# Extraction de synonymes à partir d'un corpus multilingue aligné

Jean-Luc Manguin, Tiedemann Jörg, Plas van Der

# EXTRACTION DE SYNONYMES À PARTIR D'UN CORPUS MULTILINGUE ALIGNÉ*

**Lonneke Van der Plas(1) et Jörg Tiedemann(1)**
**Jean-Luc Manguin(2)**
**(1) Université de Groningen**
**(2) CRISCO, CNRS & Université de Caen**

## 1 L'EXTRACTION AUTOMATIQUE DE SYNONYMES

L'extraction automatique de synonymes est une tâche relativement classique parmi les applications du traitement automatique des langues ; qui plus est, ses résultats et la qualité de ceux-ci vont souvent conditionner d'autres processus importants du domaine comme la fouille de textes, le résumé automatique, l'indexation de documents ou l'expansion de requêtes. Les différents procédés d'extraction ont en commun une recherche de similarités entre les unités du lexique à partir de leurs emplois dans un corpus[1]. En général, le corpus est constitué de textes, et les premières recherches ont été menées en traitant des productions spécialisées comme des comptes-rendus médicaux ou des notices techniques, tous rédigés dans la même langue. Dans ce cas, l'extraction des synonymes postule que les mots reliés sémantiquement partagent un certain nombre de contextes ; les méthodes les plus sophistiquées élaborent au préalable une analyse syntaxique des textes, ce qui permet le rapprochement de mots qui partagent non seulement le même co-texte, mais aussi la même fonction syntaxique (Bourigault, 2002). Néanmoins, et malgré tout l'intérêt des résultats de ces travaux, il faut bien reconnaître comme leurs auteurs que la moisson de synonymes s'avère relativement pauvre lorsque l'on s'attaque à des corpus généraux comme ceux constitués par plusieurs années de textes journalistiques (Bourigault & Galy, 2005). On remarque alors que les ensembles ainsi formés contiennent certes des synonymes, mais aussi des antonymes, des hyponymes et des hyperonymes, même en excluant les changements de catégorie grammaticale. Sans approfondir ces aspects d'une méthode qui n'est pas la nôtre, on peut toutefois mentionner que des travaux récents sur la polysémie verbale ont mis en évidence la flexibilité des « classes » formées à partir du partage de contextes (Jacquet & Venant, 2005). Ces mêmes travaux ont par ailleurs montré tout l'intérêt de passer d'une analyse globale sur tout un corpus, à une approche plus locale et dynamique pour mieux désambiguïser les énoncés et constituer des classes d'objets[2] qui seraient cette fois plus homogènes.

Signalons ici une variante importante et prometteuse de la méthode : elle opère un changement de corpus en s'attachant à traiter non plus des textes, mais les définitions de

* Pour faire référence à cet article : Manguin Jean-Luc, Tiedemann Jörg & Van der Plas Lonneke, « Extraction de synonymes À partir d'un corpus multilingue aligné », revue électronique *Texte et corpus*, n°3 / août 2008, Actes des Journées de la linguistique de Corpus 2007, p. 151-161 (disponible sur http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_vanderplas_tiedemann_manguin.pdf)

[1] Le terme « corpus » doit ici être entendu dans un sens large, si on l'applique à ce qui suit.

[2] « Objet » peut ici être pris dans son sens syntaxique.

dictionnaires généraux. Les résultats obtenus avec le lexique verbal, appliquant des filtrages pertinents, conduisent à des rapprochements synonymiques, voire métaphoriques tout à fait intéressants et dénués d'anomalies grossières (Gaume, 2006).

Cela dit, nous allons maintenant nous pencher sur une autre voie, qui emploie des corpus multilingues, et dont les fondements diffèrent sensiblement des méthodes précédentes.

# 2 L'EXPLOITATION DE DONNÉES MULTILINGUES

L'extraction de synonymes à partir de données (ou de corpus pris au sens large du terme) multilingues part d'un autre postulat que la méthode qui traite des données monolingues ; en effet, la similarité ne va plus se mesurer ici à l'aune du partage de contextes, mais à celle du partage de traductions. En d'autres termes, l'idée sous-jacente est que si deux mots sont souvent traduits de la même manière dans de nombreuses langues, il y a une forte probabilité pour qu'ils soient synonymes. Par exemple, les mots « accroissement » et « augmentation » sont traduits en suédois par « ökning » et en anglais par « increase » ou « rise ». Ces traductions communes, ajoutées à toutes les autres qu'ils partagent, indiquent que ces deux mots ont des sens très proches. Si l'on veut illustrer la comparaison entre le traitement monolingue et le traitement multilingue, on peut dire par exemple que dans une même langue, les mots « ville » et « région » vont partager un grand nombre de contextes verbaux, comme « aménager » en tant qu'objet et « se situer » en tant que sujet, ce qui conduira à les considérer comme proches[3]. Notre méthode, en traitant des données multilingues, ne place « région » qu'au 26ème rang des mots proches de « ville », simplement parce que ces deux mots partagent très peu de traductions.

