Longueur de branches et arbres de mots

Philippe Gambette, Núria Gala, Alexis Nasr

Longueur de branches et arbres de mots

Philippe Gambette[1], Nuria Gala[2], Alexis Nasr[2]
[1]Université Paris-Est – LIGM, [2]Université Aix-Marseille – LIF

Résumé : Les arbres de mots constituent un des outils de la statistique textuelle pour visualiser les relations sémantiques entre mots d'un texte. Les méthodes de construction de ces arbres à partir d'une distance de co-occurrence dans le texte produisent des arbres dont les longueurs d'arêtes se prêtent mal à l'analyse. Pour faciliter l'interprétation visuelle de l'arbre, l'idéal serait que des longues arêtes séparent des classes sémantiques de mots. Ainsi, découper les arêtes les plus longues de l'arbre devrait conduire à une partition de l'ensemble des mots qui fournit des classes pertinentes. À l'aide de deux corpus dont un sous-ensemble de mots a été partitionné en un ensemble de classes sémantiques, nous évaluons plusieurs formules permettant de recalculer les longueurs d'arêtes de l'arbre construit à partir des distances de co-occurrence, afin de rendre l'interprétation de l'arbre plus facile et plus fiable.

Mots-clés : classification hiérarchique, visualisation, arbre, nuage arboré, co-occurrence, partition

Branch Lengths and Word Trees

Philippe Gambette[1], Nuria Gala[2], Alexis Nasr[2]
[1]Université Paris-Est – LIGM, [2]Université Aix-Marseille – LIF

Summary: Word trees are one of the available tools in textual analysis to visualize semantic relationships between the words of a text. Tree construction methods from the co-occurrence distances between words in a text produce trees whose edge lengths are difficult to analyze. In order to make the visual interpretation of the tree easier, long edges should separate semantic classes of words. Therefore, cutting the longest edges in the tree should lead to a partition of the word set with relevant classes. Using two corpuses where a subset of words was partitioned into semantic classes, we evaluate several formulas computing new edge lengths for a tree built from co-occurrence distances, aiming at making the interpretation of the tree easier and more reliable.

Keywords: hierarchical clustering, visualization, tree, tree cloud, co-occurrence, partition

Longueur de branches et arbres de mots

Philippe Gambette[1], Nuria Gala[2], Alexis Nasr[2]
[1]Université Paris-Est – LIGM, [2]Université Aix-Marseille – LIF

1. Introduction

1.1 Les arbres de mots comme classifications

Les arbres de mots se sont ajoutés aux projections et aux réseaux de co-occurrence parmi les outils développés pour l'analyse textométrique des textes (Luong, 1989 ; Mayaffre, 2008). Ils permettent en effet de représenter de manière esthétique un nombre limité de classes de mots emboîtées (en nombre linéaire par rapport au nombre de mots), tout en laissant la possibilité de faire varier les tailles de caractères des mots, par exemple dans les nuages arborés (Gambette & Véronis, 2009), dont une illustration est donnée en Figure 1.

L'arbre est construit à partir d'une matrice de distances entre les mots, en utilisant un algorithme de classification hiérarchique. La *distance de co-occurrence* entre deux mots *a* et *b* dans cette matrice de distances est proche de 0 si les mots apparaissent souvent à proximité dans le texte, et grande s'ils apparaissent rarement ensemble. Interprétée comme une distance sémantique en suivant le principe selon lequel le sens du mot provient de ses voisins (Firth, 1957), elle conduit à l'analyse suivante de l'arbre construit pour refléter au mieux cette distance : un sous-arbre regroupe des mots dont les distances sont petites comparées aux distances avec les mots du reste de l'arbre, donc ils apparaissent plus fréquemment ensemble dans le texte qu'avec des mots du reste de l'arbre. On en déduit qu'ils constituent une classe sémantique, et donc bien

souvent représentent une thématique du texte dont ils ont été extraits.

