# Validation de la notion de similarité textuelle dans un cadre multilingue

Romaric Besançon[1], Martin Rajman[1]

[1] EPFL - DI-LIA - INR (Ecublens) - CH-1015 Lausanne - Suisse

## Abstract

We propose in this contribution a method to validate the relevance of the notion of vector based semantic similarity, using a multilingual framework. The goal is to verify that vector based semantic similarities can be reliably transfered from one language to another. More precisely, the idea is to test whether the relative positions of documents in a vector space associated with a given source language are close to the ones of their translations in the vector space associated with the target language. The experiments have been conducted using the standard Vector Space model and the DSIR model. The method used for this test can also be applied to the automated evaluation of the performance of machine translation systems or to multilingual information retrieval.

## Résumé

Dans cet article, nous proposons une approche permettant d'utiliser un corpus multilingue aligné pour la validation de la pertinence de la notion de similarité sémantique s'appuyant sur une représentation vectorielle des documents. L'idée est de vérifier, à l'aide de tests statistiques, que les positionnements relatifs des documents dans l'espace vectoriel associé à la langue source sont proches de ceux de leur traduction dans l'espace vectoriel associé à la langue cible. Les expériences présentées ici reposent sur deux modèles de représentation vectorielle des documents : le modèle vectoriel standard et le modèle DSIR. La méthodologie de cette validation peut également être appliquée à l'évaluation automatique de la performance de systèmes de traduction automatique ou dans le cadre de la recherche documentaire multilingue.

**Keywords:** Représentation vectorielle, similarité textuelle, approche multilingue

## 1. Introduction

La notion de similarité textuelle vectorielle est très souvent utilisée dans les applications de Traitement de la langue destinées à l'exploitation de corpus textuels de grande taille. Par exemple, en Recherche Documentaire, les documents pertinents retournés par le moteur de recherche sont les plus proches de la requête selon une certaine mesure de similarité (Salton and McGill, 1983) ; de même, dans le cas de la structuration automatique de bases de données textuelles (classification non supervisée), les documents sont regroupés en classes en fonction d'une mesure de similarité particulière, reposant éventuellement sur une représentation vectorielle (Salton et al., 1975a).

La notion de similarité entre documents est évidemment fortement liée au choix de la méthode de représentation des textes. La représentation la plus utilisée est la représentation vectorielle (mise en œuvre en particulier dans les systèmes de recherche documentaire tels que SMART (Salton, 1971)), dans le cadre de laquelle un document est représenté par un vecteur dans un espace vectoriel dont les dimensions sont associées à des unités linguistiques spécifiques (mot, *stems*, lemmes, etc.)

Une telle représentation vectorielle des documents, fondée sur un modèle simple et facile à mettre en œuvre, a montré son efficacité dans le cadre spécifique de plusieurs applications (en particulier en recherche documentaire ou en classification). Nous proposons dans cet article une méthode de validation de la pertinence de la mesure de similarité fondée sur une représentation vectorielle, indépendante de toute application.

L'idée de cet article est d'utiliser un corpus multilingue aligné pour valider la pertinence de la représentation vectorielle des documents, en montrant que les positionnements relatifs des représentations des documents dans l'espace vectoriel associé à une langue source sont proches de ceux de leurs traductions dans l'espace vectoriel associé à la langue cible. En d'autres termes, on cherche à vérifier que la similarité entre les représentations vectorielles de deux documents quelconques dans l'espace source est proche de la similarité entre les représentations vectorielles de leur traduction dans l'espace cible.

La suite de cette contribution s'articule de la façon suivante : nous présentons dans la section 2 les deux modèles de représentation vectorielle utilisées pour ces expériences, à savoir le modèle vectoriel standard, et le modèle DSIR intégrant plus d'information sémantique au moyen des cooccurrences ; la méthode utilisée pour effectuer les tests de validation de la notion de similarité est ensuite présentée dans la section 3 ; les données utilisées pour les tests, ainsi que les résultats obtenus sont présentés dans la section 4 ; finalement, dans la section 5, nous décrivons des applications possibles de la méthode utilisée dans cet article à l'évaluation automatique des performances des systèmes de traduction automatique.