Cela dit, avant de discuter de nos résultats, il convient de préciser quels types de données multilingues peuvent être traités afin d'extraire des relations synonymiques. En effet, par analogie avec la méthode monolingue, il est possible d'extraire les synonymes à partir soit de corpus textuels, soit de dictionnaires multilingues. Les corpus textuels multilingues doivent cependant, comme on peut le deviner, être constitués par un même texte traduit en plusieurs langues différentes ; à vrai dire, une version bilingue d'un même texte peut suffire, mais la qualité des résultats s'améliore quand on passe à une version multilingue, comme l'ont montré les premiers travaux (Van der Plas & Tiedemann, 2006). D'autre part, si le travail à partir de dictionnaires multilingues (ou de plusieurs dictionnaires bilingues) a précédé celui sur corpus textuels (Lin *et al.*, 2003), le principe reste le même, c'est-à-dire que les mots qui partagent le plus de traductions seront considérés comme les plus proches. Mais en outre, le fait de travailler sur corpus a plusieurs avantages ; citons-en deux qui nous sont utiles : l'accès aux fréquences des traductions (ce qui permet de pondérer les traitements), et la possibilité de restreindre le corpus à un domaine particulier (par exemple les notices techniques).

[3] « région » est en effet le premier voisin de « ville » trouvé par la méthode de D. Bourigault qui nous sert de comparaison monolingue.

# 3 LE CORPUS UTILISÉ

## 3.1 Description

Pour notre travail, nous avons choisi un corpus parallèle multilingue avec une condition concernant sa taille : celle-ci doit en effet être suffisamment grande pour permettre au travail d'extraction de produire assez de résultats que l'on pourra commenter par la suite. Notre choix s'est donc porté sur le corpus Europarl (Koehn, 2003). Ce corpus parallèle librement disponible sur Internet provient des actes du Parlement Européen entre mars 1996 et septembre 2003, et inclut les versions en 11 langues européennes : français, italien, espagnol, portugais, anglais, néerlandais, allemand, danois, suédois, grec et finnois. Chaque langue comprend environ 1 million de phrases, qui contiennent de l'ordre de 28 millions de mots.

En outre, il faut préciser ici que l'ensemble du corpus est aligné phrase par phrase, car cette opération est nécessaire avant de passer à l'alignement mot à mot (autrement dit à la recherche des traductions). Cet alignement phrase à phrase emploie l'algorithme conçu par Gale & Church (1993).

## 3.2 Alignement mot à mot

La seconde étape d'alignement fait passer à un niveau supérieur de précision, puisqu'il s'agit de réaliser un alignement mot à mot à l'intérieur des phrases déjà alignées. Il faut tout d'abord préciser que cela n'est possible que pour les phrases qui s'alignent une à une, comme l'indique la notice de l'outil utilisé ; en effet, il arrive dans les corpus parallèles qu'une phrase soit traduite en deux phrases distinctes, ou à l'inverse que deux phrases soient regroupées en une seule dans la traduction.

Pour réaliser cet alignement, nous avons employé un outil Open Source nommé GIZA++, élaboré par F.J. Och (2003) et basé sur les principes de traduction statistique automatique de Brown. Dans ce modèle de traduction, lorsque le programme doit aligner une phrase dans deux langues différentes, il construit des liens directionnels entre les mots de la phrase dans une des langues (qu'il va considérer comme source) et ceux la version dans l'autre langue (qui sera la cible) ; puis, le programme va inverser le rôle des langues, la cible devenant la source et vice versa. Au final, les mots des deux versions de la phrase seront reliés par des liens mono- ou bidirectionnels. C'est ici qu'interviennent le paramétrage du logiciel d'alignement et le choix d'une stratégie par l'utilisateur. Dans notre cas, nous avons choisi de privilégier la précision des résultats, et d'opter pour ce que le concepteur de GIZA++ nomme la stratégie « d'intersection » ; celle-ci consiste à ne conserver que les liens bidirectionnels, ce qui réduit notablement le nombre de paires proposées à l'alignement.

