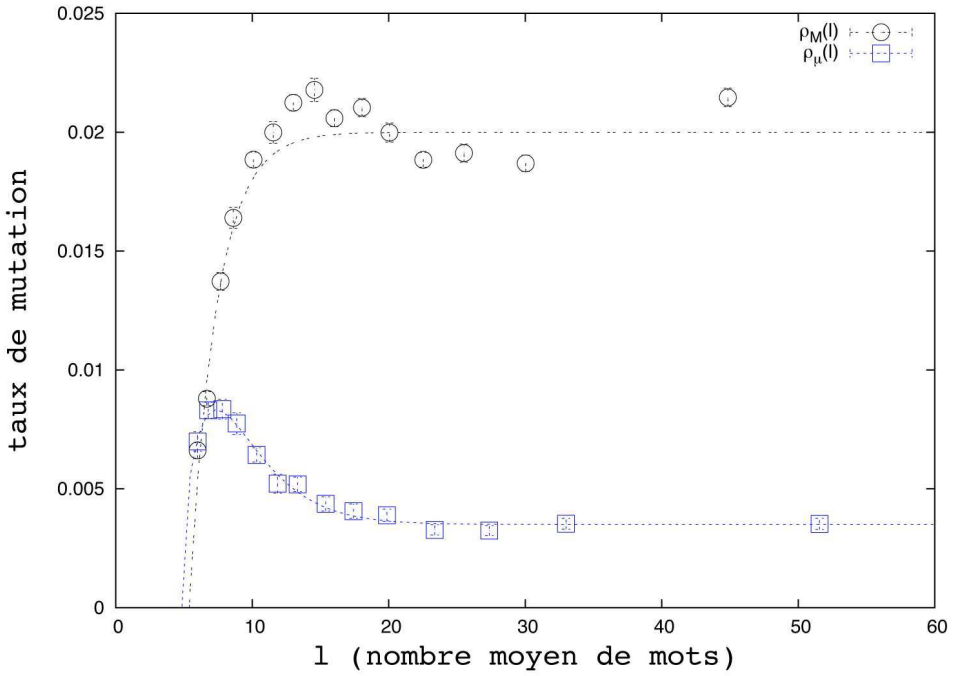
- Le taux de mutation macro est calculé pour chaque famille en divisant le nombre de sous-familles moins 1 par le nombre total de mentions moins 1.
- Le taux de mutation micro est évalué pour une sous-famille en calculant le nombre de versions distinctes dans une sous-famille moins 1 divisé par le nombre total de mentions dans la sous-famille moins 1.

40 Cependant, l'accès à la moyenne des taux de mutation n'est pas suffisant pour reconstruire de manière réaliste la morphogenèse des familles et des sous-familles : nous devons aussi prendre en compte les propriétés qui amplifient ou réduisent les taux de mutation. Dans la section précédente, nous avons montré que la stabilité des citations dépend de leur longueur et de leur popularité, ce qui suggère que le taux de mutation pourrait différer fortement en fonction de ces deux propriétés. En outre, ces propriétés dépendent elles-mêmes de façon critique des dynamiques de diffusion. C'est pour cette raison que nous devons mesurer dynamiquement la dépendance de la vitesse de mutation à ces différentes propriétés.

41 Nous définissons la méthode suivante pour mesurer la dynamique des taux de mutation. Chaque famille est considérée comme un ensemble croissant de citations qui peuplent progressivement les différentes sous-familles. Chaque fois qu'une nouvelle citation est produite, nous enregistrons si elle est une copie parfaite d'une citation déjà mentionnée précédemment, ou une nouvelle version qui n'avait jamais été observée auparavant. Dans ce dernier cas nous cherchons aussi à déterminer si la citation originale enrichit une sous-famille existante ou initie une sous-famille originale. Nous compilons ces événements en fonction de l'état d'origine de la famille et de la sous-famille, c'est-à-dire que nous énumérons les microchangements, les changements macro et les événements de copie parfaite en fonction de la taille de la famille/sous-famille et de la longueur moyenne des citations qui la peuplent. À partir de là, il est aisé de définir le taux de mutation micro/macro en fonction de la longueur moyenne donnée ou un nombre total de mentions comme la proportion d'événements de réplication produisant un changement micro/macro.

42 La figure 4 montre que la stabilité d'une citation est sensiblement liée à sa longueur *l*, mesurée en nombre de mots. Donc nous avons tracé dans la figure 7 les taux de mutation en fonction de la longueur moyenne des citations de la famille/sous-famille.