## 2. Représentation vectorielle des documents

Dans le cadre des modèles vectoriels considérés (le modèle vectoriel standard et le modèle DSIR), l'espace de représentation des documents est un espace vectoriel dont chaque dimension est associée à une unité linguistique particulière, appelée ci-après *terme d'indexation*. Nous détaillons dans les sous-sections suivantes les choix qui ont été faits pour ces expériences concernant la définition des termes d'indexation et les représentations correspondant aux deux modèles.

### *2.1. Choix des termes d'indexation*

Une série de pré-traitements automatiques a été mise en œuvre pour extraire l'ensemble des termes d'indexation, noté $T$, à partir d'un corpus bilingue (le corpus utilisé pour nos expériences est décrit plus en détails dans la section 4) :

- le corpus bilingue a été analysé par un lemmatiseur construit à l'aide de la boîte à outils logicielle SYLEX (Constant, 1995) : chaque mot a ainsi été associé à sa catégorie morpho-syntaxique et à son lemme ;
- deux lexiques ont été extraits (un pour chaque langue), contenant les lemmes des mots pleins (noms, verbes, adjectifs) ;
- un filtrage fréquentiel des termes d'indexation a été effectué sur la base de leur *fréquence en document* (*i.e.*le nombre de documents différents contenant un terme donné) ; l'ensemble d'indexation $T$ retenu a alors été l'ensemble des éléments des lexiques dont la fréquence en document est comprise dans l'intervalle de fréquences en documents $\left[\frac{|\mathcal{C}|}{100}, \frac{|\mathcal{C}|}{10}\right]$, où $|\mathcal{C}|$ est le nombre de documents dans le corpus. Cet intervalle est usuellement considéré comme adéquat pour fournir des termes avec un bon pouvoir discriminant (Salton et al., 1975b).

### *2.2. Représentation des documents dans le cadre du modèle vectoriel standard*

Dans le cadre du modèle vectoriel standard (VS), chaque document $d$ est représenté au moyen d'un vecteur $d^{VS} = (d_1^{VS}, \ldots, d_{|T|}^{VS})$, appelé *profil lexical*, dans lequel la $j^{\text{e}}$ composante $d_j^{VS}$ représente le poids (ou importance), dans le document $d$, du terme d'indexation $t_j$ associé à la $j^{\text{e}}$ dimension de l'espace vectoriel. D'une façon générale, cette mesure d'importance est le plus souvent une fonction de la fréquence du terme dans le document, intégrant de plus une pondération locale, une pondération globale et un facteur de normalisation (par rapport à la longueur du document). La fonction choisie pour nos expériences correspond au schéma de pondération *ltn* de SMART (Salton and Buckley, 1988; Singhal, 1997) :

$$d_j^{VS} = w_j = \text{idf} \times (1 + log(tf)) \tag{1}$$

où $tf$ est la fréquence du mot dans le document et $\text{idf}$ est le facteur de fréquence en document inverse $\text{idf} = \log \frac{1}{df}$, où $df$ est la fréquence en document du terme. Le facteur de pondération locale est $1 + log(tf)$ , le facteur de pondération globale est $\text{idf}$ (ce facteur permet d'accorder un poids plus important aux termes qui apparaissent moins fréquemment dans la collection et sont donc plus utiles pour la discrimination). Aucun facteur de normalisation n'est intégré directement dans cette fonction (une normalisation implicite est effectuée en utilisant la mesure de similarité du cosinus, indépendante de la norme).

### *2.3. Représentation des documents dans le cadre du modèle DSIR*

Le modèle DSIR est un modèle vectoriel permettant d'intégrer des informations sémantiques supplémentaires par l'utilisation de co-occurrences (Rajman and Bonnet, 1992; Rajman et al., 2000; Besançon, 2001).

Dans le cadre de ce modèle, les unités linguistiques $u_i$ considérées sont représentées par un vecteur $c_i = (c_{i1}, \ldots, c_{i|T|})$, appelé *profil de co-occurrence*, dont chaque composante $c_{ij}$ est la fréquence de co-occurrence de l'unité linguistique $u_i$ avec un terme d'indexation $t_j$. Un document $d$ est alors représenté comme la somme pondérée des profils de co-occurrence des unités linguistiques qu'il contient, c'est-à-dire par un vecteur $d^{DS} = (d_1^{DS}, \ldots, d_{|T|}^{DS})$ où chaque $d_j^{DS}$ est défini par :

$$d_j^{DS} = \sum_{u_i \in d} w_i \; c_{ij}$$

où la pondération $w_i$ est celle définie par l'équation (1), dans le cadre du modèle vectoriel standard.

