

Quelques variations sur les mesures de comparabilité quantitatives et évaluations sur des corpus comparables Français-Anglais synthétiques

Guiyao Ke^{1, 2}

(1) IRISA, UMR 6074

(2) Université de Bretagne Sud, 56000 Vannes

guiyao.ke@univ-ubs.fr

RÉSUMÉ

Dans la suite des travaux de (Li et Gaussier, 2010) nous abordons dans cet article l'analyse d'une famille de mesures quantitatives de comparabilité pour la construction ou l'évaluation des corpus comparables. Après avoir rappelé la définition de la mesure de comparabilité proposée par (Li et Gaussier, 2010), nous développons quelques variantes de cette mesure basées principalement sur la prise en compte des fréquences d'occurrences des entrées lexicales et du nombre de leurs traductions. Nous comparons leurs avantages et inconvénients respectifs dans le cadre d'expérimentations basées sur la dégradation progressive du corpus parallèle Europarl par remplacement de blocs selon la méthodologie suivie par (Li et Gaussier, 2010). L'impact sur ces mesures des taux de couverture des dictionnaires bilingues vis-à-vis des blocs considérés est également examiné.

ABSTRACT

Some variations on quantitative comparability measures and evaluations on synthetic French-English comparable corpora

Following the pioneering work by (Li et Gaussier, 2010) we address in this paper the analysis of a family of quantitative measures of comparability dedicated to the construction or evaluation of comparable corpora. After recalling the definition of the comparability measure proposed by (Li et Gaussier, 2010), we develop some variants of this measure based primarily on the consideration of the occurrence frequency of lexical entries and the number of their translations. We compare the respective advantages and disadvantages of these variants in the context of an experiments based on the progressive degradation of the Europarl parallel corpus, by replacing blocks according to the methodology followed by (Li et Gaussier, 2010). The impact of the coverage of bilingual dictionaries on these measures is also discussed.

MOTS-CLÉS : Corpus comparables, Mesures de comparabilité, Évaluation.

KEYWORDS: Comparable corpora, Comparability measures, Evaluation.

1 Introduction

La notion de comparabilité entre documents est assez délicate à introduire : il est communément admis de considérer que deux documents de langues différentes sont comparables lorsque ces documents traitent de sujets analogues. Par extension, la notion de corpus comparable a été introduite par (Fung et Yee, 1998), (Munteanu *et al.*, 2004) et reste assez subjective. (Déjean et Gaussier, 2002) ont proposé une définition quantitative de cette notion de comparabilité selon laquelle : *Deux corpus de deux langues \mathcal{L}_1 et \mathcal{L}_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue \mathcal{L}_1 , respectivement \mathcal{L}_2 , dont la traduction se trouve dans le corpus de langue \mathcal{L}_2 , respectivement \mathcal{L}_1 .* (Li et Gaussier, 2010) en ont dérivé une mesure qui s'appuie sur un dictionnaire de traduction parallèle. Ces auteurs ont proposé d'évaluer cette mesure en partant de documents parallèles (c'est-à-dire de traduction directe) puis de dégrader cette traduction en observant la variation produite sur leur mesure : l'idée principale étant de vérifier la cohérence de la mesure proposée quand le nombre de traductions directes des entrées lexicales diminue. Cette mesure est principalement basée sur un comptage de présence de traductions des entrées lexicales qui dépend d'une manière non explicitée à la fois du dictionnaire de traduction et de la composition des corpus étudiés. Dans cet article nous proposons d'étudier et de comparer deux variantes autour de cette mesure de comparabilité en introduisant des informations quantitatives supplémentaires concernant le nombre d'occurrences des entrées lexicales et le nombre de traductions associées, en conjecturant que ces deux grandeurs produiront des effets positifs dans certaines situations. Ces nouvelles mesures sont présentées puis évaluées par rapport à la mesure développée par (Li et Gaussier, 2010), en prenant en considération la couverture du dictionnaire de traduction exploité.