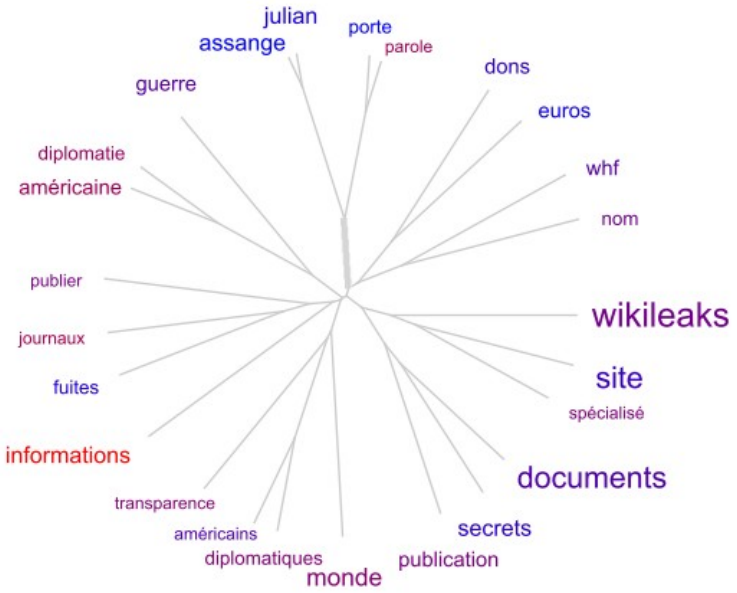

Figure 1 : Nuage arboré des 25 mots les plus fréquents (hors mots vides) du corpus Wikileaks (voir Section 4), construit par les logiciels TreeCloud (Gambette & Véronis, 2009) et SplitsTree (Huson & Bryant, 2006). L'arête en gras sépare la classe [julian, assange, porte, parole] du reste des mots.

On remarque par exemple, dans une lecture rapide du nuage arboré de la Figure 1, que le mot au coeur du texte utilisé pour construire cette visualisation est “wikileaks”, qui se trouve dans un sous-arbre aux côtés de “site” et “spécialisé”. Ceci nous permet d'esquisser une définition de la thématique principale du texte, que l'on complète en remarquant que le sous-arbre correspondant à la classe {wikileaks, site, spécialisé} est inclus dans le sous-arbre correspondant à la classe {wikileaks, site, spécialisé, publication, documents, secrets}. L'arbre rapproche également des composants du mot composé “porte-parole”, ou

le prénom “julian” du nom “assange”. Le fait que ces quatre mots sont les feuilles d'un même sous-arbre nous invite à déduire que dans le texte, Julian Assange est présenté comme le porte-parole de Wikileaks.

1.2 Interprétation des longueurs de branches

Cette interprétation d'un arbre de mots comme un simple ensemble de classes de mots emboîtées prend en compte uniquement la topologie de l'arbre, et ne fait pas intervenir les longueurs de branches. Pourtant, des longueurs de branches sont naturellement calculées par toute méthode de classification hiérarchique à partir d'une matrice de distance. La propriété attendue de ces longueurs est que les distances dans l'arbre obtenu[1] soient aussi proches que possible des distances fournies en entrée dans la matrice. Ainsi, des méthodes qui s'attachent à fournir des longueurs d'arêtes pertinentes peuvent avoir pour objectif une optimisation par les moindres carrés entre la matrice de distance fournie en entrée et les distances estimées dans l'arbre en sortie.

Ce type de méthodes, pour lesquelles les distances entre feuilles dans l'arbre calculé ont un grand intérêt, ont été particulièrement utilisées pour l'étude de l'évolution des espèces ou *phylogénie* (Felsenstein, 2004). En effet, la longueur du chemin entre deux feuilles de l'arbre a une interprétation directe : c'est la distance évolutive entre les deux espèces représentées par ces feuilles, qui se reflète généralement dans la distance entre leurs ADN.

Cette interprétation des distances entre feuilles de l'arbre est beaucoup moins pertinente pour un arbre de mots, pour trois raisons : de *modélisation*, de *fiabilité* et de *lisibilité*.

---

[1] rappelons que la distance entre deux feuilles d'un arbre est égale à la somme des longueurs des arêtes dans le chemin allant, dans l'arbre, d'une feuille à l'autre..

Tout d'abord, notons que dans un arbre phylogénétique, les sommets internes représentent des espèces ancestrales, et que les longueurs d'arêtes modélisent des distances d'évolution entre les espèces, ancestrales ou actuelles, correspondantes. En revanche, dans un arbre de mots, il est difficile d'interpréter les nœuds internes, et d'en déduire une façon naturelle d'interpréter la longueur d'une arête située entre deux nœuds internes de l'arbre.

L'absence d'un modèle d'évolution arborée implique aussi un problème de fiabilité : contrairement aux distances phylogénétiques qui ont généralement une structure proche d'une métrique d'arbre, les distances de co-occurrence entre mots peuvent être très éloignées de toute représentation arborée. Ainsi, l'approximation, fournie par l'arbre obtenu en sortie, des distances entre feuilles données dans la matrice en entrée, peut être très mauvaise, quand bien même on aurait calculé l'arbre optimal au sens de l'optimisation des moindres carrés.

Enfin, même si l'on choisit de ne pas interpréter les longueurs d'arêtes internes de l'arbre, et de se focaliser uniquement sur les distances entre feuilles, en espérant une certaine fiabilité, l'estimation visuelle de ces dernières pose un problème de lisibilité. Dans la Figure 1 par exemple, s'il est clair que dans l'arbre, “diplomatie” est plus proche d'“américaine” que de “publier”, il est en revanche difficile de comparer la distance entre “diplomatie” et “wikileaks” avec celle entre “dons” et “monde”. Un problème supplémentaire de lisibilité apparaît avec les données textuelles : la longueur excessive des branches menant aux feuilles (appelées *arêtes externes*) réduit la lisibilité de l'intérieur de l'arbre. Il s'agit là d'un défaut des méthodes de construction d'arbres à partir de distances, en particulier la méthode Neighbor-Joining de Saitou & Nei (1987) utilisée dans cet article : en raison de la structure très particulière des formules de co-occurrence de mots (Evert, 2005), on peut constater que la longueur des arêtes internes d'un

nuage arboré est souvent très petite par rapport à celle des arêtes menant aux feuilles.