Dans le cadre du modèle DSIR, les termes présents dans le document ne sont pris en compte que par le biais de leur profil de co-occurrence. Pour directement tenir compte de la présence d'un terme dans un document, un modèle DSIR hybride prenant en compte à la fois des occurrences et des co-occurrences des termes dans le document a également été proposé (Rungsawang, 1997; Rajman et al., 2000) :

$$d_j^{DS} = \alpha \; w_j + (1 - \alpha) \sum_{u_i \in d} w_i \; c_{ij} \tag{2}$$

## 3. Méthode de validation

L'idée de la méthode de validation pour vérifier la pertinence de la mesure de similarité dérivée d'une représentation vectorielle des documents repose sur l'hypothèse d'invariance suivante :

on considère que la similarité entre les représentations de deux documents (dans une langue donnée) est d'autant plus pertinente qu'elle est proche de la similarité entre les représentations des traductions des documents dans une autre langue.

Plus précisément, on cherche à tester dans quelle mesure le vecteur des similarités d'un document par rapport à un ensemble de documents de référence dans une langue source est proche du vecteur des similarités de sa traduction par rapport à l'ensemble des traductions des documents de référence dans une langue cible.

Considérons un corpus bilingue aligné composé des corpus $\mathcal{C}$ et $\mathcal{C}'$ respectivement dans les langue $L$ et $L'$. La procédure de test fonctionne alors de la façon suivante :
– on construit un corpus de test et un corpus de référence ;
– on représente les documents du corpus de test par leur positionnement par rapport aux documents du corpus de référence ;
– on compare ces représentations des documents du corpus test pour les deux langues.

**Construction des corpus de test et de référence** On extrait aléatoirement des corpus disponibles un nombre $n$ de documents ($n = 500$ dans nos expériences). Ces documents constituent le corpus de test, les $N = |\mathcal{C}| - n$ documents restants formant le corpus de référence.

On note :
– $d_i$ les documents du corpus de test TEST-$L$dans la langue source $L$, de taille $n$ ;
– $d'_i$ les documents du corpus de test TEST-$L'$dans la langue cible $L'$, de taille $n$ ;
– $D_i$ les documents du corpus de référence REF-$L$dans la langue source $L$, de taille $N$ ;
– $D'_i$ les documents du corpus de référence REF-$L'$dans la langue cible $L'$, de taille $N$ ;

La construction des corpus de test et de référence est synthétisée par la figure 1.

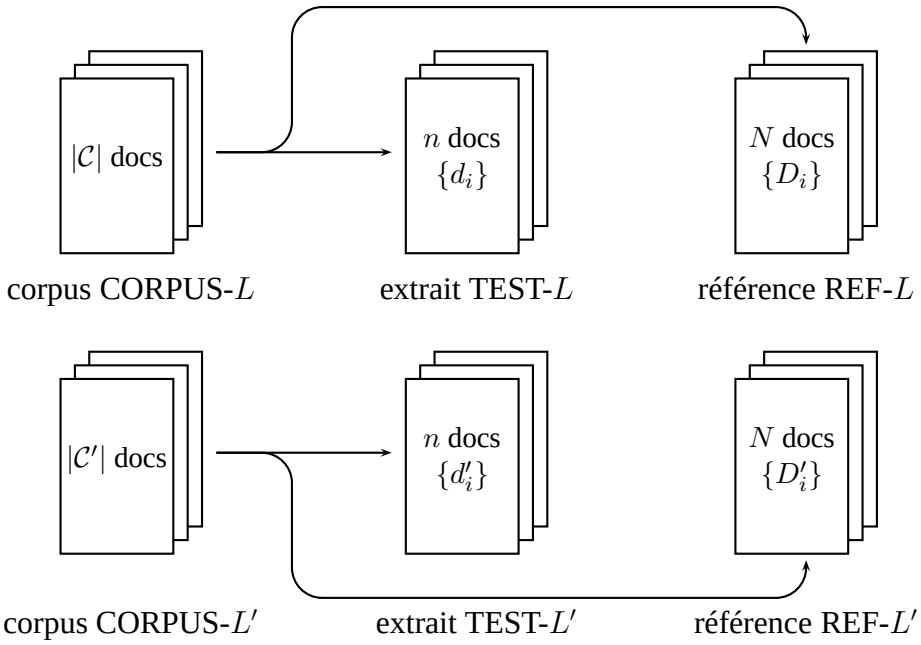


FIG. 1 – *Corpus de test et de référence*

**Représentation des documents** Tous les documents, ceux des corpus de test comme ceux des corpus de référence, sont représentés dans l'espace vectoriel associé à leur langue. On note que deux documents de langue différente ne sont donc pas directement comparables, car ils ne sont pas représentés dans le même espace.