Figure 7 – Taux de mutation micro et macro en fonction de la longueur des citations. Les points représentent les valeurs mesurées, la ligne continue représente le modèle approché ρM (l) = 0.020 − 0.292 exp (−0.499l) et ρμ (l) = 0.004 + 0.046(l − 5) exp (−0.423l)



43 On observe que le taux de mutation macro croît avec la longueur des citations. Alors que les petites phrases (moins de cinq mots) ne peuvent naturellement pas subir de macromutation compte tenu de la définition retenue pour catégoriser les familles, ce taux de mutation atteint après une croissance rapide un plateau pour les citations de taille supérieures à 20. Dans notre modèle, nous utilisons simplement une fonction exponentielle pour exprimer la dépendance du taux de mutation macro avec la longueur des citations (voir la légende de la figure 7 pour de plus amples détails). On peut interpréter cette courbe de la façon suivante : la probabilité qu'une citation puisse être décomposée en plusieurs sous-parties sémantiquement autonomes est nulle pour les phrases très petites, et beaucoup plus élevée dès que la phrase excède une vingtaine de mots.

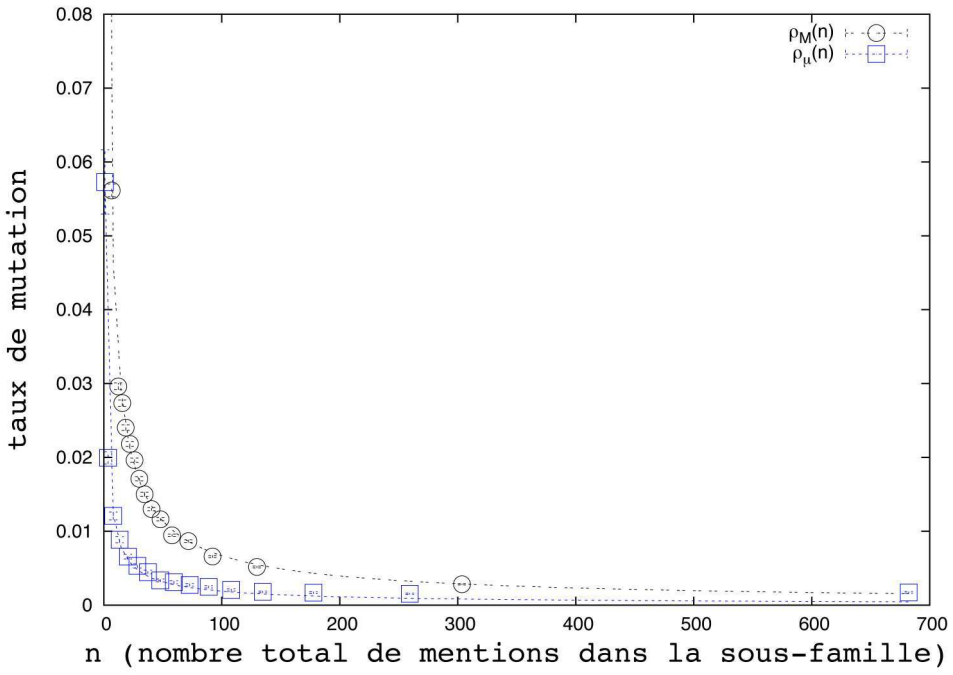
44 La corrélation entre le taux de mutation micro et la longueur des citations semble plus complexe. Le taux maximal de mutation micro est atteint pour des citations composées de huit mots. Au-delà, on observe une décroissance exponentielle jusqu'à ce que le taux de mutation atteigne un seuil minimal. Comme on avait déjà pu le suspecter lors de l'analyse de la stabilité des citations en fonction de leur taille, il semblerait que les changements micros soient guidés par deux processus distincts indiqués par la forme complexe du taux de mutation micro et l'influence de la taille des citations. Premièrement, il paraît clair que des erreurs ne peuvent être introduites que dans les cas où la citation n'est pas le résultat d'un simple copier/coller. Or, il semble raisonnable de postuler que la probabilité qu'une phrase soit recopiée de mémoire est exponentiellement décroissante avec la taille des citations. Si la citation a été recopiée de mémoire, il est alors probable que certaines erreurs aient été introduites. Si l'on se réfère aux travaux classiques de psychologie cognitive (Miller, 1956), le cerveau ne peut mémoriser que des séquences composées d'un

certain nombre d'objets. Ce « nombre magique », en-dessous duquel la mémoire à court terme est quasiment parfaitement fiable a été expérimentalement évalué à environ 5 lorsqu'il s'agit de mémoriser des séquences de mots.