Ces trois constats montrent les limites d'une utilisation des arbres de mots dont les longueurs d'arêtes sont calculées directement par l'algorithme de classification hiérarchique. Dès lors, il convient de proposer un modèle d'interprétation de l'arbre, et de calcul de ses longueurs d'arêtes, qui soit pertinent pour la textométrie.

Nous avons vu que l'interprétation la plus immédiate de l'arbre consiste à considérer chaque sous-arbre comme une classe de mots. Il convient donc de faciliter et de renforcer cette interprétation, en choisissant une méthode de calcul des longueurs d'arêtes compatible avec cet objectif. Nous proposons de recalculer les longueurs de branches après la construction de l'arbre, de telle manière qu'elles assurent sa lisibilité, tout en facilitant la lecture de l'arbre comme une partition en classes de l'ensemble des mots. Pour cela, nous proposons d'utiliser des formules proposées par Guénoche & Garreta (2002) pour évaluer la qualité des arêtes d'un arbre. Ces formules indiquent si les deux ensembles de mots séparés par une arête sont effectivement bien séparés d'après la matrice de distance. Ainsi, on attribuera à chaque arête une longueur proportionnelle à son score de qualité, et les arêtes les plus longues seront les plus discriminantes.

1.3 Évaluation par la construction automatique d'une partition

Afin d'évaluer la pertinence de ces formules, nous proposons un algorithme qui construit une partition d'un ensemble de mots, à partir d'un arbre de mots, en découpant successivement, dans l'ordre décroissant des longueurs, ses arêtes internes, jusqu'à un certain critère d'arrêt, comme illustré en Figure 2. Cet algorithme correspond donc à la démarche effectuée visuellement par l'utilisateur de l'analyse arborée, qui interprète

les arêtes les plus longues de l'arbre comme des séparations entre des classes de mots regroupés en raison de leur proximité sémantique.

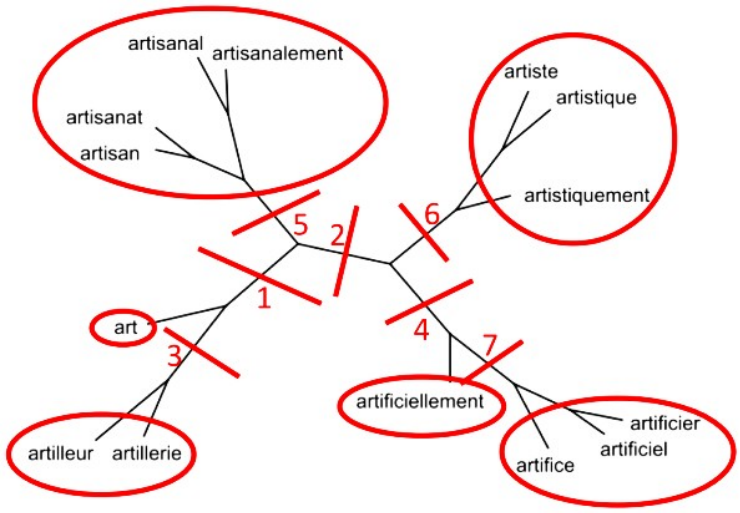

Figure 2 : Arbre de la famille de « art » dans la base de données Polymots (Gala & Rey, 2008). Les sept arêtes les plus longues sont numérotées dans l'ordre des longueurs décroissantes, et les six classes de mots obtenues après découpage de ces sept arêtes sont montrées par des ellipses rouges.

Ainsi, l'algorithme permet d'obtenir une partition de l'ensemble des mots, qui dépend directement de la longueur choisie pour les arêtes de l'arbre. C'est ce principe que nous allons utiliser pour évaluer les diverses formules permettant de calculer les longueurs d'arêtes de l'arbre. Nous allons en effet comparer des partitions de référence avec des partitions obtenues automatiquement par cette méthode.

Pour cela, nous utiliserons deux corpus d'évaluation. Le premier est constitué par 20 partitions de familles de mots de la base de données Polymots (Gala & Rey, 2008). Les

informations de distances entre les mots d'une même famille proviennent à la fois d'informations de co-occurrence dans le TLFi (Dendiel & Pierrel, 2003), et du nombre d'affixes communs (Gala et al., 2011). La partition de référence pour chacune de ces 20 familles a été construite manuellement, en effectuant des choix arbitraires de classe pour les mots polysémiques qui pourraient appartenir à plusieurs classes. Un exemple de famille et de sa partition en classes est donné au début de la Section 3.