À chaque document $d_i$ (resp. $d'_i$) du corpus de test TEST-$L$(resp. TEST-$L'$), on associe alors un vecteur des similarités dont la $j^{\text{e}}$ composante est la similarité de ce document avec le $j^{\text{e}}$ document du corpus de référence REF-$L$(resp. REF-$L'$).

Soit $V_s(d_i)$ (resp. $V_s(d'_i)$) ce vecteur des similarités, on a :

$$V_s(d_i) = (\delta(d_i, D_1), \ldots, \delta(d_i, D_N))$$
$$V_s(d'_i) = (\delta(d'_i, D'_1), \ldots, \delta(d'_i, D'_N))$$

où $\delta$ représente la similarité utilisée. Pour nos expériences, nous avons choisi la similarité du cosinus, définie, pour deux documents $d$ et $d'$, par :

$$\delta_{cos}(d, d') = \frac{d \cdot d'}{||d||\ ||d'||}$$

Le vecteur des similarités caractérise le positionnement relatif du document considéré par rapport aux documents du corpus de référence. Et, comme les corpus de référence sont alignés, les vecteurs des similarités ainsi obtenus sont, par construction, comparables entre eux, indépendamment de la langue. Ils peuvent de ce fait être utilisés pour tester l'invariance par traduction du positionnement d'un document par rapport au corpus de référence.

Cette représentation des documents de langues différentes par rapport à un corpus de référence aligné est proche de certaines méthodes utilisées pour la recherche documentaire multilingue, comme le modèle vectoriel généralisé (Carbonell et al., 1997; Yang et al., 1998), ou l'application du modèle *Latent Semantic Indexing* (LSI) à la recherche documentaire multilingue (Dumais et al., 1996; Littman and Jiang, 1998).

**Test d'invariance** Pour chaque document $d_i$ du corpus de test dans la langue $L$, on s'intéresse désormais à la similarité entre le vecteur des similarités $V_s(d_i)$qui lui est associé et chacun des vecteurs des similarités associés aux $n$ documents du corpus de test en langue $L'$. Ces similarités entre vecteurs des similarités sont également évaluées par la mesure du cosinus. L'idée, illustrée par la figure 2, est que si l'invariance par traduction est effectivement vérifiée, alors la similarité entre $V_s(d_i)$et $V_s(d'_i)$devrait être significativement plus importante que chacune des similarités entre $V_s(d_i)$et $V_s(d'_j)$ avec $j \neq i$.

Pour tester cette hypothèse, on compte le nombre de documents $d'_i$dans TEST-$L'$pour lesquels la similarité $\delta_{cos}(V_s(d_i), V_s(d'_j))$ est plus petite que la similarité $\delta_{cos}(V_s(d_i), V_s(d'_i))$, c'est-à-dire le nombre de documents $(d'_j)_{j\neq i}$ qui ont un positionnement relatif plus éloigné de celui de $d_i$ que de celui de $d'_i$.

Soit alors $f_i$ la proportion de ces documents dans TEST-$L'$, on a :

$$f_i = \frac{1}{n-1}\ |\{d'_k \in \text{TEST-}L' | k \neq i \text{ et } \delta_{cos}(V_s(d_i), V_s(d'_k)) < \delta_{cos}(V_s(d_i), V_s(d'_i))\}|$$

Pour vérifier que peu de documents en langue $L'$sont plus proches d'un document $d_i$ que sa traduction (selon la mesure choisie), on fait un test statistique sur cette proportion. Plus précisément, on souhaite vérifier que la proportion $f_i$ est significativement plus grande qu'un seuil

