Similarités induites par mesure de comparabilité : signification et utilité pour le *clustering* et l'alignement de textes comparables

Pierre-Francois Marteau[1, 2] Gildas Ménier[1, 2]
(1) IRISA, UMR 6074
(2) Université de Bretagne Sud, 56000 Vannes
pierre-francois.marteau@univ-ubs.fr, gildas.menier@univ-ubs.fr

RÉSUMÉ

En présence de corpus comparables bilingues, nous sommes confrontés à des données qu'il est naturel de plonger dans deux espaces de représentation linguistique distincts, chacun éventuellement muni d'une mesure quantifiable de similarité (ou d'une distance). Dès lors que ces données bilingues sont comparables au sens d'une mesure de comparabilité également calculable (Li et Gaussier, 2010), nous pouvons établir une connexion entre ces deux espaces de représentation linguistique en exploitant une carte d'association pondérée ("mapping") appréhendée sous la forme d'un graphe bi-directionnel dit de comparabilité. Nous abordons dans cet article les conséquences conceptuelles et pratique d'une telle connexion similarité-comparabilité en développant un algorithme (Hit-ComSim) basé sur sur le principe de similarité induite par la topologie du graphe de comparabilité. Nous essayons de qualifier qualitativement l'intérêt de cet algorithme en considérant quelques expériences préliminaires de *clustering* de documents comparables bilingues (Français/Anglais) collectés sur des flux RSS.

ABSTRACT

Similarities induced by a comparability mapping : meaning and utility in the context of the clustering of comparable texts.

In the presence of bilingual comparable corpora it is natural to embed the data in two distinct linguistic representation spaces in which a "computational" notion of similarity is potentially defined. As far as these bilingual data are comparable in the sense of a measure of comparability also computable (Li et Gaussier, 2010), we can establish a connection between these two areas of linguistic representation by exploiting a weighted mapping that can be represented in the form of a weighted bidirectional graph of comparability. We study in this paper the conceptual and practical consequences of such a similarity-comparability connection, while developing an algorithm (Hit-ComSim) based on the concept of similarities induced by the topology of the graph of comparability. We try to evaluate the benefit of this algorithm considering some preliminary categorization or clustering tasks of bilingual (English/French) documents collected from RSS feeds.

MOTS-CLÉS : Graphe de comparabilité, Similarités induites, Documents comparables, Clustering.

KEYWORDS: Comparability graph, Induced similarities, Comparable documents, Clustering.

© ATALA

1 Introduction

La construction de corpus bilingues comparables (Déjean et Gaussier, 2002), notamment spécialisés ou thématiques, fait l'objet de travaux de recherche relativement intensifs depuis plus d'une dizaine d'années dans le but de pallier en partie la pénurie de corpus parallèles, coûteux à développer et à maintenir. Les applications sont multiples mais concernent principalement l'aide à la traduction, l'extraction de terminologie, la recherche d'information multilingue. Un corpus comparable est constitué de textes "similaires" exprimés dans au moins deux langues distinctes (EAGLES, 1996). Bien qu'il n'existe pas de définition claire et partagée de la notion de comparabilité, une mesure (quantifiable) de comparabilité a été proposée par (Li et Gaussier, 2010). Cette mesure dépendante d'un lexique bilingue de traduction, évalue la comparabilité entre deux documents de langues différentes en fonction du *prorata* de termes du premier document possédant au moins une traduction dans le deuxième document et *vice-versa*. Cette notion quantifiée de comparabilité permet de construire une relation pondérée entre deux corpus de langues distinctes que nous proposons d'étudier ici dans le contexte du *clustering* de textes comparables. L'un des enjeux est de préparer l'alignement des *clusters* comparables afin de faciliter la production de sous-corpus thématiques comparables (Li *et al.*, 2011).