Le second corpus est constitué par 10 textes écrits autour de 25 mots relatifs à l'organisation Wikileaks, ces mots étant organisés en une partition de référence. Chacun des 10 textes est donc censé faire apparaître les mots d'une même classe à proximité les uns des autres, et la distance de co-occurrence entre les mots est calculée par TreeCloud.

Ces deux corpus sont utilisés dans des protocoles d'évaluation qui permettent de faire émerger deux méthodes de calcul appropriées pour le calcul des longueurs d'arête de l'arbre, parmi les 5 testées : *triples* et *lengthRatio*.

2. Les formules de longueurs d'arêtes

Notre objectif est d'attribuer des longueurs d'arêtes cohérentes avec les informations de distance sémantique entre mots, c'est-à-dire une grande longueur aux arêtes qui séparent effectivement deux groupes de mots sémantiquement éloignés l'un de l'autre, et une petite longueur aux arêtes qui séparent des mots proches sémantiquement. Pour cela, plusieurs formules sont possibles à partir d'une matrice de distances entre mots de l'arbre (Guénoche & Garreta, 2002). Nous notons $d(a,b)$ la distance sémantique entre deux mots a et b d'après cette matrice de distances. Pour être précis, il ne s'agit pas exactement d'une distance au sens mathématique du terme, puisqu'elle ne respecte pas, généralement, l'inégalité triangulaire, mais d'une *dissimilarité*. Nous comparerons cinq formules possibles,

notées *computedLength*, *triples*, *quartets*, *lengthRatio*, et *agreementPairs*.

La première, *computedLength*, consiste à considérer simplement la longueur de l'arête calculée par l'algorithme de construction de l'arbre à partir de la matrice de distance, c'est-à-dire l'algorithme Neighbor-Joining (Saitou & Nei, 1987).

La deuxième, *triples*, désigne le taux de bons triplets séparés par l'arête. Elle consiste à calculer, pour tout ensemble de trois mots $\{a,b,c\}$, où a et b sont situés d'un côté de l'arête et c de l'autre côté, la proportion de deux qui vérifient :

$$d(a,b) \leq \min(d(a,c),d(b,c)).$$

Si l'arête est cohérente avec les données de distance sémantique, le score *triples*, compris entre 0 et 1, doit être proche de 1. Inversement, si les mots de part et d'autre de l'arête sont situés à distance inférieure aux mots d'un même côté de l'arête, ce score *triples* sera proche de 0.

De manière similaire, la troisième formule, *quartets*, désigne le taux de bons quadruplets séparés par l'arête. Pour le calculer, il faut évaluer, pour tout ensemble de quatre mots $\{a,b,x,y\}$, où a et b sont situés d'un côté de l'arête, et x et y de l'autre côté, la proportion de ceux qui vérifient :

$$d(a,b)+d(x,y) \leq \min(d(a,x)+d(b,y), d(a,y)+d(b,x)).$$

De nouveau, un bon score sera proche de 1 et un mauvais sera proche de 0.

Pour calculer le taux d'accord des paires, noté *agreementPairs*, on commence par classer, dans l'ordre croissant, les distances entre paires de mots distincts. Parmi ces $n(n-1)/2$ distances (pour n mots), on s'attend à ce que celles entre deux mots d'un même côté de l'arête soient inférieures à celles entre deux mots séparés par l'arête. En appelant donc m le nombre de paires de mots d'un même côté de l'arête, et d_m la m-ième plus petite distance entre paires de mots distincts, la formule *agreementPairs* calcule la somme du nombre de paires de mots distincts d'un même côté de l'arête dont la distance est

inférieure ou égale à *d_m* d'une part (conformément à ce qui est attendu), avec d'autre part le nombre de paires de mots distincts séparés par l'arête dont la distance est supérieure ou égale à *d_m* (conformément à ce qui est attendu), somme divisée par le nombre total de distances, soit $n(n\text{-}1)/2$. Comme il s'agit d'un taux de paires de mots, les scores des arêtes cohérentes avec les données de distance seront de nouveau proches de 1, et les mauvais proches de 0.

Enfin, le ratio des longueurs moyennes, noté *lengthRatio*, est la distance moyenne entre mots séparés par l'arête, divisée par la distance moyenne entre mots d'un même côté de l'arête. Si l'arête est bien cohérente avec les données de distance sémantique, on attend que ce score soit strictement supérieur à 1, sinon, il sera inférieur à 1.

3. Évaluation sur des familles morphologiques

3.1 Protocole d'évaluation

Pour chacune des formules de longueur d'arête décrites ci-dessus, nous avons testé leur performance avec la base de données Polymots. Cette base comprend 20 000 mots regroupés en 2 000 familles, chacune centrée autour d'un mot racine. Parmi ces familles, 20 ont été partitionnées manuellement en classes sémantiques. Par exemple, voici la partition pour un extrait de la base correspondant à la famille du radical « art » : [artificier, artifice, artificiel, artificiellement] [artillerie, artilleur] [artisan, artisanal, artisanalement, artisanat] [artiste, artistique, artistiquement, art].