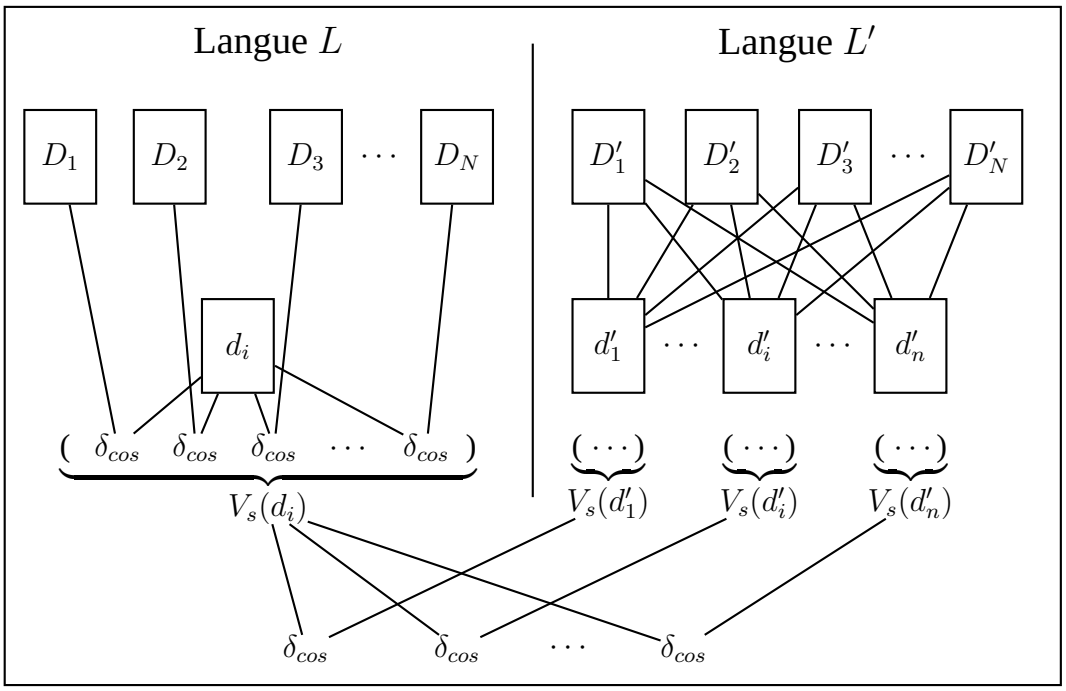


FIG. 2 – *Méthode pour la validation*

$p_0$ fixé *a priori*. Il s'agit alors de tester :

$$\begin{cases} H_0 \ : f_i = p_0 \\ H_1 \ : f_i > p_0 \end{cases}$$

Pour un échantillon assez grand (dans notre cas, $n = 500$), on rejette l'hypothèse $H_0$ au niveau $\alpha$ si (Grais, 1986, p. 261) :

$$\frac{f_i - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} > t_\alpha$$

où $t_\alpha$ est la valeur de la variable normale centrée réduite pour un pourcentage d'erreur $\alpha$. Cela revient à rejeter l'hypothèse $H_0$ si :

$$f_i > p_0 + t_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \tag{3}$$

## 4. Données et Résultats

### *4.1. Données*

Les documents utilisés sont extraits du corpus *JOC*, contenant les transcriptions des sessions de questions et de réponses du Parlement Européen, publiées en cinq langues dans le Journal Officiel de la Communauté Européenne.

Les langues retenues pour le test ont été le français et l'anglais et pour ces deux langues, les tests ont été fait dans les deux sens :
*fr-en* : {$L$=français, $L'$=anglais}
*en-fr* : {$L$=anglais, $L'$=français}.

Le corpus complet contient 6729 documents alignés dans les deux langues. Au sein de ces documents, nous avons sélectionné de façon aléatoire un corpus de test de $n = 500$ documents, les $N = 6229$ documents restant constituant donc les corpus de référence.

Les tailles des corpus en termes de nombre de mots, les tailles des lexiques (lemmes des noms, verbes et adjectifs) pour les corpus anglais et français, ainsi que les tailles des ensembles d'indexation (avec $[70, 700]$ comme intervalle de fréquence pour le filtre fréquentiel) sont données dans la table 1.

| | Corpus français | Corpus anglais |
|---|---|---|
| Nombres de mots dans le corpus | 1160877 | 1053945 |
| Taille du lexique (nombre de termes) | 25322 | 24469 |
| Taille de l'ensemble d'indexation | 1062 | 1102 |

TAB. 1 – *Taille des corpus, des lexiques et ensembles d'indexation*

### *4.2. Résultats*

Pour le test statistique, nous avons choisi $p_0 = 0.9$. L'hypothèse considérée est donc que la proportion de documents ayant un positionnement relatif plus proche que celui de la traduction est inférieur à 10%. Le taux d'erreur $\alpha$ a été quant à lui fixé à $\alpha = 2.5\%$, ce qui correspond à un $t_\alpha = 1.96$. Avec ces paramètres, l'équation (3) est donc :

$$f_i > 0.9 + 1.96\sqrt{\frac{0.09}{500}} = 0.9263 \tag{4}$$

et l'on peut vérifier que cela revient à accepter l'hypothèse $H_0$ s'il y a plus de 37 documents (sur les 500) qui sont « meilleurs » que la traduction elle-même.

