CRÉATION ET EXTENSION AUTOMATIQUES DE DICTIONNAIRES TERMINOLOGIQUES MULTI-LINGUES SPÉCIALISÉS À PARTIR DE CORPUS MONOLINGUES

Stéphane Ferrari & Violaine Prince

LIMSI-CNRS
BP 133
F-91403 ORSAY CEDEX
FRANCE
[ferrari, prince]@limsi.fr

Résumé

Dans cet article, nous proposons une méthode pour la création et l'extension de dictionnaires terminologiques multi-lingues spécialisés. Cete méthode présente la particularité de ne pas nécessiter de corpus alignés mais de simples corpus monolingues thématiques. Nous présentons le prototype de traitement de corpus français implémentant les outils nécessaires à sa mise en oeuvre. Nous exposons ensuite les différentes phases et algorithmes de passage du monolingue au multilingue.

Mots-clef

dictionnaires spécialisés multi-lingues, corpus monolingues.

1 Introduction

L'augmentation de dictionnaires électroniques à partir d'une étude automatique de corpus a déjà été utilisée dans de nombreux travaux, notamment dans (Daille et al. 94) en ce qui concerne la mise en correspondance de terminologie bilingue. Dans le même esprit, mais de manière complémentaire, nous avons pu, dans (Tanaka et Prince 95), proposer une amélioration d'un dictionnaire bilingue général Japonais-Anglais de près de 20% en nombre de "bons équivalents", et ce, à partir d'un corpus monolingue en langue source (Japonais). C'est là que nous avons constaté que des corpus spécialisés avaient tendance à "spécialiser" le dictionnaire en sortie, et que cette particularité pouvait être utilisée pour créer ou étendre des dictionnaires bilingues de spécialité, outils interessants *per se* pour la traduction technique de documents à l'aide d'ordinateurs. Nous nous sommes plus particulièrement focalisés sur les dictionnaires économiques multilingues, étant donné l'abondance des corpus, et notamment des corpus journalistiques, dans ce domaine, et dans différentes langues. Par ailleurs, la terminologie économique est relativement prisée lorsqu'il s'agit de couplage de langues indo-européennes (telles l'Anglais, le Français, l'Espagnol) avec des langues d'autres groupements, comme l'Arabe, le Japonais, le Coréen, etc...

Dans le cadre des projets de l'Association des Universités de Langue Partiellement ou Entièrement Française (AUPELF), nous menons actuellement un travail d'extension d'un dictionnaire électronique économique et social quadrilingue (Français, Anglais, Arabe, Japonais), par application de méthodes d'extraction automatiques ou semi-automatiques sur des corpus hétérogènes. Notre travail en est à sa première phase : celle où les corpus monolingues de la langue source (ici le Français) doivent être étiquetés et préparés pour l'extraction des termes économiques fréquents, qui recevront donc l'attention des experts pour la traduction.

Dans cette contribution, nous parlerons des résultats obtenus durant cette phase, et qui nous ont paru mériter de retenir l'attention. Nous commencerons dans la section 2, par expliquer comment nous procédons pour le traitement du corpus source en Français, et les résultats obtenus sur les études de termes fréquents. Nous décrirons ensuite, en section 3, la méthode utilisée dans (Tanaka et Prince 95) pour réaliser les correspondances pour les termes déjà existants dans la base dictionnairique de départ (nous partons d'un ensemble minimal implanté, pré-extrait à partir d'un dictionnaire général). Puis nous conclurons en section 4 sur l'état actuel des travaux, les perspectives qu'ils ouvrent, et les étapes ultérieures vers lesquelles nous nous engagerons.

2 Traitement automatique de corpus monolingues en français

Nous avons constitué un corpus thématique d'environ 450 000 mots, recueil d'articles de Bourse ou d'économie, extraits automatiquement[1] du journal "Le Monde sur CD-ROM". Ce corpus thématique doit permettre d'amorcer la spécialisation des dictionnaires ayant le français comme langue source. De manière à optimiser les procédures de passage du monololingue au multilingue, il est nécessaire de développer en premier lieu les outils de traitement de corpus adaptés aux langues utilisées.

Les prérequis au traitement multilingue sont essentiellement l'étiquetage et la lemmatisation. STK, un outil intégrant des ressources déjà existantes, est actuellement en cours de développement pour le traitement des corpus en français.

2.1 Étiquetage

Dans sa version actuelle, le prototype STK intègre l'étiqueteur initiallement développé par Éric Brill (Brill 92) pour l'anglais. Certaines contraintes accompagnent ce choix : format du texte issu de la segmentation et format du lexique.