Pour chacune des formules de longueur d'arêtes, nous effectuons, pour chaque famille de mots, une comparaison, à l'aide de l'indice de Rand, ou de l'indice de Rand corrigé, entre les partitions construites manuellement et celles construites automatiquement à partir d'une distance sémantique prenant en compte les co-occurrences des mots dans le TLFi ainsi que leur

nombre d'affixes communs (Gala et al, 2011). Les partitions construites automatiquement le sont en utilisant un algorithme de découpage des arêtes par longueur décroissante, illustré en Figure 2. Au *k*-ième découpage d'arête, on considère que chaque composante connexe obtenue, dans l'arbre ainsi découpé, fournit une classe de la partition. Nous obtenons ainsi une partition d'au plus *k*+1 classes, dont nous pouvons calculer un score de similarité avec la partition manuelle, comme montré en Figure 3.

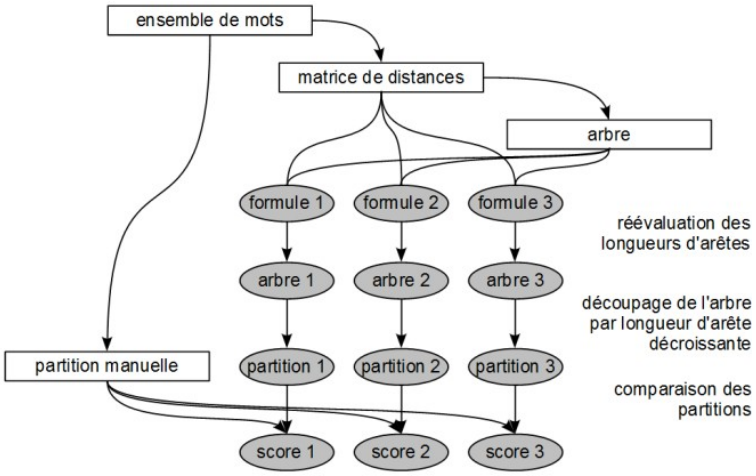

Figure 3 : Méthodologie d'évaluation des formules de calcul de longueur des arêtes de l'arbre.

Plus précisément, pour obtenir un score de qualité, comme il y a *n*-3 arêtes internes dans un arbre de *n* mots, nous sélectionnons parmi ces partitions obtenues après 1, 2, 3... *n*-3 découpages, celle qui est la plus proche de la partition de référence, selon l'indice de Rand, ou selon l'indice de Rand corrigé. Nous obtenons de cette façon deux scores de qualité

(Rand et Rand corrigé) pour chacune des formules de calcul des longueurs d'arêtes de l'arbre.

Rappelons que ces deux indices sont au plus égaux à 1, valeur atteinte pour deux partitions identiques. L'indice de Rand entre deux partitions *P1* et *P2* correspond à la proportion de paires d'éléments qui sont dans la même classe à la fois dans *P1* et dans *P2*, ou dans des classes distinctes à la fois dans *P1* et dans *P2* (Rand, 1971). Comme cet indice a tendance à surestimer la similitude entre deux partitions, l'indice de Rand corrigé (Hubert & Arabie, 1985) a été proposé. Son espérance pour deux partitions aléatoires est nulle, il soustrait donc la part de similitude due au hasard.

3.2 Résultats

Les moyennes des scores de Rand, données en Figure 4, et plus encore celles des scores de Rand corrigé, montrent que pour ces 20 familles de mots, l'utilisation des formules *triples* et *lengthRatio* fournit les meilleures résultats pour fixer les longueurs d'arête.

Formule :	*computed Length*	*triples*	*quartets*	*lengthRatio*	*agreement Pairs*
Rand	0.783	**0.791**	0.762	**0.792**	0.765
Rand corr.	0.354	**0.396**	0.254	**0.392**	0.270

Figure 4 : Moyenne, pour 20 familles de mots de la base Polymots, des scores de Rand et de Rand corrigé pour la meilleure partition construite automatiquement en fonction de la formule choisie pour calculer les longueurs des arêtes.

Pour plus de détails, les scores de Rand et Rand corrigé obtenus pour les 10 premières familles de mots sont donnés en Figures 5 et 6, respectivement.

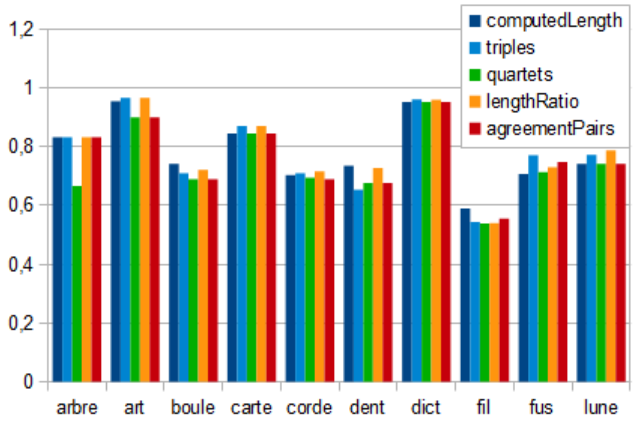

Figure 5 : Score de Rand pour la meilleure partition construite automatiquement en fonction de la formule choisie pour calculer les longueurs des arêtes, pour 10 familles de mots de la base Polymots.