45 C'est la raison pour laquelle nous avons choisi de reproduire la corrélation entre le taux de mutation micro et la longueur des citations avec une équation plus complexe composée par le produit de deux probabilités : la probabilité qu'une phrase citée ne soit pas répliquée par copier/coller et la probabilité qu'une altération soit introduite par erreur (que l'on considère comme linéaire en fonction du nombre de mots qui excède ce *fameux chiffre magique*) si la citation est répliquée de mémoire. Ce produit modélise la probabilité qu'un blogueur ou un journaliste introduise un changement « involontaire » dans la citation (voir la légende de la figure 7 pour une description plus détaillée).

46 Le nombre de mentions déjà recueillies par une citation pourrait également être un paramètre déterminant pour la vitesse de mutation. La figure 5 montre en tout cas que les citations qui sont largement dupliquées sont également les plus stables. Fort de cette intuition, nous représentons les taux de mutation en fonction du nombre total de mentions déjà reçues dans la famille/sous-famille. Nous observons que les taux de mutation diminuent considérablement en fonction du nombre de mentions. Ce comportement confirme notre hypothèse que les citations plus populaires sont moins sensibles aux changements. Les citations très populaires pourraient être si présentes dans l'ensemble de l'espace médiatique que la probabilité d'introduire des micromutations par erreur est fortement diminuée (le grand nombre de copies « rappelant » les agents à une formulation « correcte » de la citation). Autre hypothèse possible, les citations les plus populaires, rencontrant le plus de succès, sont naturellement dotées d'une « fitness » supérieure, si bien qu'elles constituent des attracteurs cognitifs naturels.

Figure 8 – Taux de mutation micro et macro en fonction du nombre total de mentions présentes dans la sous-famille/famille, et leurs modèles approchés. ρM(n) = 0.225n−0.763 and ρμ(n) = 0.057n−0.739



Simulation

47 Nous proposons le processus génératif suivant pour tâcher de reconstruire des familles de citations. Notre objectif est de définir un modèle agent-centré réaliste qui rende compte de la distribution des tailles des familles/sous-familles et de la dynamique de notre mesure de diversité dans le temps. Nous nous appuyons sur le principe classique de l'urne de Polya (Simmons et Adamic, 2011) en faisant l'hypothèse que chaque sous-famille s'organise autour d'un noyau de sens original et une dynamique propre indépendante des autres sous-familles. C'est pourquoi, nous postulons le processus itératif suivant : à chaque pas de temps, une sous-famille est d'abord choisie aléatoirement puis une citation de cette sous-famille est choisie au hasard (avec une probabilité proportionnelle à son nombre de mentions) comme candidate à une réplication, cette réplication peut alors donner lieu à une mutation de type micro ou macro en fonction de la longueur de la citation ou du nombre de fois où elle a été mentionnée. Dans la simulation, nous utilisons les fonctions approchées des taux de mutation micro ($\rho\mu(l)$, $\rho\mu(n)$) et macro ($\rho M(l)$, $\rho M(n)$) estimées dans la section précédente. En fonction de ces probabilités, une nouvelle citation est reversée dans la famille, soit sous une forme inchangée, soit avec une mutation micro (créant une nouvelle version dans la sous-famille choisie), soit avec une mutation macro (créant une nouvelle sous-famille). Pour chaque famille, ce processus est répété jusqu'à ce que la famille reçoive exactement le même nombre de citations que dans la distribution empirique originale.

Résultats du modèle

48 Notre simulation montre que le modèle proposé reconstruit avec succès la distribution des tailles de sous-familles, ainsi que la diversité des familles et sous-familles, calculée avec la mesure d'entropie. La figure 9 montre un très bon accord entre les distributions de tailles de sous-familles empiriques et simulées, suggérant que notre modèle parvient à reproduire des sous-familles composées de citations très ressemblantes. De plus la figure 10 montre que notre modèle parvient également à reproduire la distribution d'entropie des familles et des sous-familles illustrant, à cardinalité égale, la plus grande diversité des familles par rapport aux sous-familles.

Figure 9 – Comparaison de la distribution des tailles de sous-familles dans les données empiriques (points noirs) et produites par notre modèle (losange bleu).