2 Variations autour d'une mesure quantitative de comparabilité

2.1 Mesure de comparabilité de Li et Gaussier (Cmp_{LG})

Cette mesure fait intervenir un comptage du nombre des entrées lexicales passerelles permettant de *coupler* deux corpus de langues distinctes via un lexique de traduction. Notons C_1 un corpus en langue \mathcal{L}_1 et C_2 un corpus en langue \mathcal{L}_2 . La mesure de similarité définie par (Li et Gaussier, 2010) se présente formellement sous la forme :

$$Cmp_{LG}(C_1, C_2) = \frac{\sum_{w_1 \in WC_1 \cap WD_1} \sigma(w_1) + \sum_{w_2 \in WC_2 \cap WD_2} \sigma(w_2)}{|WC_1 \cap WD_1| + |WC_2 \cap WD_2|} \quad (1)$$

où : $WC_i, i \in \{1, 2\}$ est le vocabulaire en langue \mathcal{L}_i associé au corpus C_i ; WD_i est l'ensemble des entrées en langue \mathcal{L}_i du dictionnaire bilingue utilisé présentes dans WC_i ; $\sigma(w_i)$ est une fonction indicatrice qui prend la valeur 1 si au moins une traduction de l'entrée lexicale $w_i \in WC_i$ en langue \mathcal{L}_i existe dans le vocabulaire associé au corpus de l'autre langue, 0 sinon.

2.2 Enrichissement de la mesure LG

La mesure LG ne prend ni en compte le nombre d'occurrences des entrées lexicales dans les documents ni leurs nombres de traductions. Nous proposons ci-après deux variantes de la mesure LG qui font intervenir explicitement ces deux grandeurs en conjecturant que leur prise en compte produira dans certaines situations un effet positif.

2.2.1 Première variante : Cmp_{VA_1}

Cette première variante met en exergue de manière symétrique entre langue cible et langue source les trois éléments suivants : le nombre d'occurrences des entrées lexicales w pris dans le vocabulaire du corpus de la langue source, le nombre de leurs traductions dans le dictionnaire bilingue et la présence d'au moins une de leurs traductions dans le vocabulaire du corpus de la langue cible.

$$Cmp_{VA_1} = 1/2 \cdot (Cmp_{1,2}(C_1, C_2) + Cmp_{2,1}(C_1, C_2)) \quad (2)$$

où :

$$Cmp_{1,2}(C_1, C_2) = \frac{\sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{tf(w_1, C_1)}{\tau(w_1, WD_1)} \cdot \sigma(w_1) \right)}{\sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{tf(w_1, C_1)}{\tau(w_1, WD_1)} \right)}$$

$$Cmp_{2,1}(C_1, C_2) = \frac{\sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{tf(w_2, C_2)}{\tau(w_2, WD_2)} \cdot \sigma(w_2) \right)}{\sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{tf(w_2, C_2)}{\tau(w_2, WD_2)} \right)} \quad (3)$$

avec $tf(w_i, C_i)$ le nombre d'occurrences de l'entrée lexicale w_i dans le corpus C_i de la langue $i \in \{1, 2\}$; $\tau(w_i, WD_i)$ le nombre de traductions de l'entrée lexicale w_i du corpus C_i dans le dictionnaire WD_i . $\sigma(w_i)$ est défini comme précédemment.

2.2.2 Deuxième variante : Cmp_{VA_2}

Cette deuxième variante est très proche de la précédente, elle se distingue essentiellement sur la manière de symétriser la mesure.

$$Cmp_{VA_2}(C_1, C_2) = \frac{\sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{tf(w_1, C_1)}{\tau(w_1, WD_1)} \cdot \sigma(w_1) \right) + \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{tf(w_2, C_2)}{\tau(w_2, WD_2)} \cdot \sigma(w_2) \right)}{\sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{tf(w_1, C_1)}{\tau(w_1, WD_1)} \right) + \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{tf(w_2, C_2)}{\tau(w_2, WD_2)} \right)} \quad (4)$$

où $tf(w_i, C_i)$, $\tau(w_i, WD_i)$ et $\sigma(w_i)$ ont la même signification que précédemment.

3 Protocole d'évaluation

Nos expérimentations se sont focalisées sur les langues Anglaise et Française et suivent globalement le protocole proposé dans (Li et Gaussier, 2010). Ce protocole est construit sur le principe d'une dégradation progressive d'un corpus parallèle par remplacement déterministe par blocs de lignes. Nous avons complété ce protocole en développant une approche non-déterministe pour le remplacement des blocs afin d'évaluer l'impact de la procédure de remplacement des blocs sur la qualité observée des mesures.

3.1 Mesure d'évaluation

3.1.1 Référence empirique étalon

La référence empirique est construite sur la base du pourcentage de dégradation du corpus Europarl. Par exemple, si nous considérons 100 lignes par bloc, pour chaque bloc et pour chaque test, nous obtenons un vecteur de 101 valeurs (en partant de 0% de remplacement pour aboutir à 100% de remplacements). Nous obtenons ainsi une mesure de référence empirique, dite étalon, caractérisée par un vecteur (0%, 1%, 2%...100%) de $N = 101$ coordonnées.