Nous pouvons aussi préciser que nous n'avons traité que du texte « brut », autrement dit sans marquage d'aucune sorte, et que le finnois a été éliminé de la liste des langues en raison de problèmes d'alignement.

# 4 L'EXTRACTION PROPREMENT DITE

Avant de procéder à la mise en ordre des données pour commencer l'extraction, nous avons traité l'ensemble des paires retenues à l'étape précédente par un lemmatiseur de la langue source (ici le français), en l'occurrence Treetagger. Les mots pour lesquels nous cherchons des synonymes seront ainsi mis sous une forme canonique semblable à celle du

dictionnaire de référence, ce qui facilitera les comparaisons lors de la phase d'évaluation des résultats. En outre, nous seront à même de choisir la catégorie des mots étudiés.

## 4.1 Les vecteurs caractéristiques

Le principe de notre méthode est d'établir une similarité entre des vecteurs ; chaque mot de la langue source, repéré dans le corpus, va être doté d'un vecteur caractéristique dont les composantes sont les traductions de ce mot dans les langues cibles. Les coordonnées du vecteur sont alors égales aux fréquences de ces traductions, comme dans l'exemple ci-après, où figure le mot français « automne » :

| | *herfst_NL* | *outono_PT* | *autumn_EN* | *fall_ENG* |
|---|---|---|---|---|
| *automne* | *102* | *92* | *75* | *67* |

*Tableau 1 : Les traductions du mot français « automne »*

## 4.2 Les traitements

Pour calculer la similarité entre deux vecteurs, autrement dit entre deux mots de la langue source, nous utilisons l'indice de Dice « pondéré », dont la formule est donnée ci-dessous (Curran & Moens, 2002) :

$$Sim(mot1, mot2) = \frac{2\sum min\ (poids(mots1,trad), poids(mot2,trad))}{\sum\ (\ poids(mot1,trad) + poids(mot2,trad)\ )}$$

La fonction de pondération que nous avons choisie est l'information mutuelle spécifique, dont la formule bien connue est :

$$I(mot,trad) = \log \frac{P(mot,trad)}{P(mot)P(trad)}$$

Nous donnons ci-après un exemple de calcul à partir des données suivantes :

| | *river_EN* | *fiumel_IT* | *rio_PT* | *rivier_NL* |
|---|---|---|---|---|
| *fleuve* | *102* | *111* | *107* | *82* |
| *rivière* | *74* | *69* | *62* | *55* |

*Tableau 2 : Les traductions des mots français « fleuve » et « rivière »*

Les mots « fleuve » et « rivière » sont souvent proches et substituables ; pour calculer l'information mutuelle spécifique entre « fleuve » et « river », les probabilités présentes dans la formule de l'information mutuelle spécifique seront assimilées aux fréquences divisées par le nombre de paires de mots anglais – français ; s'il y a par exemple 10 000 paires, alors le numérateur sera égal à 102 / 10 000. Autrement dit, l'information mutuelle

sera toujours calculée pour la paire de langues considérée. Il faut ajouter que nous pouvons mettre une condition de fréquence qui filtre ces vecteurs, et imposer ainsi un seuil sur la somme de la « ligne », et un autre sur chaque « cellule ». Cette donnée de filtrage interviendra dans la présentation des résultats.

À l'issue de tous ces traitements, les mots de la langue source sont pourvus d'un certain nombre de mots proches, avec chacun une similarité comprise entre 0 et 1. Par exemple, voici les résultats pour le mot « accident » (pour des valeurs de similarité supérieures à 0.015) :

> accident : (0.172) catastrophe, (0.172) incident, (0.134) naufrage, (0.110) désastre, (0.103) malheur, (0.096) sinistre, (0.089) tragédie, (0.068) drame, (0.056) événement, (0.051) calamité, (0.044) épisode, (0.033) catastrophique, (0.032) désastreux, (0.031) cataclysme, (0.025) hasard, (0.023) ravage, (0.022) lésion, (0.022) dommage, (0.019) blessure, (0.018) catastropher, (0.018) route, (0.018) mégarde, (0.017) malheureux, (0.016) fléau, (0.016) affaire, (0.015) blessé, (0.015) tort, (0.015) débâcle, (0.015) dégât

# 5 L'ÉVALUATION DES RÉSULTATS

Même si le simple examen de l'exemple ci-dessus donne une idée de la qualité des résultats, il importe d'effectuer une évaluation aussi objective que possible des sorties proposées par ce traitement automatique.