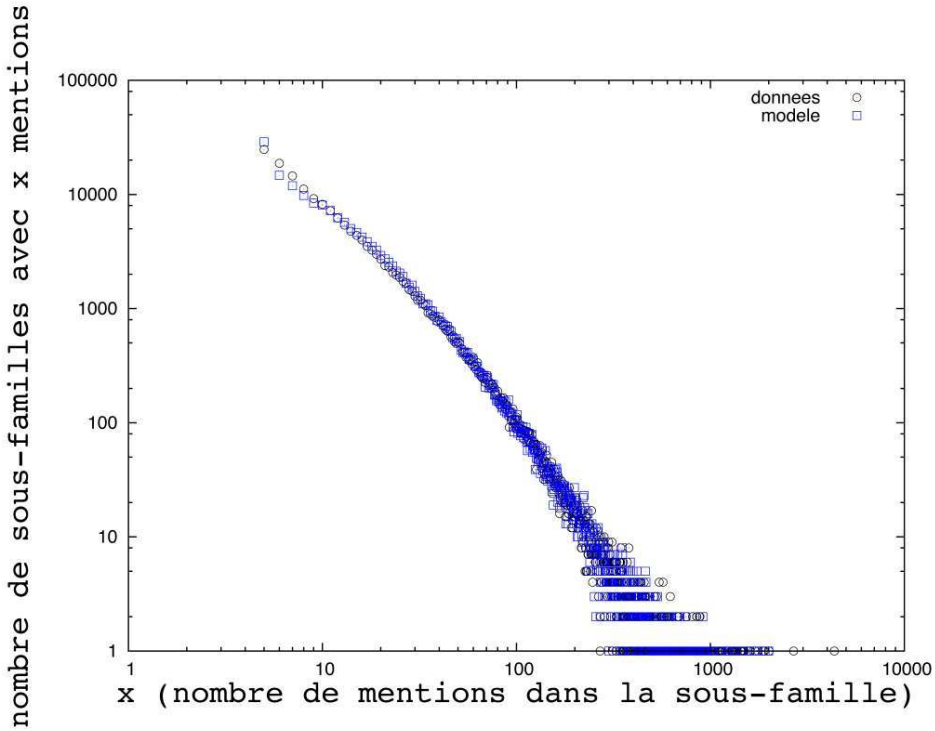
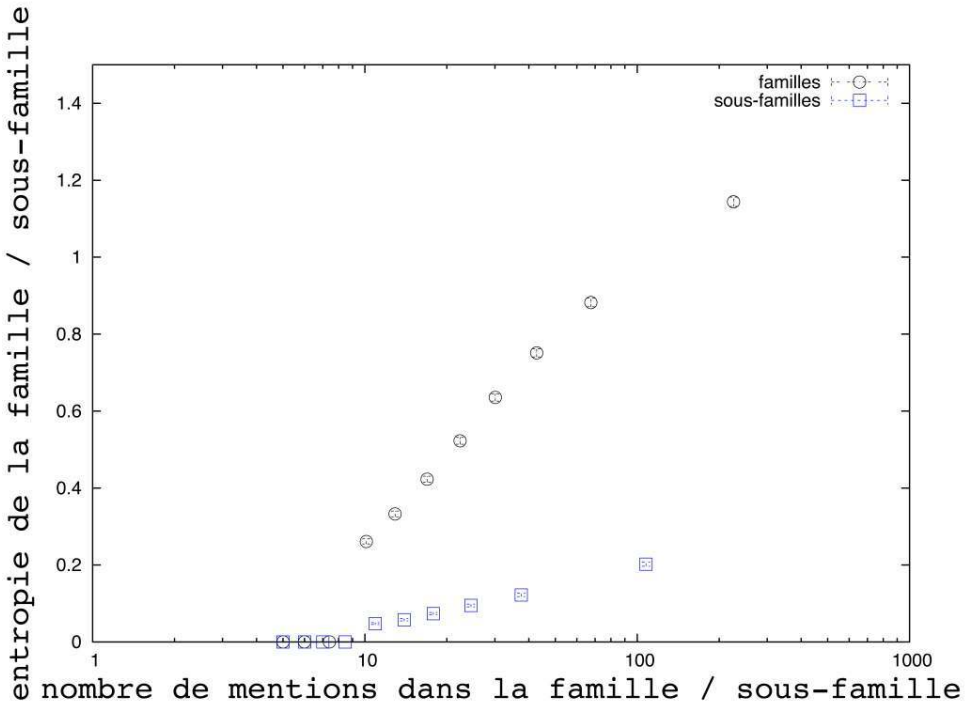


Figure 10 – Entropie des familles et sous-familles produites par le modèle en fonction du nombre total de mentions de leurs citations.