3.1.2 Comparaison d'une mesure de comparabilité à la référence empirique

Pour établir le degré d'adéquation/d'inadéquation d'une mesure à la référence empirique, nous utilisons le coefficient de corrélation de Pearson. Celui-ci estime le degré de corrélation

entre une mesure de comparabilité X et la référence empirique Y de la manière suivante :

$$r_p = \frac{\sum_{n=1}^N (X_n - \bar{X}) \cdot (Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^N (X_n - \bar{X})^2} \sqrt{\sum_{n=1}^N (Y_n - \bar{Y})^2}} \quad (5)$$

Parmi d'autres estimateurs de corrélation, le coefficient de corrélation de Pearson est en général utilisé lorsque les variables X et Y sont supposées suivre des lois normales. En l'absence de contre indication particulière ce coefficient nous semble constituer ici un compromis acceptable.

3.1.3 Taux de couverture

Les taux de couverture du dictionnaire et des corpus sont des paramètres qui influencent grandement les mesures de comparabilité. Nous les définissons de la manière suivante :

- on définit le taux de couverture d'un dictionnaire D vis à vis du vocabulaire V associé à un corpus (i.e. ici à un bloc) par la quantité $T_D = \frac{|V \cap D|}{|V|}$.
- on définit le taux de couverture d'un vocabulaire V associé à un corpus (i.e. à un bloc) vis à vis d'un dictionnaire D par la quantité $T_V = \frac{|V \cap D|}{|D|}$.

3.2 Prétraitements et principes d'évaluation

3.2.1 Dictionnaires bilingues utilisés

Nous avons exploité deux dictionnaires bilingues dans le cadre de cette étude pour évaluer l'impact du dictionnaire de sa couverture sur les mesures de comparabilité.

Le premier dictionnaire référencé sous l'intitulé *fullDicText* est un dictionnaire propriétaire qui contient 74921 paires d'entrées lexicales français/anglais, se décomposant en 32767 d'entrées lexicales en langue anglaise, et 27511 d'entrées lexicales en langue française.

Le deuxième dictionnaire référencé sous l'intitulé *dicElra*, et disponible sous la référence ELRA-M0033, contient 243580 paires d'entrées lexicales en langues française et anglaise, se décomposant en 110541 entrées lexicales en langue anglaise et 109196 entrées lexicales en langue française.

3.2.2 Prétraitements

Nous disposons de deux corpus : un corpus parallèle «français-anglais Europarl corpus» (Koehn, 2005) et un corpus anglais «Associated Press corpus : AP». Ces corpus sont lemmati-

sés en exploitant le TreeTagger (Schmid, 1994) (Schmid, 2009) puis segmentés en phrases (une phrase par ligne). A l'issue de ce prétraitement, nous disposons ainsi de trois documents contenant chacun plusieurs millions de lignes : un document parallèle français EPE, un document parallèle anglais EPE et un document anglais AP.

3.2.3 Principes d'évaluation

En suivant les travaux de (Li et Gaussier, 2010), nous partitionnons le corpus parallèle Europarl en sélectionnant un nombre variable de lignes : 1000 lignes, 10000 lignes, 100000 lignes et 1428000 lignes (ce qui correspond à l'intégralité du corpus Europarl). Chaque élément de la partition obtenue est ensuite divisée en 10 blocs, chaque bloc contenant le même nombre de lignes (100 lignes, 1000 lignes, 10000 lignes, et 142800 lignes). Nous calculons ensuite les mesures au niveau des blocs alignés.

Nous proposons deux séries d'expérience qui se distinguent par le mode de remplacement : déterministe ou aléatoire. Pour chacune de ces séries, trois tests différents sont effectués selon les principes décrits ci-après. L'évaluation des mesures de comparabilité consiste à quantifier la corrélation entre leur décroissance observée et la décroissance attendue d'une mesure *empirique* quantifiant le degré de dégradation du corpus parallèle initial.

3.2.4 Remplacement déterministe

Pour le premier test, nous construisons les corpus référencés par *GAd* en remplaçant par permutation un certain nombre de lignes issues d'un bloc (le nombre de lignes est fonction du pourcentage de dégradation du corpus parallèle 0%, 1%... 100%) par le même nombre de lignes issues d'un autre bloc. La permutation des blocs est prédéfinie, par exemple : le bloc 1 <-> le bloc 6, bloc 2 <-> le bloc 7, etc.

Pour le deuxième test, nous construisons les corpus référencés par *GBd*, en remplaçant certaines lignes issues d'un bloc (le nombre de lignes est fonction du pourcentage de dégradation du corpus parallèle souhaité) par le même nombre de lignes extraites du document *AP*.