## 5.1 La méthode

L'évaluation que nous proposons comprend deux volets : l'estimation chiffrée de la qualité des résultats, et la comparaison avec une méthode basée sur l'analyse distributionnelle de données monolingues dans la même langue source. En effet, grâce à Didier Bourigault que nous remercions d'avoir mis ses données à notre disposition, nous possédons toutes les paires rapprochées par l'outil Upery à partir des textes des dix années du journal *Le Monde* (200 millions de mots), du moins toutes celles qui ont une similarité supérieure à 0.1.

Afin de comparer ces résultats à ceux de notre méthode, nous avons choisi tout d'abord de traiter une liste d'environ 1 000 noms, c'est pourquoi nous avons fixé un seuil à 0.16 pour les paires issues des données monolingues ; en éliminant les paires de similarité inférieures, seuls 950 noms communs[4] restent ainsi dotés de mots apparentés.

Nous avons ensuite extrait des résultats de notre traitement multilingue les mots proches attribués à ces 950 noms ; enfin, pour placer les deux méthodes sur un pied d'égalité, nous avons filtré tous les résultats au moyen d'un dictionnaire de noms communs avant de commencer les mesures[5].

## 5.2 Les mesures utilisées

Nous avons désormais à notre disposition deux listes d'environ 1 000 mots, chacun de ceux-ci étant assorti d'une liste de mots proches et munis d'une valeur de similarité ; nous ferons sur ces deux listes trois mesures, en effectuant un filtrage « passe-haut » sur ces

[4] Pour être précis, il s'agit ici de noms « arguments », si l'on se réfère à la distinction opérée par D. Bourigault.

[5] On notera au passage que nous avons donné pour « accident » le résultat AVANT ce filtrage, afin de montrer certaines erreurs que l'on pourra discuter par la suite.

valeurs de similarité, c'est-à-dire que seuls seront retenus les mots ayant une valeur supérieure ou égale à celle du seuil.

Les trois mesures sont classiques : il s'agit de la couverture, de la précision et du rappel. La couverture se calcule simplement : pour une valeur du seuil, c'est le nombre de mots pourvus d'une liste non vide, divisé par le nombre de mots dans la liste de référence (ici 950). On s'attend bien entendu à ce que cette couverture diminue quand le seuil augmente.

Les deux autres mesures exigent l'emploi d'une référence en matière de synonymes ; de la même manière que D. Bourigault et E. Galy avaient évalué leurs propres résultats selon un point de vue un peu différent, nous avons opté pour le dictionnaire électronique des synonymes du CRISCO (que nous noterons par la suite DES) comme référence en matière de synonymes. À partir de là, nous pouvons calculer la précision et le rappel pour chaque mot auquel le traitement a attribué un ou des mots proches. Ainsi, pour un mot, la précision est la proportion de synonymes considérés comme « bons » par le DES parmi les mots proches fournis par le système. Le rappel est ce même nombre de « bons » synonymes fournis, divisé cette fois par le nombre total de synonymes donnés par le DES pour le même mot. Pour une valeur donnée du seuil, nous ferons ensuite la moyenne des précisions et rappels pour tous les mots munis de mots proches[6].

## 5.3 Les résultats

Nous donnons ci-après les courbes de variation de la couverture, de la précision et du rappel en fonction du seuil de similarité, et ce pour les deux types de traitement (multilingue et monolingue). Les résultats présentés ici emploient un seuil de 10 pour les lignes et de 4 pour les cellules.