2.1.1 Segmentation

L'étiqueteur de Brill requiert une entrée constituée d'une seule phrase par ligne, et préalablement segmentée. Une procédure particulière de segmentation de texte a de ce fait été développée . Une segmentation simple inclue la séparation des différentes formes fléchies et ponctuation. Ainsi, une entrée du type :

"*L'enfant de Pierre lit son livre.*"

devient

"*_l'_enfant_de_Pierre_lit_son_livre_._*"[2]

Comme le montre l'exemple précédent, certains problèmes sont inhérents à une segmentation simple : traitement des majuscules (ici, le déterminant "*L'*" doit basculer en minuscule, tandis que le nom propre "*Pierre*" est à conserver), ellipses (nous préférons conserver l'apostrophe dans "*l'*" de manière à ne pas ajouter d'ambiguïtés artificielles à l'étiquetage qui suivra, "*l*" seul pouvant être une simple lettre), ponctuations (un point peut être utilisé dans un sigle tel "*C.N.R.S.*" ou en fin de phrase), etc.

La nécessité d'une entrée au format *une phrase par ligne* constitue une contrainte d'un ordre supplémentaire. La reconnaissance de l'unité phrase peut en effet être améliorée par le résultat d'un étiquetage, voire d'une analyse syntaxique lors d'imbrications du type :

Pierre s'exclama :

"Cet enfant est bien sage !", tandis que son fils poursuivait sa lecture.

Dans le cas de l'étiqueteur de Brill, la situation est inversée, et la recherche de l'unité phrastique ne peut se fonder que sur le texte brut. De ce fait, la procédure de segmentation développée dans STK inclue un traitement spécial des insertions (parenthèses, tirets, dialogues, listes) de manière à produire la sortie adéquate.

2.1.2 Lexique

Le lexique utilisé, tant pour l'étiquetage que pour la lemmatisation, est adapté de BDLEX, initialement développé par G. Perrenou et M. De Calmès à l'IRIT, Toulouse. Il contient environ 250 000 formes fléchies correspondant à 23 000 formes canoniques. Le jeu d'étiquettes utilisé pour l'assignation des catégories syntaxiques dérive des informations contenues dans ce lexique, soit, principalement :

- genre et nombre pour les noms, adjectifs et participes passés,
- personne et nombre pour les verbes conjugués.

L'étiqueteur de Brill contraint le format lexical de la manière suivante :

à une entrée lexicale (forme fléchie) doit correspondre l'ensemble des étiquettes possibles, la première de la liste étant la plus fréquente.

Le lexique BDLEX ne contenant pas d'information sur les fréquences d'utilisation, nous avons ordonné les liste selon quelques heuristiques simples. Les erreurs conséquentes sont actuellement corrigées manuellement (p.e. *son* est plus souvent un *possessif* qu'un *nom*).

2.2 Lemmatisation

La lemmatisation consiste à associer une forme canonique aux différentes formes fléchies rencontrées (p.e. "*vouloir*" pour "*veux*", "*voulu*", "*voulait*", ...). Nous exploitons ici les résultats de l'assignation de catégories syntaxiques effectuée par l'étiqueteur de Brill, ainsi que le lexique BDLEX initial. Ainsi, à la forme fléchie "*été*" sera

---

[1] Le corpus du Monde permet de catégoriser les différents articles selon de multiple critères (titre, auteur, secteur de rédaction, etc...). Les articles extraits avaient ici pour thème commun soit *Sujet en France : bourse*, soit *Secteur Rédaction : économie*.

[2] Nous avons choisi d'utiliser des blancs soulignés pour rendre plus visibles les séparations. Les séparateurs utilisés sont en réalité le blanc (espace séparateur de formes fléchies) et le caractère de fin de ligne (séparateur de phrases).

associé soit l'information *verbe*, soit l'information *nom* par l'étiqueteur, permettant alors d'y associer les formes canoniques respectives "*être*" ou "*été*". Le principal problème auquel nous sommes actuellement confronté concerne la recherche d'une forme canonique pour les mots absents du lexique. Sans automatisation d'une telle procédure, nous avons effectués nos premiers test sur de simples couples (forme fléchie, étiquette). Cette approche expérimentale nous a permis de constater une distribution de certains noms spécialisés fortement dépendante du nombre, information présente dans notre jeu d'étiquettes. Nous avons de ce fait utilisé ce même format (forme, étiquette) pour l'ensemble des termes, et avons pu conforter notre observation. Ainsi, en économie, le mot "*barre*" est employé couramment au singulier, pour signifier la notion de "seuil de valeur", tandis qu'aucune utilisation de ce même mot au pluriel n'apparaît dans notre corpus.

Ces constatations nous ont amené à constituer en premier lieu des lexiques de couples (forme fléchie, étiquette), couvrant l'ensemble du corpus, la lemmatisation devant suivre désormais cette phase initiale de repérage des termes spécialisés.