Pour le troisième test, nous construisons les corpus référencés par *GCd*, en remplaçant toutes les lignes d'un bloc par toutes les lignes d'un autre bloc, c'est-à-dire par exemple, le bloc 1 devient le bloc 6 et le bloc 2 devient le bloc 7, etc. A ce stade, et dans chaque bloc, un certain nombre de lignes (fonction du pourcentage de dégradation du corpus parallèle souhaité) sont remplacées par un même nombre de lignes extraites du fichier *AP*.

3.2.5 Remplacement aléatoire

Pour le premier test, nous construisons les corpus référencés par *GAa* en remplaçant aléatoirement selon une loi uniforme un certain nombre de lignes, en fonction du pourcentage de dégradation du corpus parallèle souhaité, par le même nombre de lignes extraites (sans remise pour garantir que les remplacements concernent systématiquement des lignes différentes) du reste des lignes non exploitées du corpus parallèle.

Pour le deuxième test, nous construisons les corpus référencés par *corpus GBa* en remplaçant aléatoirement selon une loi uniforme un certain nombre de lignes, en fonction du pourcentage de dégradation du corpus parallèle souhaité, par le même nombre de lignes extraites du document *AP*, en supprimant les lignes de remplacement déjà exploitées du document *AP*.

Pour le troisième test, nous construisons le corpus référencé par *corpus GCa*, en remplaçant d'abord toutes les lignes d'un bloc par le même nombre de lignes issues du complément du bloc dans l'ensemble des lignes du corpus Europarl (sans remplacement). Ensuite, au sein de chaque bloc, nous effectuons le remplacement aléatoire selon une loi uniforme d'un nombre de lignes donné (qui dépend du pourcentage de dégradation du corpus Europarl souhaité) par le même nombre de lignes extraites du corpus *AP* sans remplacement.

Ainsi, pour les deux séries de trois tests, le degré de comparabilité moyen décroît, en principe, de $GA_{d|a}$ à $GC_{d|a}$, en passant par $GB_{d|a}$.

4 Expérimentations

4.1 Influence de la taille des blocs sur les corrélations moyennes

Nous étudions ici les corrélations moyennes et leurs écarts-types entre les mesures de comparabilité et la référence empirique lorsque la taille des blocs exprimée en nombre de lignes varie dans l'ensemble $\{10^2, 10^3, 10^4, 10^5\}$.

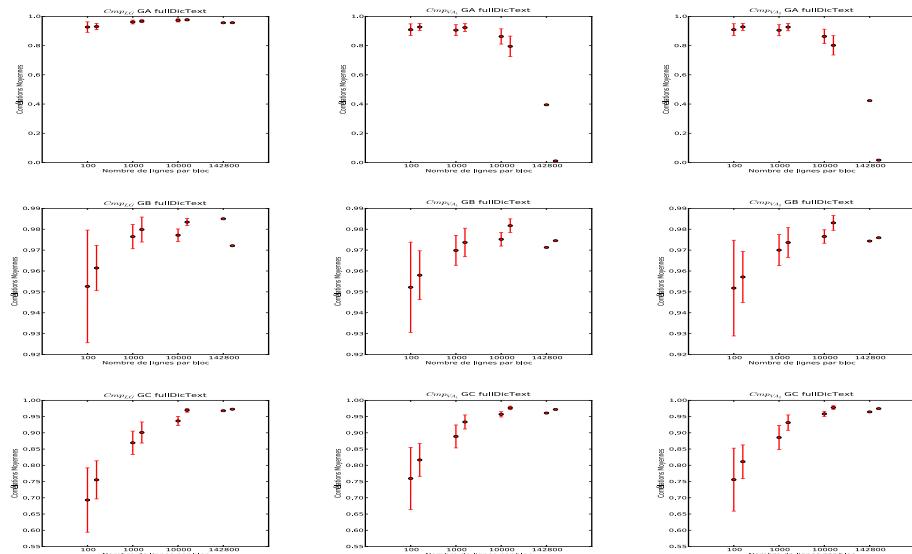


FIGURE 1 – Influence de la taille des blocs de corpus sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire bilingue *fullDicText*. Les deux modes de remplacement sont représentés pour chaque taille de bloc avec un léger décalage : déterministe à gauche et aléatoire à droite