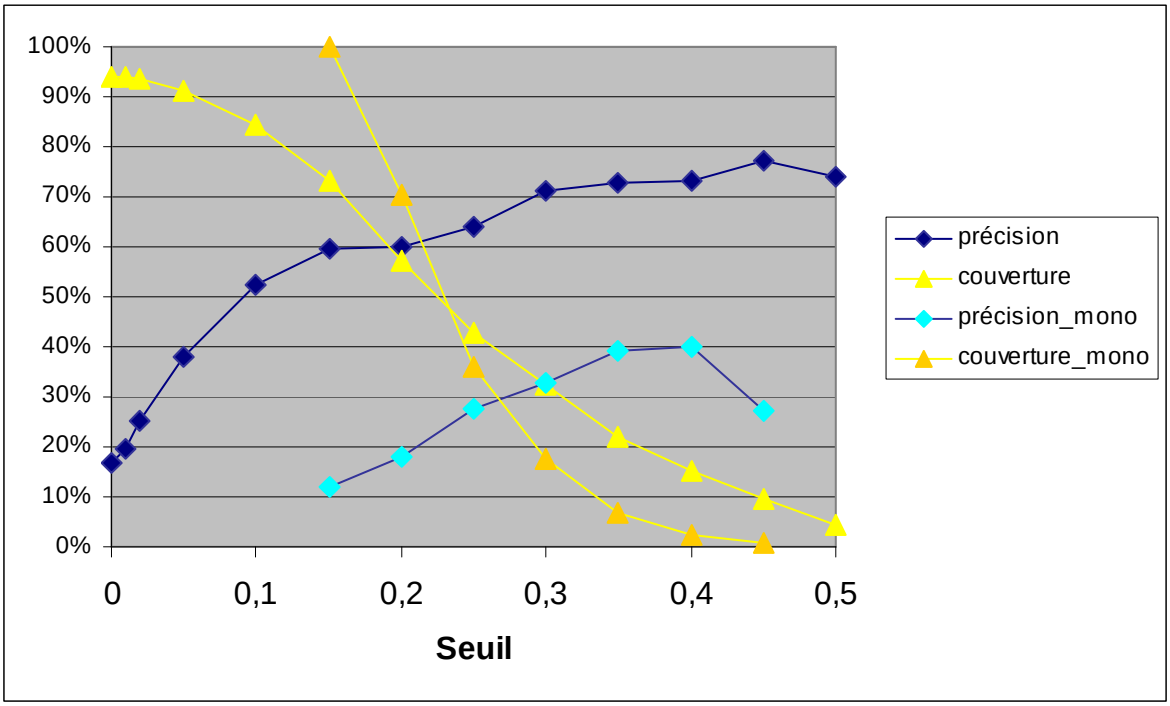
Concernant les données monolingues, nous n'avons pas de mesure pour les valeurs de similarité inférieures à 0.16, car c'est au moyen de cette valeur que nous avons déterminé notre liste de 950 noms ; d'autre part, au-delà de 0.45, la couverture du système monolingue est nulle.

D'une manière générale, la couverture des deux systèmes décroît quand le seuil de similarité augmente, mais cette décroissance est moins rapide dans le cas du système monolingue. On peut ainsi comprendre que le système multilingue est moins « productif » que le système monolingue, en ce sens qu'il rapproche moins facilement les mots entre eux. Ceci peut s'expliquer d'abord par la taille du corpus (la partie française ne représente que 28 millions de mots) et ensuite par le domaine qui est plus restreint qu'un ensemble de textes journalistiques. D'autre part, on peut remarquer que la couverture n'atteint jamais 100% (elle vaut ici 94% si l'on prend en compte tous les mots proposés), et l'abaissement des seuils à 2 par cellule et 2 par ligne ne la fait passer qu'à 97%.