Conclusion

49 Dans cet article nous avons introduit un nouvel algorithme de *clustering* de citations comme première étape pour l'analyse de la structure de familles de citations et de leurs transformations. Nous avons montré comment la structure mésoscopique de ces familles peut être décrite par leurs sous-familles. La différence en matière de diversités (entropies) des citations calculées à ces deux échelles montre que deux processus dynamiques distincts sont à l'œuvre : la compétition que se livrent des citations très semblables au sein d'une sous-famille produit des ensembles plus homogènes qu'à l'échelle des familles elles-mêmes, naturellement plus propices à produire de l'hétérogénéité. Nous avons également présenté un modèle de morphogenèse des familles rendant compte de la diversité des tailles de sous-familles et de la différence de diversité aux deux échelles. Ce modèle s'appuie sur l'analyse de la stabilité des citations qui semble fortement liée à leur taille et à leur popularité. Dans des travaux futurs, il serait également souhaitable de prendre en compte la structure du réseau social sous-jacent, qui permettrait sans nul doute d'enrichir l'analyse des déterminants de la stabilité des citations. On peut notamment penser que des citations étant recopiées de sources « autorisées » subiront relativement moins de transformations. Une description plus fine de la dynamique de diffusion des citations telle qu'elle est réalisée dans Leskovec *et al.* (2009) nous permettrait également de produire un modèle de morphogenèse des familles plus exhaustif.

Remerciements
Les auteurs remercient Tommaso Brotto, Zorana Ratkovic pour avoir jouer le rôle de juges dans l'évaluation des méthodes de clustering.
Nous remercions également chaleureusement Sébastien Lerique, Camille Roth, Andrei Mogoutov, Benjamin Fagard, Isabelle Tellier et Jonathan Platkiewicz pour leurs précieux commentaires.

BIBLIOGRAPHIE

Adar, E., Zhang, L., Adamic, L. A. et Lukose, R., « Implicit structure and the dynamics of blogspace », *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*, 2004.

Aunger, R., « Cultural transmission and diffusion », *Encyclopedia of cognitive science*, Londres, MacMillan, 2003.

Coleman, J. S., Katz, E. et Menzel, H., « The Diffusion of an Innovation Among Physicians », *Sociometry*, vol. 20, n° 4, 1957.

Dawkins, R., *The selfish gene*, Oxford, Oxford University Press, 2006.

Edmonds, B., « Three challenges for the survival of memetics », *Journal of Memetics-Evolutionary models of information transmission*, vol. 6, n° 2, 2002, p. 45–50.

Kristeva, J., « Word, dialogue and novel », in T. Moi, *The Kristeva Reader*, Oxford, Blackwell, 1966.

Lancichinetti, A. et Fortunato, S., « Community detection algorithms : A comparative analysis », *Phys. Rev. E*, 2009.

Lerique, S. et Roth, C., « How do brains copy and paste ? the semantic drift of quotes in blogspace », à paraître.

Leskovec, J., Adamic, L. A. et Huberman, B. A., « The dynamics of viral marketing », portal.acm.org, 2007.

Leskovec, J., Backstrom, L. et Kleinberg, J. M., « Meme-tracking and the Dynamics of the News Cycle », *Proceedings of The Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09)*, 2009.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. et Hurst, M., « Cascading behavior in large blog graphs », *SIAM International Conference on Data Mining (SDM 2007)*, 2007.

Lieberman, E., Michel, J., Jackson, J., Tang, T. et Nowak, M., « Quantifying the evolutionary dynamics of language », *Nature*, vol. 449, n° 7163, 2007, p. 713–716.

Mihalcea, R., Corley, C. et Strapparava, C., « Corpus-based and knowledge-based measures of text semantic similarity », in *Proceedings of the American Association for Artificial Intelligence* (AAAI 2006), Boston, 2006.

Miller, G., « The magical number seven, plus or minus two : Some limits on our capacity for processing information », *Psychological review*, vol. 101, n° 2, 1956, p. 343.

Padó, S., *User's guide to sigf : Significance testing by approximate randomisation*, 2006.