Lors de nos expériences, nous avons produit les résultats suivants :
– le nombre de documents $N_R$ (parmi les 500) pour lesquels l'hypothèse $H_0$ est acceptée, *i.e.*les documents pour lesquels l'hypothèse d'invariance n'est pas vérifiée ;
– le nombre de documents $N_0$ (parmi les 500) pour lesquels la traduction est *le* document dont le positionnement relatif est le plus semblable à celui du document original ($f_i = 0$), *i.e.*les documents pour lesquels l'hypothèse d'invariance est parfaitement vérifiée ;

La table 2 présente les résultats obtenus d'une part avec une représentation vectorielle standard, en utilisant le schéma de pondération *ltn* de SMART et d'autre part avec une représentation vectorielle hybride du modèle DSIR, avec un coefficient d'hybridation $\alpha$, pour plusieurs valeurs de $\alpha$. La valeur $\alpha = 1$ correspond au modèle vectoriel standard. Les valeurs indiquées sont les moyennes obtenues sur 30 corpus de test extraits de manière aléatoire et indépendante.

L'hypothèse d'une proportion inférieure à 10% des documents est donc acceptée, en utilisant une représentation vectorielle standard, en moyenne pour plus de 99% des documents (99.4% pour le sens *fr-en*, 99.6% pour le sens *en-fr*). De façon plus précise, les résultats montrent que pour environ 95% des documents, la traduction est effectivement *le* document le meilleur (selon la mesure utilisée) parmi les 500 documents testés.

| | *en-fr* | | *fr-en* | |
|---|---|---|---|---|
| $\alpha$ | $N_R$ | $N_0$ | $N_R$ | $N_0$ |
| 1 (VS) | 2.03 (0.41%) | 476.47 (95.3%) | 2.9 (0.58%) | 473.7 (94.74%) |
| 0.9 | 1.67 (0.33%) | 476.3 (95.26%) | 2.57 (0.51%) | 474.4 (94.88%) |
| 0.8 | 1.47 (0.29%) | 477.5 (95.5%) | 1.97 (0.39%) | 475.77 (95.15%) |
| 0.7 | 1.23 (0.25%) | **478.67 (95.73%)** | 1.73 (0.35%) | 476.63 (95.33%) |
| 0.6 | **0.9 (0.18%)** | 478.33 (95.67%) | 1.4 (0.28%) | **476.7 (95.34%)** |
| 0.5 | 1 (0.2%) | 474.6 (94.92%) | **1.33 (0.27%)** | 473.37 (94.67%) |
| 0.4 | 4.9 (0.98%) | 449.83 (89.97%) | 3.2 (0.64%) | 452.1 (90.42%) |
| 0 | 352.43 (70.49%) | 40.8 (8.16%) | 344.87 (68.97%) | 52.1 (10.42%) |

TAB. 2 – *Résultats de la validation de la similarité*

Ces résultats permettent donc incontestablement de postuler que le positionnement relatif des documents les uns par rapport aux autres dans l'espace vectoriel est stable lors du passage d'une langue à l'autre. Cela est en soi une confirmation intéressante de la pertinence des mesures de similarité entre documents fondées sur une représentation vectorielle simple des documents.

Notons au passage que ce résultat n'était pas évident, en particulier parce que les lexiques ne sont pas alignés. En effet, Les deux lexiques (anglais et français) sont choisis selon la même procédure, en fonction des fréquences en documents des unités linguistiques sélectionnées, mais chacun uniquement en fonction du corpus dans la langue considérée. De ce fait, le terme associé à la $i^{\text{e}}$ dimension d'un espace vectoriel ne correspond pas à la traduction du terme associé à la $i^{\text{e}}$ dimension de l'autre espace vectoriel. En d'autres termes, bien que la représentation des documents de chaque langue ne dépende que du corpus de cette langue, la méthode de représentation aboutit à une représentation du corpus par un nuage de points qui est stable d'une langue à l'autre.