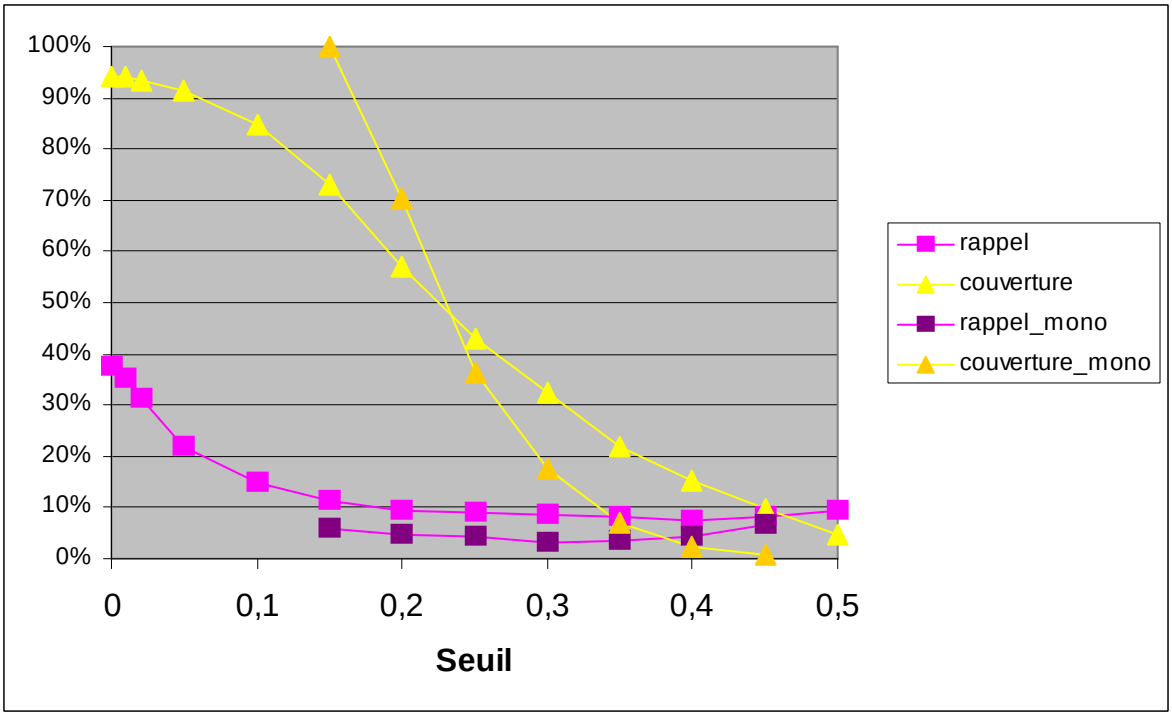
Les courbes de la précision mettent clairement en évidence la qualité des synonymes produits. En effet, si l'on compare les deux méthodes pour une même couverture de 50% par exemple, la méthode multilingue produit des synonymes avec une précision de l'ordre de 60 à 65%, tandis que la méthode monolingue n'atteint que 25% ; en outre, les valeurs maximales manifestent elles aussi la qualité de la production. Il est à noter que cette précision ne varie quasiment pas (de l'ordre de 1%) quand on passe à des seuils valant 2 par cellule et par ligne. Enfin, comme nous le verrons dans la discussion, ces valeurs de

[6] En d'autres termes, cela signifie que si la couverture est de 400 mots sur 950 pour une valeur du seuil, la précision et le rappel sont calculés sur ces 400 mots.

précision peuvent même être revues à la hausse de quelques pourcents, puisque parmi les synonymes proposés, certains peuvent être tout à fait acceptables alors qu'ils ne sont pas mentionnés dans le dictionnaire de référence.




*Graphique 1 : Couverture et précision pour les deux méthodes d'extraction*




*Graphique 2 : Couverture et rappel pour les deux méthodes d'extraction*

Les chiffres du rappel apportent un peu moins de satisfaction, même si la méthode multilingue reste encore dans ce domaine supérieure à la méthode monolingue ; il est en effet assez décevant de voir que la valeur maximale atteinte n'est que de 38%, et que l'abaissement des seuils pour les vecteurs caractéristiques ne la fait grimper qu'à 44%. Dans tous les cas, lorsque le seuil de similarité franchit la valeur 0.1, le rappel se situe déjà sous les 20%, ce qui est tout de même peu. Cela dit, nous verrons dans les paragraphes 5.5 et 5.6 que cette relative déception s'explique par d'autres phénomènes liés à la substitution, et qu'elle est compensée par d'autres satisfactions.

## 5.4 Problèmes rencontrés

Reprenons l'exemple des synonymes proposés pour le mot « accident » :

> accident : (0.172) catastrophe, (0.172) incident, (0.134) naufrage, (0.110) désastre, (0.103) malheur, (0.096) sinistre, (0.089) tragédie, (0.068) drame, (0.056) événement, (0.051) calamité, (0.044) épisode, (0.033) catastrophique, (0.032) désastreux, (0.031) cataclysme, (0.025) hasard, (0.023) ravage, (0.022) lésion, (0.022) dommage, (0.019) blessure, (0.018) catastropher, (0.018) route, (0.018) mégarde, (0.017) malheureux, (0.016) fléau, (0.016) affaire, (0.015) blessé, (0.015) tort, (0.015) débâcle, (0.015) dégât

Nous avons fait figurer en vert les synonymes donnés par le DES ; à côté de ceux-ci, nous trouvons en noir des synonymes acceptables ou des mots proches, sur la présence desquels nous reviendrons dans les paragraphes suivants, car par exemple « sinistre » semble tout à fait approprié comme synonyme d'« accident ». Enfin, nous avons marqué en rouge les erreurs évidentes produites par le traitement. À l'exception de « catastropher » qui est probablement une lemmatisation erronée de « catastrophe », ces erreurs proviennent de problèmes d'alignement, car le caractère collocatif de ces mots avec « accident » est relativement manifeste.