On peut également s'interroger sur la similarité entre ces formules de calcul des distances d'arêtes. Pour en savoir plus, il est possible de comparer les ensembles de longueurs d'arêtes obtenues pour chacune de ces cinq formules. Si l'on se focalise sur l'arbre de la famille du mot “art”, on constate que les longueurs d'arêtes calculées par la formule *lengthRatio* présentent une corrélation avec celles calculées par la formule *triples*, le coefficient de corrélation entre ces ensembles de distances étant de 0,865. Les autres choix de paires de formules ne font en revanche pas apparaître de corrélations aussi nettes. On note également, toujours sur l'arbre de la famille “art”, que les arêtes internes ont une longueur généralement plus importante que les arêtes externes, tant par le calcul avec la formule *triples* que celui avec *lengthRatio*. Ceci est un avantage pour la lisibilité de l'arbre.

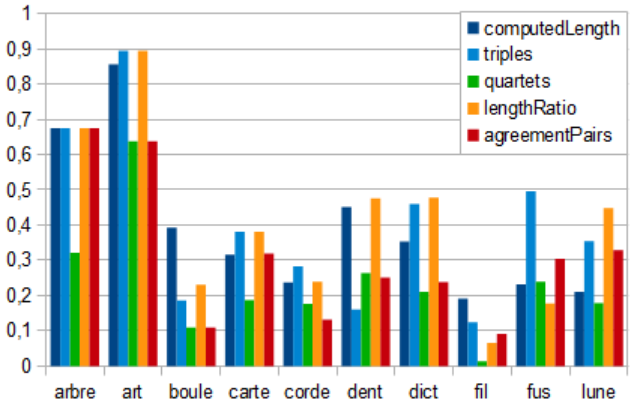

Figure 6 : Score de Rand corrigé pour la meilleure partition construite automatiquement en fonction de la formule choisie pour calculer les longueurs des arêtes, pour 10 familles de mots de la base Polymots.

En revanche, par rapport à la formule *triples*, la formule *lengthRatio* fournit une variance des longueurs d'arêtes moins importante. Pour améliorer la lisibilité, il faudra donc effectuer une transformation monotone des distances (par exemple, une transformation affine) qui augmente cette variance.

4. Évaluation sur un corpus textuel

4.1 Protocole d'évaluation

Dans l'évaluation sur des familles morphologiques ci-dessus, la distance entre les mots de la famille est une composition entre une distance sur les co-occurrents communs au sein du TLFi et une distance des affixes communs. Ainsi, cette distance n'est pas calculée directement à partir d'un corpus textuel, alors que c'est le cas pour les distances de co-occurrence utilisées comme base de la construction des nuages arborés.

Nous proposons donc un second protocole d'évaluation de la qualité de la partition qui se base sur les distances de co-

occurrence entre mots d'un texte implémentées dans TreeCloud. A partir d'une partition d'un ensemble de 25 mots liés à Wikileaks ([julian, assange, porte, parole], [dons, euros, nom, whf], [membres], [wikileaks, site, documents, publication, spécialisé, secrets], [guerre, diplomatie, américaine], [américains, diplomatiques, transparence, monde], [fuites, journaux, publier]), dix textes d'environ 300 mots ont été rédigés. Ce corpus est disponible sur le site treecloud.org.

La consigne suivante a été donnée pour l'élaboration des textes par dix groupes d'étudiants : rédiger "*un texte de plus de 300 mots qui fait obligatoirement apparaître les 25 mots voulus, en tentant de rapprocher les mots contenus dans une même classe de la partition*". La partition de départ provenait d'un découpage thématique en sous-arbres, réalisé manuellement, d'un nuage arboré. Ce nuage arboré montré en Figure 7 a été construit par le logiciel TreeCloud à partir de la concaténation de trois articles de presse : "WikiLeaks : une transparence qui fait débat", dans *Le Monde* du 30 novembre 2010, "WikiLeaks change la donne de la diplomatie et des médias", dans *Les Echos* du 29 novembre 2010, et "Wikileaks, une nébuleuse si peu transparente...", dans *Les Echos* du 9 décembre 2010.

Les 10 textes sont alors concaténés pour former le corpus d'évaluation, et c'est sur ce corpus qu'on applique la méthode de création d'un nuage arboré (fenêtre glissante de 10 mots et pas de glissement de 1 mot pour le calcul des co-occurrences, méthode Neighbor-Joining de Saitou & Nei (1987) pour la construction de l'arbre), puis la méthode indiquée dans la Figure 3 pour le calcul des longueurs d'arêtes, la création de la partition par découpage des arêtes les plus longues, et la comparaison avec la partition de référence, pour chaque formule de calcul des longueurs d'arête. Comme la partition de référence a 7 classes, nous arrêtons le processus de découpage

après 6 découpages d'arêtes, afin de créer la partition dont on évaluera la qualité par rapport à la partition de référence.