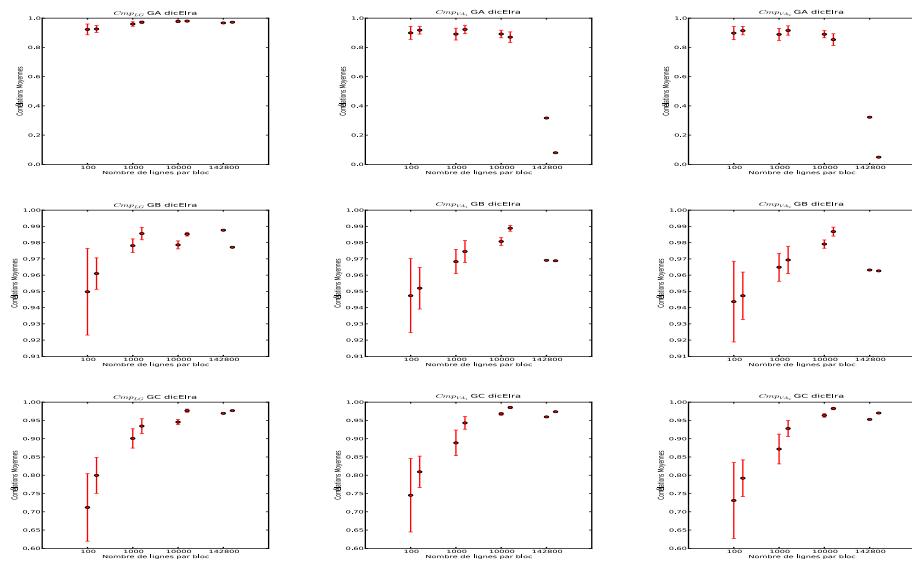


FIGURE 2 – Influence de la taille des blocs de corpus sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire bilingue *Elra*. Les deux modes de remplacement sont représentés pour chaque taille de bloc avec un léger décalage : déterministe à gauche et aléatoire à droite

Les figures 1 et 2 montrent, pour les deux modes de remplacement, que la mesure Cmp_{LG} est plus en adéquation avec la référence empirique au sens du coefficient de corrélation de Pearson sur les expériences *GA* que ses variantes Cmp_{VA_1} et Cmp_{VA_2} , en particulier pour des tailles de blocs importantes. Pour les expériences *GB*, les trois mesures atteignent quasiment le même niveau de corrélation vis-à-vis de la référence empirique. Enfin, sur les expériences *GC*, les deux variantes Cmp_{VA_1} et Cmp_{VA_2} semblent être légèrement plus robustes que Cmp_{LG} , principalement pour des tailles de bloc petites. Les deux dictionnaires bilingues utilisés conduisent à des résultats très voisins. Par contre, la procédure de remplacement aléatoire semble améliorer pour toutes les mesures et pour les deux dictionnaires la corrélation avec la référence empirique étalon, tant en moyenne qu'en écart type.

4.2 Influence des taux de couverture sur les corrélations moyennes des mesures avec la référence empirique

Nous étudions ici l'influence des taux de couverture (des dictionnaires et des vocabulaires en faisant varier la taille des blocs) sur les corrélations moyennes vis-à-vis de la référence empirique étalon obtenues sur la base des corpus dégradés par remplacement déterministe ou aléatoire, ceci pour les trois mesures Cmp_{LG} , Cmp_{VA_1} et Cmp_{VA_2} . Les figures 3 et 4 présentent ces corrélations moyennes pour les deux dictionnaires *fullDicText* et *dicElra* et pour les deux modes de remplacement, aléatoire et déterministe.

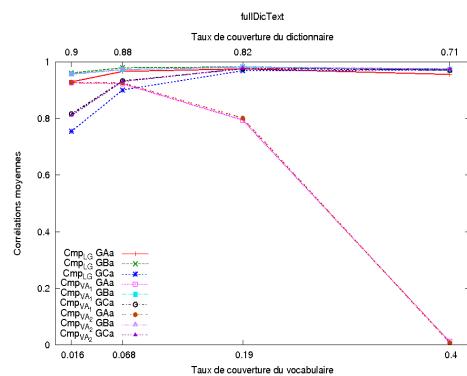
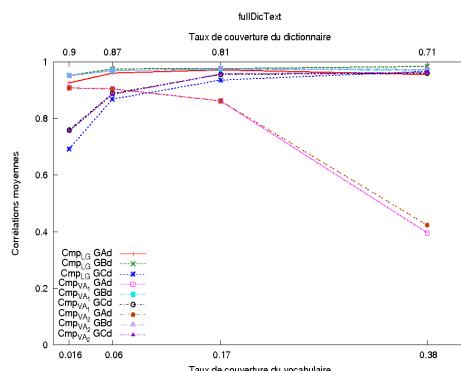