2.3 Outils annexes, évaluation

De manière à permettre une gestion plus souple des corpus, un découpage en unités logiques balisées en SGML est en cours de développement (unités *articles* avec *entêtes* formatés, *titre*, *auteur*, etc., pour le corpus journalistique, *paragraphes* pour tous corpus). Cette option permet une navigation HTML classique et une visualisation plus efficace des différents résultats.

Des lexiques de couples (forme, étiquette) sont constitués automatiquement, ordonnés par fréquences d'emploi et catégories, de manière à déterminer le lexique terminologique. STK est soumis à une évaluation dans le cadre d'une ARC (action de recherche concertée) de l'AUPELF pour la partie extraction de lexique terminologique spécialisé. La méthodologie d'extraction combine les fréquences d'occurrence des termes à un marquage textuel (Ferrari 96), de manière à distinguer notamment les termes métaphoriques du vocabulaire spécifique au domaine du corpus étudié.

2.4 Premiers résultats et commentaires

Les premiers résultats obtenus sur le corpus extrait du journal Le Monde nous ont permis de constater l'utilisation d'un vocabulaire fortement polysémique, notamment en ce qui concerne les formes nominales les plus fréquentes (voir Tableau 1), justifiant la nécessité de leur affecter un sens en contexte. Ainsi, le terme "*vendredi*" correspond, dans notre corpus, tantôt à la date des bilans hebdomadaires journalistiques, tantôt à ce que les spécialistes du "forex"[3] nomment l'*effet week-end*. Nous avons ainsi entendu parler récemment de "*vendredi noir*" aux informations pour qualifier les retombées de l'anonce d'un taux de chômage inattendu aux États-Unis.

score total	catégorie grammaticale	forme fléchie
1575	NMS	marché
861	NMP	francs
659	NFS	semaine
605	NFS	hausse
589	NMS	groupe
589	NFS	page
577	NMS	mois
573	NFP	actions
571	NMP	milliards
552	NFP	valeurs
547	NMP	titres
503	NFS	baisse
502	NMP	marchés
481	NFS	année
459	NMS	octobre
440	NMS	cours
436	NMP	millions
434	NMS	vendredi
403	NFS	fin
396	NFS	société
392	NMS	terme
362	NFS	place
357	NFS	fois
351	NFP	entreprises
349	NMP	taux
344	NMS	lundi
340	NMP	jours
338	NFP	sociétés
335	NMS	change
332	NMS	début

Légende

score : nombre d'occurrence sur l'ensemble du corpus.
catégories : **N** pour *nom*, **M** et **F** pour *masculin* ou *féminin*, **S** ou **P** pour *singulier* ou *pluriel*

Tableau 1 : Les 30 noms les plus fréquents extraits du corpus d'articles d'économies du journal << Le Monde sur CD-ROM >> par STK *v 0-1.1*

D'une manière plus générale, des termes comme *marché, groupe, action, valeur, titre, cours, société, place,* etc., apparaissant dans le tableau précédent, voient leur polysémie naturelle

[3] forex : mis pour Foreign Exchange Market

fortement réduite dans le domaine économique. Le nombre d'acceptions peut ici être réduit pour approcher la monosémie dans la plupart des cas. Nous remarquons qu'un terme tel *vendredi* constitue en quelque sorte une exception, car sa polysémie dans le corpus peut être vue comme une conséquence de la co-existence de deux domaines sémantiques distincts : l'un concernant l'économie, l'autre concernant le journalisme. Une automatisation de la spécialisation du lexique n'est de ce fait pas remise en question par ces premiers résultats expérimentaux. Nous constatons simplement une dépendance de leur validité *a priori* liée au choix du corpus source.

L'existence d'exceptions telles celle exposée précédemment nous a conduit à considérer la constitution des correspondances entre lexiques bilingues spécialisé de manière semi-automatique, l'avis d'un expert étant requis, ne serait-ce que pour lever les ambiguïtés dues à la présence de domaines sémantiques résiduels ou conflictuels.

3 Description des outils de passage monolingue-multilingue

Les outils de passage entre le niveau monolingue et la création ou l'extension d'un dictionnaire multilingue sont de type semi-automatique, c'est-à-dire qu'ils font appel en dernier ressort à un jugement d'expert (traducteur, lexicographe). La partie automatique consiste à fournir justement à cet expert le meilleur sous-ensemble possible de termes pour la mise en équivalence, facilitant par là très largement le travail du lexicographe, qui recherche les différents sens d'un même terme pour en proposer des traductions. Cette partie automatique commence donc par l'application de l'outil STK pour la mise en forme des corpus et la recherche des termes les plus fréquents dans un domaine de spécialité donné (ici l'économie). Ces termes devront donc, en tout ou partie, prioritairement figurer dans un dictionnaire spécialisé multilingue "à jour". Ensuite, nous appliquons des méthodes fondées sur des mises en correspondance terminologiques inter-linguistiques que nous exposons ci-après.