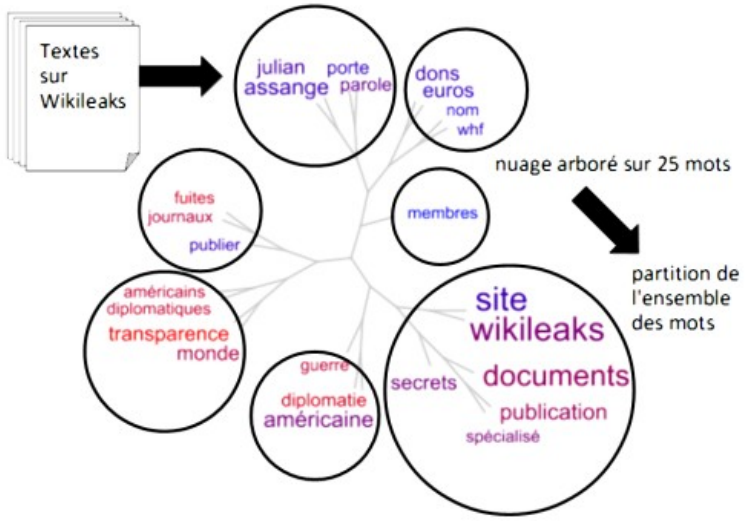

Figure 7 : Nuage arboré de trois articles de presse concaténés permettant de construire une partition de 25 mots liés à Wikileaks (chacune des 7 classes de la partition est contenue dans un cercle) fournie comme base de la rédaction des 10 textes du corpus d'évaluation.

Comme 13 distances de co-occurrence entre mots sont implémentées dans TreeCloud, nous avons choisi de nous focaliser sur les 7 dont la robustesse pour construire des nuages arborés était la meilleure, d'après l'analyse de Gambette & Véronis (2009) : *liddell*, *gmean*, *jaccard*, *dice*, *ms* (minimum sensitivity), *zscore* et *hyperlex* (Gambette, 2010).

4.2 Résultats

Nous obtenons les résultats montrés en Figure 8 pour le score de Rand corrigé. Les formules *lengthRatio* et *triplets* apparaissent encore une fois comme les meilleures, comme on le voit dans la Figure 9 avec les moyennes de score de Rand corrigé.

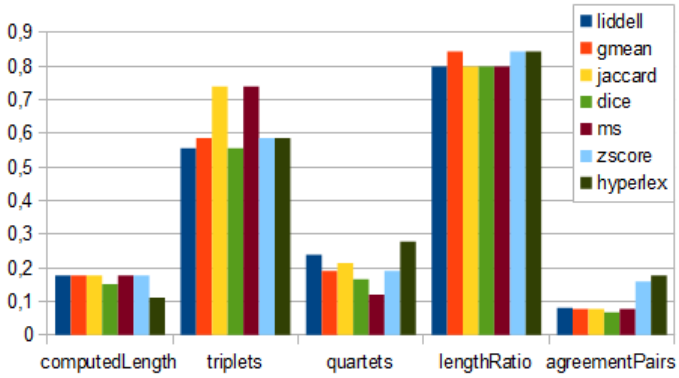

Figure 8 : Score de Rand corrigé des partitions construites automatiquement sur le corpus d'évaluation "Wikileaks", en fonction de 7 distances de co-occurrence et de 5 formules de calcul des longueurs d'arêtes.

Formule :	*computed Length*	*triples*	*quartets*	*lengthRatio*	*agreement Pairs*
Rand corr.	0.164	**0.621**	0.200	**0.818**	0.102

Figure 9 : Moyenne, pour 7 distances de co-occurrence, du score de Rand corrigé des partitions construites automatiquement sur le corpus d'évaluation "Wikileaks", en fonction de 5 formules de calcul des longueurs d'arêtes.

Les partitions obtenues sont très proches de la partition d'origine, voici par exemple celle obtenue avec la formule de co-occurrence *gmean* et la formule de longueurs d'arêtes *lengthRatio*, à comparer avec la partition originale de la Figure 7 : [assange, julian, porte, parole], [site, wikileaks, documents, secrets, publication, spécialisé], [américaine, guerre, diplomatie], [membres], [monde, fuites, diplomatiques, américains, transparence, journaux, publier], [dons, whf, euros, nom].

On peut remarquer que cette partition n'a que 6 classes, car un des 6 découpages n'a pas induit de séparation d'une

classe en deux. En effectuant le découpage de l'arête suivante la plus longue avec ces paramètres (*gmean* et *lengthRatio*), on retrouve exactement la partition originale.