Qiu, L., Kan, M.-Y. et Chua, T.-S., « Paraphrase recognition via dissimilarity significance classification », in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006.

Rogers, E. M., « New Product Adoption and Diffusion », *The Journal of Consumer Research*, 1976.

Rosvall, M. et Bergstrom, C. T., « Maps of random walks on complex networks reveal community structure », *PNAS*, 2008.

Salton, G. et McGill, M., *Introduction to modern information retrieval*, New York, McGraw-Hill, 1983.

Schmid, H., « Probabilistic part-of-speech tagging using decision trees », in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Shannon, C. E., « A Mathematical Theory of Communication », *Bell Sys. Tech. J.*, vol. 27, 1948, p. 379.

Simmons, M. et Adamic, L., « Memes Online: Extracted, Subtracted, Injected, and Recollected », *ICWSM 2011*, 2011.

Sperber, D., *La Contagion des idées*, Odile Jacob, 1996.

RÉSUMÉS

Dans cet article, nous étudions les dynamiques de prolifération et de diversification des « citations » dans la blogosphère. Dans la continuité des travaux séminaux de Leskovec et Simmons sur les dynamiques « culturelles » dans les médias sociaux, nous analysons en profondeur les transformations que les citations subissent au cours de leur diffusion en ligne. Nous ne visons pas dans notre approche à modéliser la dynamique temporelle du processus de diffusion mais plutôt de décrire finement la nature des changements qui affectent les expressions placées entre guillemets. Quelles sont les grands types de transformations observées et quelles propriétés des citations les rendent plus ou moins sensibles à ces mutations ? En poursuivant la métaphore biologique, nous essayons de comprendre comment des mutations à différentes échelles génèrent des « espèces » de citations (familles).

In this paper we study the dynamics of growth and diversification of quotations in the blogosphere. In line with the seminal work of Leskovec and Simmons on cultural dynamic in social media, we analyze in depth the changes that quotations undergo during their dispersal. In our approach we do not aim to model the temporal dynamics of the diffusion process but rather to accurately describe the nature of the changes that affect quoted texts. What are the major types of changes observed and what properties of the quotations make them more or less prone to these changes ? Following a biological metaphor, we try to understand which way these changes at different scales.

INDEX

Mots-clés : dynamique culturelle, médias sociaux, mémétique, diffusion de l'information
Keywords : cultural dynamics, social media, memetics, information diffusion

AUTEURS

ELISA OMODEI

Doctorante au Laboratoire Langues, Textes, Traitements informatiques, Cognition de l'ENS.
Parmi les publications :
Avec A. Bazzani, S. Rambaldi, P. Michieletto et B. Giorgini, « The physics of the city : pedestrians dynamics and crowding panic equation in Venezia », *Quality & Quantity*, 2012.
Avec T. Poibeau et J.-Ph. Cointet, « Multi-Level Modeling of Quotation Families Morphogenesis », *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing*, 2012.
Avec T. Poibeau et J.-Ph. Cointet, « A symmetric approach to understand the dynamics of scientific collaborations and knowledge production », *Proceedings of MARAMI 2013* (Modèle & Analyse de réseaux : Approches mathématique & informatiques), 2013.

JEAN-PHILIPPE COINTET

Chercheur à l'Inra dans l'unité SenS (Sciences en Société).
Parmi les publications :
Avec C. Roth, « Social and Semantic Coevolution in Knowledge Networks », *Social Networks*, vol. 32, n° 1, 2010, p. 16-29.
Avec C. Taramasco et C. Roth, « Academic team formation as evolving hypergraphs », *Scientometrics*, vol. 85, n° 3, 2010, p. 721-740.
Avec L. Tabourier et C. Roth, « Generating constrained random graphs using multiple edge switches », *Journal of Experimental Algorithmics*, vol. 16, n° 1, 2011.
Avec S. Parasie, « Online press serving local democracy », *RFSP*, vol. 62, n° 1, 2012, p. 41-66.
Avec P. Keating, A. Cambrosio, N. Nelson et A. Mogoutov, « Therapy's shadow: a short history of the study of resistance to cancer chemotherapy », *Frontiers in Pharmacology*, n° 58, vol. 4, 2013.
Avec D. Chavalarias, « Phylomemetic Patterns in Science Evolution – The Rise and Fall of Scientific Fields », *PLoS ONE*, 8(2):e54847, 2013.