Les résultats obtenus sont positifs pour la représentation vectorielle standard, mais il est intéressant de noter que la représentation hybride du modèle DSIR permet d'améliorer encore la proportion de documents pour lesquels l'hypothèse est acceptée, pour arriver jusqu'à près de 99.8% de documents pour lesquels l'hypothèse est vérifiée (cf. figure 3). La valeur d'hybridation la meilleure semble être entre 0.5 et 0.6. Par contre, la représentation par le seul modèle DSIR ($\alpha = 0$) donne de très mauvais résultats : ceci peut être expliqué par le fait que la représentation par les seuls profils de co-occurrences a l'inconvénient de « lisser » les représentations des documents : au contraire du modèle vectoriel standard dans lequel les vecteurs sont très creux et les différences très marquées, les vecteurs de la représentation par le modèle DSIR sont des moyennes de profils de co-occurrences et ont des profils moins discriminants. Par contre, cette expérience montre qu'ils contiennent néanmoins des informations intéressantes pour la représentation, qui peuvent être capturées par l'utilisation du modèle hybride.

## 5. Applications

La technique présentée pour valider la notion de similarité textuelle fondée sur une représentation vectorielle des documents peut être directement utilisée pour estimer la performance des systèmes de traduction automatique, plus précisément pour établir une comparaison entre différents systèmes de traduction.

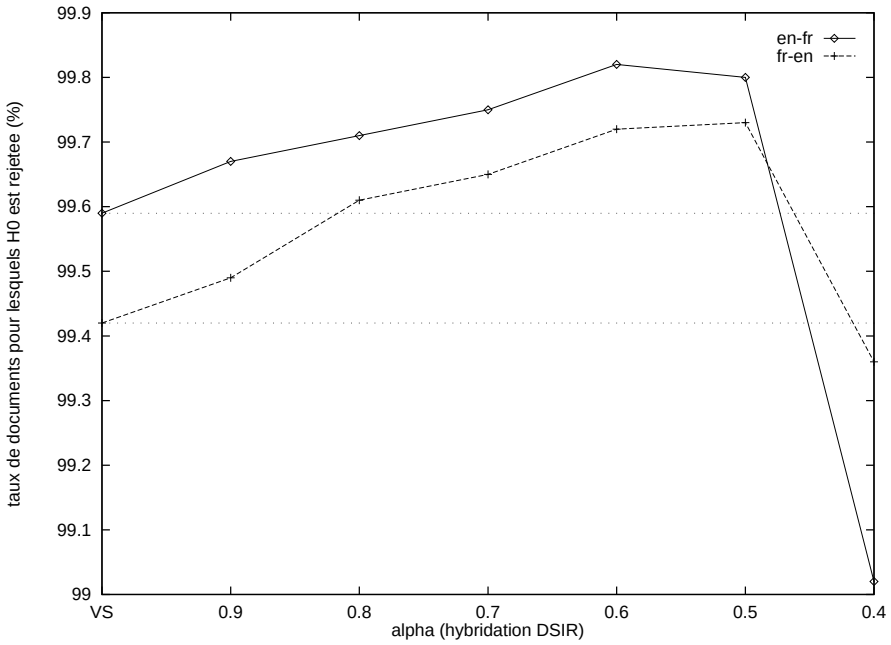


FIG. 3 – *Effet de l'hybridation avec le modèle DSIR sur la pertinence de la similarité textuelle dans un cadre multilingue*

Chaque système de traduction associe à un texte dans une langue source un texte traduit dans la langue cible. L'idée pour comparer la qualité sémantique des traductions (*i.e.*« est-ce que le texte traduit a conservé l'information présente dans le texte source ? ») est alors :

1. de calculer le vecteur des similarités caractérisant le positionnement du texte source par rapport à un corpus de référence (dans la langue source) ;
2. de calculer les vecteurs des similarités caractérisant les positionnements respectifs des traductions produites par les systèmes par rapport aux traductions des textes du corpus de référence ;
3. d'établir un classement selon la similarité entre ces vecteurs et le vecteur représentant le texte source.