5. Conclusions et perspectives

Cet article fournit une méthodologie de calcul des longueurs d'arêtes d'un arbre de mots qui permet de l'interpréter comme un ensemble de classes, les classes les mieux séparées l'étant par les arêtes les plus longues.

La proposition d'une méthode de partitionnement de l'ensemble des mots aux feuilles de l'arbre par découpage successif des arêtes dans un ordre de longueur décroissant permet de mettre en application ce principe. Comparer les partitions ainsi obtenues à une partition de référence nous a permis de déterminer que deux formules de calcul des longueurs d'arêtes semblent fournir de bons résultats : *triples* (le taux de triplets de mots séparés par cette arête qui le sont également d'après la matrice de distance) et *lengthRatio* (la distance moyenne entre mots de part et d'autre de l'arête, divisée par la distance moyenne entre mots d'un même côté de l'arête).

Ainsi, ces deux formules, ou toute transformation affine (ou plus généralement toute transformation qui respecte l'ordre relatif des arêtes en fonction de leur longueur ainsi calculée), permettent d'obtenir un arbre interprétable comme un ensemble de classes de mots plus ou moins séparées les unes des autres. Ce système augmente la fiabilité de l'interprétation de l'arbre, là où les longueurs directement calculées par la méthode de classification hiérarchique choisie pour construire l'arbre peuvent induire en erreur l'utilisateur qui tente d'interpréter l'arbre.

Comme nous l'avons vu avec la qualité des scores obtenus dans l'évaluation sur un corpus textuel simulant un ensemble d'articles portant sur un même sujet, la méthode de classification non supervisée d'un ensemble de mots en fonction

de leur distance sémantique proposée dans cet article permet d'obtenir de bons résultats, et a donc un intérêt en tant que telle. Une comparaison plus poussée avec d'autres méthodes, et sur d'autres corpus construits de la même manière, permettrait de confirmer la qualité de la méthode.

Remerciements

Nous remercions Alain Guénoche pour les discussions qui sont à l'origine de cet article, et les outils indiqués en réponse à nos besoins méthodologiques. Les remarques du relecteur anonyme de l'article ont également permis de l'améliorer. Le colloque *La co-occurrence : du fait statistique au fait textuel* a également permis de nourrir cet article grâce à plusieurs discussions et présentations. En particulier, les informations données par Jean-Marie Leblanc à propos de sa méthodologie de validation, sur des textes générés à partir du résultat attendu, ont inspiré le second protocole expérimental de cet article. Nous remercions enfin les étudiants du module d'Ingéniérie Linguistique du master 1 mention informatique de l'Université Paris-Est Marne-la-Vallée en 2011-2012, pour leur participation au projet *Classes de mots / Infoling 2012* qui a permis la constitution du corpus utilisé dans ce protocole expérimental, à partir des textes qu'ils ont rédigés.

Références

Dendien J. & Pierrel J.M. (2003). « Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence ». *Traitement automatique des langues* 44(2) : 11–37.

Evert S. (2005). *The Statistics of Word Cooccurrences, Word Pairs and Collocations*. Thèse de l'Université de Stuttgart, pp. 75–91.

Felsenstein J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Firth J.R. (1957). « A synopsis of linguistic theory, 1930–1955 ». *Studies in Linguistic Analysis*, pp. 1–32. Special Volume, Philological Society.

Gala N., Hathout N., Nasr A., Rey V. & Seppälä S. (2011). « Création de clusters sémantiques dans des familles morphologiques à partir du TLFi », In *Actes de TALN'11*.

Gala N. & Rey V. (2008). « Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques ». In *Actes de TALN'08*.

Gambette P. & Véronis J. (2009). « Visualising a Text with a Tree Cloud ». In Locarek-Junge H. and Weihs C., éditeurs, *Classification as a Tool of Research, Proc. of IFCS'09*, pp. 561–570.

Gambette P. (2010). « User manual for TreeCloud ». Manuscrit, http://manual.treecloud.com.

Guénoche A. & Garreta H. (2002). « Representation and Evaluation of Partitions ». In *Classification, clustering and data analysis, Proc. of IFCS'02*.

Hubert L. & Arabie P. (1971), « Comparing Partitions », *Journal of Classification* 2(1) : 193–218.

Huson D.H. & Bryant D. (2006). « Application of Phylogenetic Networks in Evolutionary Studies ». *Molecular Biology and Evolution* 23(2) : 254–267, logiciel disponible sur www.splitstree.org.

Luong X. (1989). *Analyse arborée des données textuelles*. CUMFID 16.

Mayaffre D. (2008). « Quand “travail”, “famille”, “patrie” co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence ». In Heiden

S., Pincemin B., éditeurs, *Actes des JADT'08*, pp. 811–822.

Rand W.M. (1971), « Objective criteria for the evaluation of clustering methods », *Journal of the American Statistical Association* 66(336) : 846–850.

Saitou N. & Nei M. (1987), « The neighbor-joining method: a new method for reconstructing phylogenetic trees », *Molecular Biology and Evolution* 4(4) : 406–425.