2 Motivation

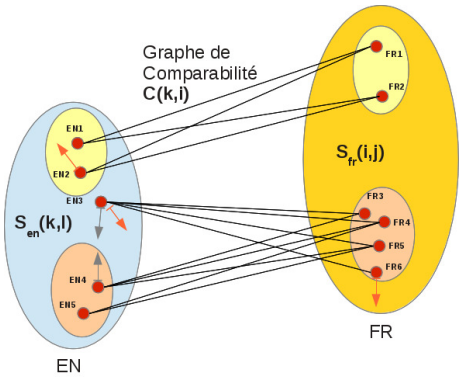

FIGURE 1 – Couplage de deux espaces de similarité par graphe de comparabilité

La figure 1 présente deux ensembles de documents *EN* (anglais) et *FR* (français) munis respectivement des fonctions de similarité $S_{en}(.,.)$ et $S_{fr}(.,.)$ et mis en relation par un graphe de comparabilité défini par la fonction de comparabilité $C(.,.)$. Les arcs de ce graphe sont bidirectionnels et pondérés par une valeur de comparabilité comprise dans l'intervalle $[0,1]$. L'idée principale que nous développons dans cet article est celle du renforcement de la similarité par la comparabilité : autrement dit, si deux documents du corpus *EN* sont comparables à un sous-ensemble de documents du corpus *FR* fortement similaires, alors leur similarité devrait être renforcée (et réciproquement). *A contrario*, si deux documents du corpus *EN* sont comparables à un sous-ensemble de documents du corpus *FR* faiblement similaires, alors leur similarité devrait décroître (et réciproquement). Ainsi, sur la figure 1, du point de vue de la similarité appréhendée

© ATALA

sous l'angle de la comparabilité, le document EN3 devrait s'éloigner du document EN2 pour se rapprocher des documents EN4 et EN5. De même, le document FR6 devrait s'éloigner des documents FR5,FR4 et FR3. L'utilité escomptée d'un tel renforcement/affaiblissement est une forme de filtrage du bruit inhérent aux modèles de représentation des documents caractérisés par un manque de connaissance (linguistique ou terminologique entre autres).

3 Modélisation : l'algorithme Hit-ComSim

Le modèle *bi-espace* (ici bilingue *EN* et *FR*) *d'induction de la similarité par la comparabilité* proposé repose sur un algorithme itératif décrit par les deux équations suivantes :

$$\begin{aligned} S^t_{en}(k,l) &= \sigma(k,l) \cdot \sum_{i=1}^{|FR|} \sum_{j=1}^{|FR|} C(k,i)C(k,j)S^{t-1}_{fr}(i,j)C(l,i)C(l,j) \\ S^t_{fr}(i,j) &= \sigma(i,j) \cdot \sum_{k=1}^{|EN|} \sum_{l=1}^{|EN|} C(k,i)C(l,i)S^{t-1}_{en}(k,l)C(k,j)C(l,j) \end{aligned} \quad (1)$$

– $C(k,i)$ est la comparabilité entre le $k^{\text{ième}}$ objet du corpus *EN* et le $i^{\text{ième}}$ objet du corpus *FR*
– $S^t_{en}(k,l)$ est la similarité entre les $k^{\text{ième}}$ et $l^{\text{ième}}$ objets du corpus *EN* à l'itération t
– $S^t_{fr}(i,j)$ est la similarité entre les $i^{\text{ième}}$ et $j^{\text{ième}}$ objets du corpus *FR* à l'itération t
– $\sigma(.,.)$ est une fonction de normalisation définie comme suit :

$$\sigma(k,l) = \left((\sqrt{\nu_{en}(k).\nu_{en}(l)} \right)^{-1}, \sigma(i,j) = \left((\sqrt{\nu_{fr}(i).\nu_{fr}(j)} \right)^{-1}$$

$$\nu_{en}(k) = \textstyle\sum_{i,j}(C(k,i) \cdot C(k,j))^2 \cdot S^{t-1}_{fr}(i,j), \quad \nu_{fr}(i) = \sum_{k,l}(C(k,i) \cdot C(l,i))^2 \cdot S^{t-1}_{en}(k,l)$$

L'initialisation de l'équation (3) pour $t = 0$ est effectuée par exemple à partir des matrices de similarités initiales S^0_{en} et S^0_{fr} calculées dans les ensembles *EN* et *FR* respectivement.

3.1 Convergence

La preuve de convergence (sous conditions en pratique satisfaisables) de l'algorithme présente peu d'intérêt et nous ne la détaillerons pas. Celle-ci est une conséquence du théorème de Perron-Frobenius (Perron, 1907) (Frobenius, 1912). Il en découle l'existence d'un **point fixe unique** pour cet algorithme ainsi qu'une vitesse de **convergence exponentielle**, particulièrement utile compte tenu de la complexité de l'algorithme (cf. paragraphe 3.3). Enfin, la convergence de l'algorithme est **indépendante des conditions initiales** (i.e. des matrices de similarités initiales).