En outre, le rapprochement de « mégarde » avec « accident », bien que correct selon la référence qui inclut des sens anciens, est sans doute le fruit des traductions communes aux deux locutions « par mégarde » et « par accident » ; un autre exemple de ce phénomène nous est donné par les synonymes proposés pour « majorité » :

> majorité : (0.330) majoritaire, (0.325) plupart, (0.234) majoritairement, (0.151) majeur, (0.079) gros, (0.064) largement, (0.058) partie, (0.058) écrasant, (0.053) essentiel, (0.052) essentiellement, (0.050) principalement, (0.050) nombre, (0.049) grand, (0.049) unanimité, (0.040) consensus, (0.036) large, (0.035) unanime, (0.033) partiellement, (0.032) généralement, (0.032) section, (0.032) fraction,…

On y remarque encore des collocatifs comme « large » ou « écrasant », mais surtout des adverbes synonymes de la locution « en majorité » comme « majoritairement », « largement » ou « principalement ». On peut ainsi imaginer que la détection et le marquage des locutions avant le processus d'alignement pourrait améliorer la précision des résultats.

## 5.5 Compléments sur la précision

Comme nous l'avons dit plus haut, la précision obtenue est tributaire dans une certaine mesure de la référence choisie, en l'occurrence le DES. En allant plus loin, l'examen des synonymes proposés par le traitement automatique que nous avons employé nous révèle la présence de relations souvent pertinentes, mais que le dictionnaire de référence considère comme « mauvaises », puisqu'elles ne font pas partie des relations synonymiques directes.

Cependant, comme le montrent les exemples dans le tableau ci-dessous (Tableau 3, qui donne en même temps la similarité calculée), la qualité des relations est manifeste :

| | | |
|---|---|---|
| affection | pathologie | 0,377 |
| affrontement | confrontation | 0,589 |
| aggravation | détérioration | 0,339 |
| ampleur | taille | 0,321 |
| assemblée | hémicycle | 0,446 |
| avancée | progrès | 0,498 |
| barrière | entrave | 0,367 |
| carence | déficit | 0,314 |
| collaborateur | employé | 0,309 |
| compensation | indemnisation | 0,317 |
| composant | ingrédient | 0,387 |
| dédommagement | indemnisation | 0,543 |
| démarrage | lancement | 0,353 |
| documentaire | reportage | 0,302 |
| expert | spécialiste | 0,322 |
| fiscalité | imposition | 0,366 |
| gens | population | 0,312 |
| honnêteté | sincérité | 0,419 |
| poulet | volaille | 0,343 |
| refonte | remaniement | 0,333 |
| suppression | élimination | 0,376 |

*Tableau 3 : Exemples de relations non prévues par le dictionnaire de référence*

L'apparition de relations comme « affection : pathologie » ou « documentaire : reportage » et leur rejet par le dictionnaire de référence soulève ici le problème de la fiabilité des références ; en effet, même si le DES résulte de la fusion de sept dictionnaires, ce n'est pas pour autant que les lexicographes ont fixé dans leurs ouvrages les sens les plus récents, même si les dictionnaires généraux les admettent. L'extraction automatique pourrait donc révéler des emplois synonymiques nouveaux, pourvu que le corpus soit assez proche dans le temps.

## 5.6 Compléments sur le rappel

La valeur relativement faible du rappel pose également le problème du contenu du dictionnaire de référence comme nous allons le voir dans ce qui suit, mais elle s'explique aussi par un phénomène lié à la substitution dans un énoncé. En principe, des synonymes sont des substituts, autrement dit au sein d'un même énoncé, l'un peut remplacer l'autre et vice versa, sans que le sens de l'énoncé soit modifié de manière significative. Malheureusement, les dictionnaires de synonymes proposent souvent des termes dont la substitution au mot choisi s'accompagne d'un marquage ou d'un changement de niveau de langue ; et dans ce cas, il est compréhensible que la traduction tienne compte de ces

modifications collatérales. C'est pourquoi, dans le traitement multilingue, certains synonymes argotiques ou familiers ne sont pas rappelés. Par exemple, les synonymes de « chaussure » donnés par le DES sont[7] :

*chaussure : bas, basket, bateau, botte, caoutchouc, chaussette, pompe, soulier, spartiate, tennis*

En dehors des synonymes comme « bas » ou « chaussette » qui ne sont sans doute pas les meilleurs, il est évident que « bateau » ou « pompe » sont des substituts (très) familiers, et qu'ils partageront rarement leurs traductions avec un terme aussi neutre que « chaussure ». Ainsi le traitement automatique que nous avons décrit attribue à « bas » des synonymes comme « baisse » ou « chute », à « bateau » des synonymes du même champ que « navire », et parvient à associer « pompe » à « machine » aussi bien qu'à « cérémonie » ; mais il ne réussit pas à les détecter en tant que synonymes de « chaussure ».