On peut ainsi établir un classement permettant d'estimer (de façon automatique) les performances relatives des systèmes de traduction automatique. Cette technique a été utilisée dans (Rajman and Hartley, 2001), et elle a montré que, si la technique est assez peu fiable pour un document particulier, elle est robuste pour le classement moyen (*i.e.*sur l'ensemble des documents) et permet effectivement une évaluation des performances des systèmes de traduction, relativement bien corrélée avec les mesures d'adéquation (*adequacy*) et d'informativité (*informativeness*) produites par des experts humains (Carroll, 1966).

Par ailleurs, la méthode de représentation des documents par un vecteur des similarités indiquant leur positionnement par rapport à un corpus de référence étant indépendante de la langue, elle peut également être utilisée pour calculer des similarités entre documents de langue différente dans des applications destinées à l'exploitation de corpus multilingues, comme par exemple la recherche documentaire dans des corpus multilingues (Oard, 1997).

## 6. Conclusion

La notion de similarité textuelle, fondamentale pour de nombreux systèmes de traitement de la langue, reste encore à valider. Nous proposons ici une validation reposant sur l'invariance des positionnements respectifs des documents, représentés dans un espace vectoriel, pour des langues différentes. Cette validation est encourageante car elle confirme les hypothèses faites sur la pertinence de la similarité fondée sur une représentation vectorielle. La méthode de validation proposée peut également être utilisée pour comparer différents modèles de représentation : on a par exemple montré dans cet article que le modèle DSIR intégrant des co-occurrences dans la représentation vectorielle permet d'améliorer la pertinence de la similarité textuelle.

Toutefois, les résultats présentés ici sont des résultats quantitatifs globaux. Une étude plus approfondie de ces résultats devrait être entreprise, en particulier pour mettre en évidence de façon détaillée les propriétés des documents pour lesquels l'hypothèse nulle ne peut être rejetée, ainsi que l'apport de la représentation à l'aide du modèle DSIR pour ces documents.

D'autre part, cette validation a été effectuée seulement pour un corpus bilingue anglais–français. Des tests avec d'autres langues devraient également être envisagées pour vérifier la validité générale de la similarité, en particulier pour des langues assez éloignées.

Enfin, la technique utilisée pour effectuer cette validation trouve des applications pour la comparaison automatique des performances de différents systèmes de traduction automatique ou pour la recherche documentaire multilingue.

## Références

Besançon R. (2001). *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*. PhD thesis, École Polytechnique Fédérale de Lausanne.

Carbonell J. G., Yang Y., Frederking R. E., Brown R. D., Geng Y., and Lee D. (1997). Translingual information retrieval : A comparative evaluation. In *IJCAI (1)*, pages 708–715.

Carroll J. (1966). An experiment in evaluating the quality of translations. In Pierce. J. editor, *Language and machines. Report by ALPAC. NASNRC*, pages 67–75.

Constant P. (1995). *Manuel de développement SYLEX-BASE*. INGÉNIA-LN, Paris, France.

Dumais S., Landauer T., and Littman M. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, pages 16–23.

Grais B. (1986). *Méthodes Statistiques*. Dunod.

Littman M. and Jiang F. (1998). A comparison of two corpus-bases methods for translingual information retrieval. Technical Report CS-1998-11, Department of Computer Science, Duke University, Durham, North Carolina 27708-0129.

Oard D. W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*.

Rajman M., Besançon R., and Chappelier J.-C. (2000). Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2).

Rajman M. and Bonnet A. (1992). Corpora-base linguistics : new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.

Rajman M. and Hartley T. (2001). Automatically predicting mt systems rankings compatible with fluency, adequacy and informativeness scores. In *MT evaluation workshop, MT Summit VIII*, Siantiago de Compostela, Spain.

Rungsawang A. (1997). *Recherche Documentaire à base de sémantique distributionnelle*. PhD thesis, ENST, Paris.

Salton G. editor (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall.

Salton G. and Buckley C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24 :513–523.

Salton G. and McGill M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill.

Salton G., Wong A., and Yang C. S. (1975a). A vector space model for automatic indexing. *Communications of the ACM*, 18(11) :613–620.

Salton G., Yang C. S., and Yu C. T. (1975b). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1) :33–44.

Singhal A. (1997). *Term Weighting Revisited*. PhD thesis, Department of Computer Science, Cornell University.

Yang Y., Carbonell J. G., Brown R. D., and Frederking R. E. (1998). Translingual information retrieval : Learning from bilingual corpora. *Artificial Intelligence*, 103(1–2) :323–345.