FIGURE 3 – Influence du taux de couverture sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire *fullDicText*, à gauche pour les corpus dégradés par remplacement déterministe, à droite pour les corpus dégradés par remplacement aléatoire

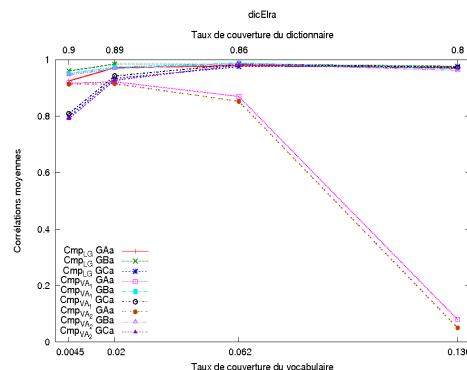
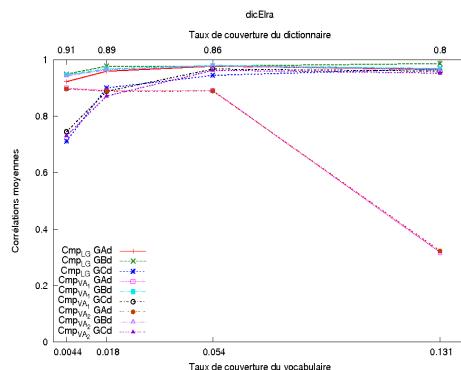


FIGURE 4 – Influence du taux de couverture sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire *dicElra*, à gauche pour les corpus dégradés par remplacement déterministe, à droite pour les corpus dégradés par remplacement aléatoire

On constate sur les figures 3 et 4 une meilleure corrélation moyenne pour la mesure Cmp_{LG} sur les corpus *GA*, tandis que les variantes Cmp_{VA_1} et Cmp_{VA_2} voient leurs corrélations s'effondrer sur ce même corpus lorsque le taux de couverture du dictionnaire croît. Sur les corpus *GB*, les trois mesures ont des performances très voisines, tandis que, sur les corpus *GC*, les deux variantes sont un peu mieux corrélées à la référence que la mesure Cmp_{LG} . Nous notons également une légère baisse en corrélation moyenne qui s'observe pour les trois mesures lorsque le taux de couverture du vocabulaire est très faible. Ces résultats sont analogues pour les deux dictionnaires *fullDicText* et *Elra* ainsi que pour les deux modes de remplacement déterministe et aléatoire.

4.3 Capacités des mesures à discriminer les degrés de dégradation du corpus parallèle Europarl

Afin de quantifier la capacité des mesures à discriminer les différents niveaux de dégradation du corpus parallèle Europarl au fur et à mesure des remplacements, que ceux-ci soient déterministes ou aléatoires, nous utilisons la mesure de discrimination suivante :

$$\Delta(i) = \frac{|\sigma_i + \sigma_{i+1} + 2 \cdot (m_i - \sigma_i/2 - (m_{i+1} + \sigma_{i+1}/2))|}{\sigma_i + \sigma_{i+1}} = \frac{2 \cdot |m_i - m_{i+1}|}{\sigma_i + \sigma_{i+1}} \quad (6)$$

où m_i et σ_i sont les moyennes et écarts types des valeurs de comparabilité associées aux niveaux (de 0%, 1%, ..., 100%) de dégradation du corpus Europarl indexés par $i \in \{1, \dots, 101\}$. En pratique, on observe que $\forall i, m_i \geq m_{i+1}$ et la valeur absolue n'est pas requise. $\Delta(i) \in [0, \infty]$ est d'autant plus grande que l'écart entre les comparabilités moyennes successives est grand et que la somme des écarts types associés est faible. Ainsi, plus la fonction $\Delta(i)$ est élevée, mieux le niveau i de dégradation du corpus est discriminé par la mesure de comparabilité.