*Voyons un autre exemple ; le DES donne comme synonymes d'« arabe » les mots suivants :*

*arabe : arabesque, arabique, bédouin, beur, maure, nedjdi, sarrasin*

Sans entrer dans des explications détaillées, il est facile de comprendre que dans le texte des actes du Parlement Européen, on ne rencontre que « maure » parmi les éléments de cette liste[8]. Notre processus d'extraction donne plusieurs synonymes qui ne sont pas tous acceptables, mais dont les deux premiers sont néanmoins valables :

*arabe : (0.044) musulman, (0.021) saoudien*

Force est de constater que ces deux termes proposés sont tout à fait acceptables dans certains contextes, et cela montre une fois encore que la référence ne reflète pas forcément l'usage contemporain de certains mots. Par conséquent, la valeur du rappel est toute relative, et ne sert en fait que pour la comparaison avec la méthode monolingue.

# 6 CONCLUSION ET PERSPECTIVES

Comme nous l'avons vu, l'extraction automatique de synonymes à partir de corpus multilingues alignés permet d'aboutir à une précision plus satisfaisante qu'à partir d'un corpus monolingue. Certes la difficulté réside dans la recherche d'un corpus multilingue fiable, ainsi que dans le développement d'un système d'alignement mot à mot, mais la précision des résultats et les perspectives lexicographiques renforcent l'intérêt de cet instrument. En effet, nous avons vu que la qualité des synonymes produits, bien qu'améliorable par la prise en compte des unités multiples en amont de l'étape d'alignement mot à mot, est suffisante pour réfléchir à la pertinence du dictionnaire d'évaluation. En inversant ainsi la finalité de ces travaux, c'est-à-dire en passant du développement d'un instrument à l'examen de ses résultats, nous pouvons imaginer dans un premier temps apporter ainsi à la lexicographie un outil qui actualiserait objectivement les données des dictionnaires. En allant plus loin, et en extrapolant légèrement cette méthode à la traductologie, il nous paraît envisageable d'élargir le concept de dictionnaire

[7] On ne donne pas ici TOUS les synonymes proposés par le DES, mais seulement ceux dont la présence est attestée dans le corpus *Europarl*.

[8] À la rigueur, « bédouin » aurait sans doute pu apparaître, de même que « sarrasin » mais ce dernier probablement en tant que végétal.

de synonymes à celui de dictionnaire de paraphrases, ce qui constituerait un outil intéressant pour tous les apprenants de la langue.

# 6 REMERCIEMENTS

Cet ouvrage est basé sur une recherche qui s'est déroulée dans le cadre du projet 'Question answering using dependency relations' qui fait partie du programme de recherche 'Interactive Multimedia Information eXtraction', IMIX, qui a été financé par NWO, l'organisation néerlandaise de la recherche scientifique.

# 7 RÉFÉRENCES

Bourigault D. (2002). « Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », in : *Actes TALN 2002*, Nancy.

Bourigault D. & Galy E. (2005). « Analyse distributionnelle de corpus de langue générale et synonymie », in : Williams G. (ed), *Actes des 4èmes Journées de la Linguistique de Corpus* [http://web.univ-ubs.fr/corpus/jlc4.html#publi2005]

Curran J.R. & Moens M. (2002). « Improvements in automatic thesaurus extraction », in : *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, p. 59-67.

Gale W. & Church K. (1993). « A program for aligning sentences in bilingual corpora. », *Computational Linguistics*, vol. 19/1.

Jacquet G. & Venant F. (2005). « Construction automatique de classes de sélections distributionnelles », *12ème conférence annuelle sur le traitement automatique des langues* (TALN 05), Dourdan.

Koehn Ph. (2003). « Europarl: A multilingual corpus for evaluation of machine translation », Travail non publié, disponible sur http://www.statmt.org/europarl/.

Lin D., Zhao S., Qin L. & Zhou M. (2003). « Identifying synonyms among distributionally similar words », in : *IJCAI*, p. 1492-1493.

Van der Plas L. & Tiedemann J. (2006). « Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity », in : *Proceedings of Coling-ACL*.

Och F.J. (2003). « GIZA++: Training of statistical translation models », Disponible sur http://www.fjoch.com/GIZA++.html.