3.2 Interprétation

Le graphe de comparabilité est représenté par la matrice de comparabilité C dont l'élément $C(k,i)$ représente la comparabilité entre le $k^{\text{ième}}$ document du corpus *EN* (anglais) et le $i^{\text{ième}}$ document du corpus *FR* (français). Cette matrice définit un graphe dont les nœuds représentent les documents et dont les arcs bidirectionnels sont pondérés par les éléments $C(k,i)$. A $t = 0$ les matrices de

© ATALA

similarités S^t_{en} et S^t_{fr} sont initialisées (conformément aux mesures de similarités propres aux espaces de représentation dans lesquels sont plongés les documents, si celles-ci existent, à défaut toute matrice positive symétrique de dimension adéquate convient). La notion de sous-ensemble de documents comparables communs (intersection) est appréhendée de manière "floue" par le biais des produits $C(k,i).C(l,i)$. Ainsi, pour $t > 0$, S^t_{en} (resp. S^t_{fr}) est mise à jour à partir de S^{t-1}_{fr} (resp. S^{t-1}_{en}). La convergence et l'existence d'un point fixe unique indépendant des conditions initiales donne une valeur particulière aux limites $\lim_{t\to+\infty} S^t_{en}$ et $\lim_{t\to+\infty} S^t_{fr}$ qui ne dépendent que du graphe de comparabilité caractérisé par la matrice C. On peut ainsi qualifier ces limites de ***similarités induites par une mesure de comparabilité***.

3.3 Complexité et simplification algorithmique

L'algorithme exploité directement dans la formulation proposée par l'équation (3) est en complexité $O(|EN|^2 \cdot |FR|^2)$ ce qui limite son exploitabilité aux petits corpus. Une simplification immédiate consiste à diminuer la complexité du graphe de comparabilité en limitant le nombre de connexions issues d'un document dans une langue donnée en ne considérant que le sous ensemble de documents qui lui sont les plus comparables dans le corpus de l'autre langue. La cardinalité des sous-ensembles de documents les plus comparables devient ainsi un paramètre de l'algorithme. Un deuxième paramètre régulant la connectivité du graphe est un seuil de comparabilité permettant d'élaguer les arcs du graphe associés à une comparabilité en dessous de ce seuil.

$$\begin{aligned} S^t_{en}(k,l) &= \sigma(k,l) \cdot \sum_{i,j \in NPC_{fr}(k) \cap NPC_{fr}(l)} C(k,i)C(k,j)S^{t-1}_{fr}(i,j)C(l,i)C(l,j) \\ S^t_{fr}(i,j) &= \sigma(i,j) \cdot \sum_{k,l \in NPC_{en}(i) \cap NPC_{en}(j)} C(k,i)C(l,i)S^{t-1}_{en}(k,l)C(k,j)C(l,j) \end{aligned} \quad (2)$$

où $NPC_{fr}(k)$ (resp. $NPC_{en}(i)$) est l'ensemble des indices des documents du corpus FR (resp. EN) les plus comparables au document k (resp. i)du corpus EN (resp. FR). Les coefficients de normalisation ν_{fr} et ν_{fr} sont ajustés en conséquence. En pratique, on peut limiter le nombre maximum de documents les plus comparables à une constante N_{pc}. L'algorithme simplifié caractérisé par l'équation (2) est en complexité inférieure à $O(max(|EN|,|FR|)^2 \cdot N^2_{pc})$. Cette simplification n'affecte que la complexité du graphe de comparabilité et n'a donc aucune incidence sur la convergence de l'algorithme, mais peut en avoir sur sa vitesse de convergence.

4 Expérimentations

Nos expérimentations (mis à part le cas d'école) portent sur des documents en langues anglaise et française collectés sur le web à partir de flux RSS mis à disposition par des quotidiens ou des chaînes de télévision. Nous avons exploité 12 sources pour le corpus EN pour un total de 1702 documents en langue anglaise et 11 sources pour le corpus FR, pour un total de 2466 documents en langue française.

Les documents collectés sont constitués en principe des champs *titre*, *description* (résumé), *date* et *texte*, ce dernier correspondant à la source citée par l'item RSS collecté. Les champs textuels

© ATALA

sont ensuite "*stemmés*", filtrés par l'utilisation d'une *stoplist* pour produire pour chaque document collecté un modèle de type *sac-de-mots* exploitant l'heuristique *tf-idf* (Spärck Jones, 1972) (bien entendu d'autres heuristiques de pondérations des mots réputées plus robustes sont possibles, par exemple l'heuristique BM25 (Robertson et Spärck Jones, 1976)). La similarité *cosinus* est utilisée dans les espaces *EN* et *FR* et le graphe de comparabilité est construit en exploitant la mesure de comparabilité définie par (Li et Gaussier, 2010) ainsi que le dictionnaire ELRA contenant 238742 couples de termes français/anglais. La comparabilité des corpus est faible puisque la mesure ne dépasse pas .35 pour toute paire de documents EN/FR. Nous avons utilisé un seuil d'*élagage* fixé à .2 en dessous duquel les liens de comparabilité sont considérés comme non existants.