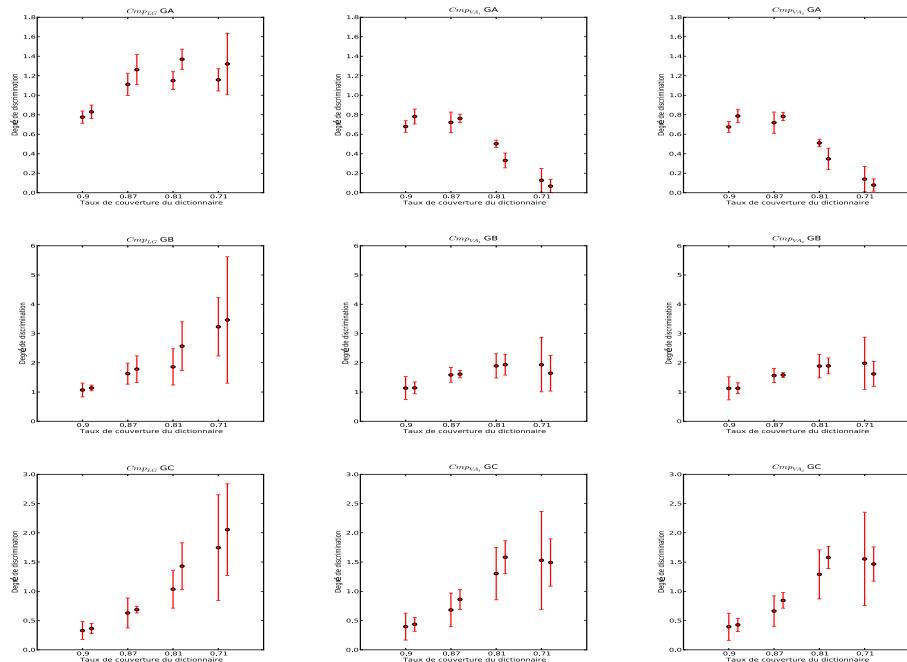


FIGURE 5 – Capacité des mesures de comparabilité à discriminer les degrés de dégradation du corpus Europarl : moyennes et écarts-types de $\Delta(\cdot)$ en fonction des taux de couverture du dictionnaire *fullDicText* exploité sur les corpus produits par remplacements déterministe (décalages à gauche) et aléatoire (décalages à droite).

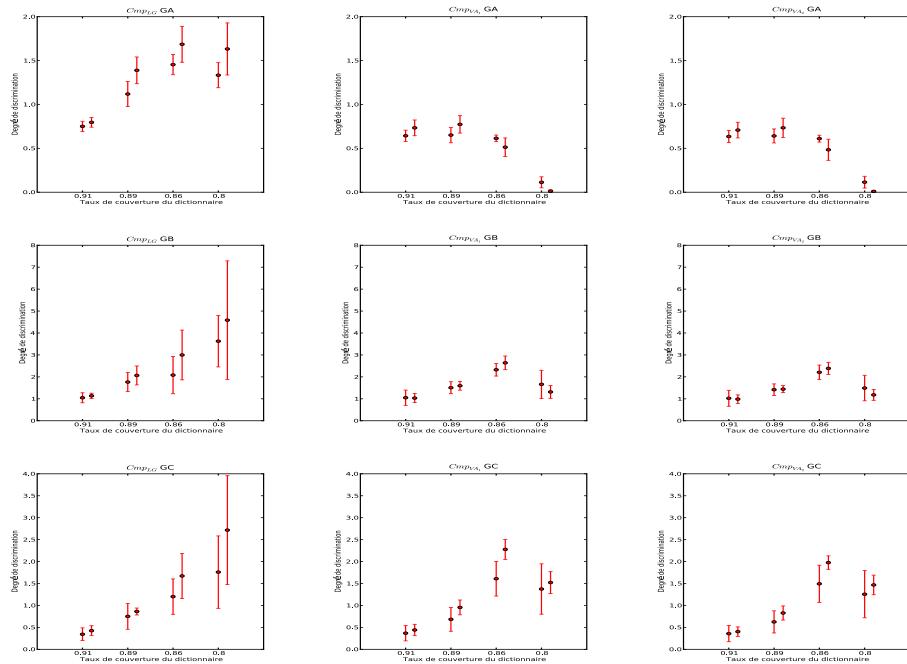


FIGURE 6 – Capacité des mesures de comparabilité à discriminer les degrés de dégradation du corpus Europarl : moyennes et écarts-types de $\Delta(.)$ en fonction des taux de couverture du dictionnaire *dicElra* exploité sur les corpus produits par remplacements déterministe (décalages à gauche) et aléatoire (décalages à droite).

Les figures 5 et 6 présentent pour les trois mesures Cmp_{LG} , Cmp_{VA_1} et Cmp_{VA_2} , sur les trois types de corpus (GA, GB et GC) la valeur moyenne et l'écart type de la mesure de discrimination Δ en fonction du taux de couverture des dictionnaire *fullDicText* et *dicElra* respectivement. Ici également, on constate que les variantes Cmp_{VA_1} et Cmp_{VA_2} sont moins discriminantes que la mesure Cmp_{LG} sur les corpus GA surtout pour les taux de couverture faible. Sur les corpus GB et GC, les mesures ont des niveaux de corrélation très voisins, surtout pour les taux de couverture les plus élevés du dictionnaire. Enfin, Sur les corpus GC, les variantes semblent légèrement plus robustes, notamment pour les taux de couverture élevés du dictionnaire. A noter que la capacité de discrimination moyenne augmente lorsque le taux de couverture du dictionnaire diminue dans la plupart des cas, mais sa variance augmente également en proportion également dans la plupart des cas.

5 Analyse et conclusions

Les résultats obtenus montrent que la mesure Cmp_{LG} et ses variantes Cmp_{VA_1} , Cmp_{VA_2} sont relativement voisines du point de vue de leur corrélation vis-à-vis de la mesure empirique étalon définie dans le contexte du protocole d'évaluation mis en œuvre. Il ressort néanmoins

que la mesure Cmp_{LG} est bien mieux corrélée à la mesure étalon sur les corpus les plus proches du corpus parallèle initial (Europarl) GAd et GAq , tandis que les variantes Cmp_{VA_1} , Cmp_{VA_2} sont légèrement plus robustes lorsque les mesures sont confrontées aux corpus GCd et GCa , les plus éloignés du corpus Europarl et sans doute les plus proches des corpus *bruités* tels que ceux constitués à partir de données collectées sur le Web par exemple. Sur les corpus intermédiaires GBd et GBa les trois mesures atteignent des niveaux de corrélation comparables vis-à-vis de la mesure empirique étalon.

Les dictionnaires ont un léger effet sur la corrélation entre nos deux variantes de comparabilité et la mesure empirique étalon : pour le dictionnaire *fullDicText*, Cmp_{VA_2} est légèrement mieux corrélée à la mesure étalon, tandis que pour le dictionnaire *dicElra*, c'est la variante Cmp_{VA_1} qui semble mieux corrélée.

Les degrés de corrélation de ces mesures augmentent lorsque le nombre de lignes par bloc augmente, en particulier pour le corpus *GC* (augmentation de plus de 20% entre la configuration 100 lignes par bloc et la configuration 142800 lignes par bloc). Par exemple, pour deux documents d'environ 100 lignes chacun, si la valeur de comparabilité est supérieure à 0,7, les deux documents sont probablement très comparables et pour deux documents de plus de 1000 lignes chacun, si la valeur de comparaison est supérieure à 0,8, les deux documents sont probablement comparables au même degré que les précédents. A l'appui de ce résultat, nous pouvons espérer proposer une référence raisonnablement stable pour la comparabilité des documents en fonction de leur nombre de phrases afin de juger si les documents sont suffisamment comparables ou non pour la tâche considérée.

Par ailleurs, les capacités des mesures à discriminer les niveaux successifs de dégradation du corpus parallèle que nous proposons est également un critère de comparaison intéressant nous semble-t-il. Sur ce critère, les tendances précédemment évoquées restent en vigueur. La mesure Cmp_{LG} se comporte mieux sur les corpus *GA* tandis que les variantes Cmp_{VA_1} et Cmp_{VA_2} semblent plus discriminantes sur les corpus *GC* et peut être également *GB* compte tenu des variances plus faibles observées sur ce critère pour les deux variantes.

Les modes de remplacement aléatoire ou déterministe semblent avoir un impact assez significatif au vu des résultats. Sur le corpus Europarl, le protocole déterministe de dégradation du remplacement proposé par (Li et Gaussier, 2010) engendre, en général, une baisse en moyenne des corrélations des trois mesures évaluées ainsi qu'un accroissement des écarts types, surtout sur les corpus s'éloignant du corpus parallèle Europarl (i.e. *GB* et *GC*). Cela amène à privilégier le mode de remplacement aléatoire par rapport au mode déterministe.

En matière de perspective, d'une part, nous allons essayer d'améliorer la précision lorsque le taux de couverture du dictionnaire est faible ; et d'autre part, nous allons exploiter et évaluer ces mesures de comparabilité dans le cadre d'expérimentations portant sur des réelles, en particulier sur des tâches de *bi-classification* et de *bi-clustering* de données thématiques bilingues.

Remerciements

Ces travaux ont été partiellement financés dans le cadre du projet ANR-08-CORD-009 METTRICC.

Références

- DÉJEAN, H. et GAUSSIER, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Numéro spécial, corpus alignés:1–22.
- FUNG, P et YEE, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.
- KOEHN, P (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.
- LI, B. et GAUSSIER, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING*, pages 644–652.
- MUNTEANU, D. S., FRASER, A. et MARCU, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pages 265–272.
- SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.
- SCHMID, H. (2009). TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.