La mesure de comparabilité exploitée fait intervenir un comptage du nombre des entrées lexicales *passerelles* permettant de *coupler* deux documents de langues distinctes via un lexique de traduction. Notons d_{en} un document en langue anglaise et d_{fr} un document en langue française. La mesure de similarité définie par (Li et Gaussier, 2010) se présente formellement sous la forme :

$$Cmp_{LG}(d_{en}, d_{fr}) = \frac{\sum_{w_1 \in Wd_{en} \cap WD_{en}} \sigma(w_1) + \sum_{w_2 \in Wd_{fr} \cap WD_{fr}} \sigma(w_2)}{|Wd_{en} \cap WD_{en}| + |Wd_{fr} \cap WD_{fr}|} \quad (3)$$

où : $Wd_i, i \in \{en, fr\}$ est le vocabulaire en langue $\mathscr{L}_i$ associé au document d_i ; WD_i est l'ensemble des entrées en langue $\mathscr{L}_i$ du dictionnaire bilingue utilisé présentes dans Wd_i ; $\sigma(w_i)$ est une fonction indicatrice qui prend la valeur 1 si au moins une traduction de l'entrée lexicale $w_i \in Wd_i$ en langue $\mathscr{L}_i$ existe dans le vocabulaire associé au corpus de l'autre langue, 0 sinon.

4.1 Cas d'école construit à la main

Nous reprenons ici l'exemple proposé en Figure 1 avec les données initiales ($C(k,i)$, S^0_{en},S^0_{fr}) et finales obtenues à l'issue de 4 itérations de l'algorithme décrit par l'équation (3) (S^4_{en}, S^4_{fr}) :

$$C = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}, S^0_{en} = \begin{pmatrix} 1. & .8 & .6 & .2 & 0. \\ .8 & 1. & .8 & .4 & .2 \\ .6 & .8 & 1. & .6 & .4 \\ .2 & .4 & .6 & 1. & .8 \\ 0. & .2 & .4 & .8 & 1. \end{pmatrix}, S^4_{en} = \begin{pmatrix} 1. & 1. & 0. & 0. & .0 \\ 1. & 1. & 0. & 0. & .0 \\ 0. & 0. & 1. & .8 & .6 \\ 0. & 0. & .8 & 1. & .8 \\ 0. & 0. & .6 & .8 & 1. \end{pmatrix},$$

$$S^0_{fr} = \begin{pmatrix} 1. & .8 & .3 & .3 & .2 & 0. \\ .8 & 1. & .5 & .5 & .3 & .2 \\ .3 & .5 & 1. & .8 & .8 & .7 \\ .3 & .5 & .8 & 1. & .8 & .8 \\ .2 & .3 & .8 & .8 & 1. & .8 \\ 0. & .2 & .7 & .8 & .8 & 1. \end{pmatrix} S^4_{fr} = \begin{pmatrix} 1. & 1. & 0. & 0. & 0. & 0. \\ 1. & 1. & 0. & 0. & 0. & 0. \\ 0. & 0. & 1. & .8 & .9 & .6 \\ 0. & 0. & .8 & 1. & 1. & .5 \\ 0. & 0. & .9 & 1. & 1. & .5 \\ 0. & 0. & .6 & .5 & .5 & 1. \end{pmatrix}$$

Les attentes initiales synthétisées en Figure 1 sont *a priori* bien satisfaites à l'issue des 4 premières itérations de l'algorithme, puisque le document *FR6* de l'ensemble *FR* s'éloigne en terme de similarité des documents *FR3*, *FR4* et *FR5*. De même, le document *EN3* de l'ensemble *EN* s'éloigne des documents *EN1* et *EN2* pour rejoindre le *cluster* constitué des documents *EN4* et *EN5*.

© ATALA

4.2 Expérimentations sur un petit corpus de 4 classes

Classe	#EN	#FR	code
Mali	10	10	o
Syria/Syrie	9	9	◇
gay	7	11	□
Beckham	4	3	⋆

TABLE 1 – Petits Corpus *EN* (30 documents) et *FR* (33 documents) catégorisables en quatre classes

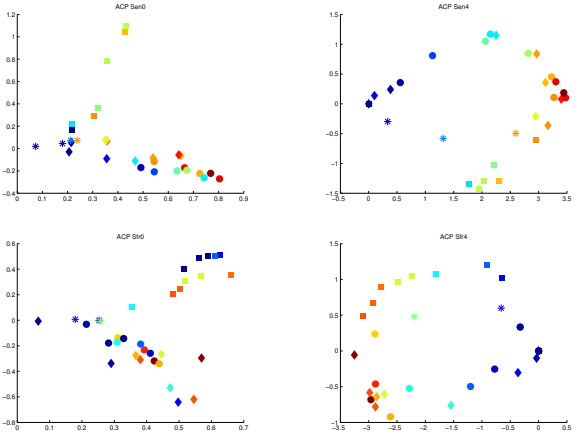

FIGURE 2 – ACP des matrices de similarités S^0_{en}, S^4_{en} (haut) et S^0_{fr}, S^4_{fr} (bas) sur le petit corpus

Cette expérimentation sur un nombre réduit de données a pour but d'illustrer le comportement de l'algorithme et ses potentialités, notamment sur des tâches de *clustering* et/ou d'alignement de *clusters*. Les corpus *EN* et *FR* sont constitués de documents extraits de flux RSS et sont catégorisés en quatre classes comme indiqué en table 1. Chaque classe est associée à un mot clé unique (Mali, Syrie/Syria, gay, Beckham) et chaque document possédant au moins une occurrence de mot clé correspondant à une classe est rattaché à cette classe. Un post-traitement a permis de vérifier que chaque document retenu du corpus n'appartient qu'à une seule classe. A partir des matrices de similarité initiales S^0_{en} et S^0_{fr}, quatre itérations de l'algorithme décrit par l'équation (3) sont appliquées pour obtenir les matrices de similarités consolidées par la comparabilité S^4_{en} et S^4_{fr}. Une analyse en composantes principales (ACP) pour chacune des quatre matrices est ensuite proposée pour visualiser les projections des documents sur les deux axes principaux. Un code de couleur est proposé pour quantifier la comparabilité moyenne des documents : rouge/forte, bleu/faible. Les résultats obtenus sont présentés en figure 2. L'effet sur le *clustering* est un éloignement des documents faiblement comparables et un rapprochement des documents comparables, tout en maintenant une proximité thématique. La comparabilité (articles de guerre) des *clusters Mali* et *Syria/Syrie* ressort dans les ACP des matrices produites par Hit-ComSim. Les axes principaux produits sur ces matrices semblent plus 'stables' que ceux produits sur les

© ATALA

matrices initiales, caractéristique qui, si cela est reproductible, pourrait s'avérer utile pour aligner des *clusters* ou des documents et isoler les données peu comparables.

4.3 Expérimentations sur un corpus plus large

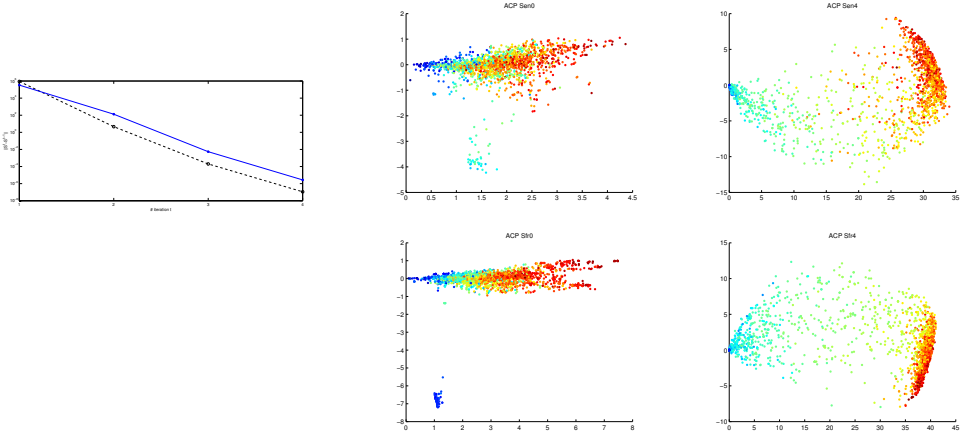

FIGURE 3 – Vitesse de convergence de l'algorithme (en haut à gauche) en échelle semi-log (en abscisse le nombre d'itérations et en ordonnée la norme des différences $||S_{en}^t - S_{en}^{t-1}||$ en bleu continu et $||S_{fr}^t - S_{fr}^{t-1}||$ en noir pointillé), et ACP pour le corpus plus large des matrices de similarités S_{en}^0, S_{en}^4 (en haut) et S_{fr}^0, S_{fr}^4 (en bas).

Cette expérience concerne l'ensemble des corpus EN (1702 documents) et FR (2466 documents) collectés. L'algorithme dans sa version simplifiée définie par l'équation (2) a été exploité sur 4 itérations avec une taille de voisinage de 32 (seuls les 32 plus proches voisins définissent la connectivité du graphe de comparabilité). La vitesse de convergence (exponentielle) de l'algorithme simplifié est présentée en figure 3 en haut à gauche sur une échelle semi-logarithmique. Les ACP effectuées sur les 4 matrices de similarité S_{en}^0, S_{en}^4, S_{fr}^0 et S_{fr}^4 sont également présentées en figure 3 en utilisant un code de couleur pour représenter la comparabilité moyenne associée à chaque document projeté sur les deux axes principaux (rouge/forte comparabilité, bleu/faible comparabilité). Cette mesure de comparabilité moyenne a du sens lorsque l'on envisage construire des corpus comparables. L'effet d'Hit-ComSim est ici encore une distribution des documents en fonction de leur comparabilité moyenne. Ainsi, les clusters isolés à comparabilité moyenne faible sont regroupés (autour de (0,0)) et séparés des documents à plus forte comparabilité moyenne distribués sur un arc de cercle centré en (0,0). Pour les matrices de similarité initiales, l'axe principal des ACP sont fortement corrélés à la comparabilité moyenne, mais des clusters à faible comparabilité moyenne sont isolés et justifient du 2ème axe principal.

© ATALA

5 Conclusion et perspectives

Nous avons proposé un algorithme (Hit-ComSim) pour étudier le couplage similarité/comparabilité, la comparabilité étant considérée sous la forme d'un graphe pondéré liant deux espaces de représentation d'objets distincts. Cet algorithme permet d'aboutir à la notion de *similarités induites par mesure de comparabilité* qui peuvent être aisément fusionnées aux similarités natives des espaces liés par comparabilité si celles-ci existent. La complexité élevée ($O(N^4)$) de l'algorithme peut être significativement simplifié en réduisant la connectivité du graphe de comparabilité. Par ailleurs, il est très possible que les matrices de similarité (les limites) produites par l'algorithme puissent être *approchées* de manière directe et calculatoire en $O(N^2)$. Au delà d'une curiosité algorithmique, les premières expériences relatives au *clustering* de données bilingues comparables montrent des potentialités utiles pour aligner des groupes de documents comparables et thématiquement cohérents. La confirmation de ces résultats préliminaires reste à établir par le biais d'expérimentations plus poussées et quantifiées, notamment pour mieux étudier l'impact des paramètres de l'algorithme, en particulier ceux qui déterminent la connectivité du graphe de comparabilité. L'enrichissement des mesures de comparabilité en tenant compte des fréquences d'occurrence et de traduction ou encore des relations sémantiques (synonymie, hyponymie et hyperonymie) induites (c.f. (Apidianaki et He, 2010) par exemple) est naturellement possible. D'autres tâches comme la *bi-classification* de documents bilingues sont également envisagés. Enfin, une approche complémentaire relative au *renforcement de la comparabilité par les similarités* peut être considérée.

Références

Apidianaki, M. et He, Y. (2010). An algorithm for cross-lingual sense clustering tested in a mt evaluation setting. *In Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT 2010)*, pages 219–226.

Déjean, H. et Gaussier, E. (2002). Une nouvelle approche a l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Numéro spécial, corpus alignés:1–22.

EAGLES (1996). Expert advisory group on language engineering standards guidelines : http ://www.ilc.pi.cnr.it/eagles96/browse.html. Rapport technique.

Frobenius, G. (1912). *Über Matrizen aus nicht negativen Elementen*. Reimer.

Li, B. et Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In COLING*, pages 644–652.

Li, B., Gaussier, E. et Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers - Volume 2*, HLT '11, pages 473–478, Stroudsburg, PA, USA. Association for Computational Linguistics.

Perron, O. (1907). Zur theorie der matrices. *Mathematische Annalen*, 64:248–263.

Robertson, S. E. et Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, n.3:129—-146.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

 © ATALA