3.1 Hypothèse de l'équivalence polysémie monolingue-multilingue : énoncé et restrictions

Les travaux de Church (Church et al. 91) aussi bien que ceux de Daille (op.cit.) admettent l'hypothèse suivante, que nous reprenons pour notre part, parce qu'elle semble raisonnable dans le cadre d'un domaine terminologique relativement restreint :

- si un terme *t* dans une langue source L_S possède $(y_1, y_2, ...y_n)$ comme synonymes alors les traductions dans la langue cible L_C des y_i sont les équivalents dans L_C de *t*.

Cette hypothèse est aménagée pour répondre aux besoins de l'extraction de sens associés. En particulier, nous posons les restrictions suivantes :

- les synonymes dans la langue source ne sont pas forcément les traductions stables dans la langue cible : c'est le problème de la conservation de l'assignation du sens depuis L_S vers L_C et inversement depuis $L_C$vers L_S qui a été largement débattu dans (Tanaka et Umemura 94). Par conséquent, si les synonymes peuvent être des équivalents, cela n'est vrai que dans le contexte d'expression de *t*, c'est-à-dire étant donné le domaine auquel appartient le texte où l'on doit traduire *t*, le thème du paragraphe et celui de la phrase, mais surtout, étant donné les usages d'une langue à l'autre. Ainsi par exemple, les équivalents (*citadin, poli)* de *urbain* en Français, seront traduits de manière respective par *city dweller* et *urbane* en Anglais selon les contextes locaux, mais il existe aussi de nombreux cas où *urbain* est directement traduit par *urban*, et cela d'une façon non régulière.
- la méthode la plus couramment adoptée pour rechercher des équivalents en contexte est la méthode dite de l'*information mutuelle* utilisée par Church (op.cit.). Elle s'établit selon la formule suivante :
  soient a et b deux termes dont on veut juger de la co-relation. L'information mutuelle établie entre a et b dans un corpus donné C est

$$I(a,b) = \log_2 \frac{P(a,b)}{P(a)P(b)}$$

où P(a) est la probabilité de trouver a dans, le corpus et P(b) celle de trouver b, et P(a,b) étant la probabilité d'avoir a si b.
Le problème est que l'information mutuelle dénote une propriété de forte corrélation, mais elle n'indique pas de quelle nature est cette relation, en particulier, elle ne permet pas de discriminer les synonymes de *a*, souvent co-utilisés dans une relation anaphorique pour éviter la répétition, d'autres termes reliés à *a* qui peuvent être de l'ordre des *differentiae* de *a*, des antonymes, des descripteurs thématiques etc...C'est le cas par exemple des co-occurrents de *docteur* qui peuvent être (*médecin, praticien, hopital, santé, soins, ordonnance...*) où le premier terme est synonyme, mais où les suivants sont respectivement un hyponyme, et des arguments d'une description prédicative ou schématique d'un des sens du terme *docteur*

(qui est alors traduisible en Anglais par *doctor/docteur* ou *physician/médecin*).

Ces défauts sont certes notables, néanmoins, l'information mutuelle est une méthode simple qu'il convient de conserver, mais en en restreignant les conclusions. Sinon, il faut verser dans une théorie de la synonymie, et donc, soit se référer à une hiérarchie dite "ontologique" des termes où seuls certains hyperonymes et hyponymes pourraient avoir le statut de synonymes (la limite restant à définir), soit utiliser un dictionnaire des synonymes dans la langue source et n'admettre que l'intersection entre les co-occurrents et les termes du dictionnaire. Ce dernier cas peut être aussi très restrictif et faire perdre l'acquis dynamique et innovant de l'extraction terminologique en discours.

Le moyen terme à trouver est celui dans lequel deux phases sont nécessaires pour réaliser une bonne extension de dictionnaire à partir de l'hypothèse de l'équivalence polysémie monolingue-multilingue.

- une première phase dans laquelle, à l'aide de l'information mutuelle, on constitue des "matrices de proximité" entre termes, ces termes étant préalablement lemmatisés par STK, de façon à conserver des catégories grammaticales signifiantes (noms, verbes, adjectifs) et surtout homogènes avec le terme à traduire. Les corpus considérés sont forcément homogènes eux aussi, c'est-à-dire relevant d'une spécialité, pour circonscrire, ne serait-ce que grossièrement, la dérive de sens de certains mots polysèmes.
- une deuxième phase dans laquelle on extrait les termes ayant les plus forts coefficients dans la matrice et une expertise humaine leur assigne ou non le statut d'équivalent acceptable.

Il est clair que dans ce cas, l'extension d'un dictionnaire existant est beaucoup plus facile à réaliser que sa création : l'automatisation ne fait que rendre moins fastidieux le travail lexicographique, mais elle ne propose aucunement de se susbstituer au rôle décisionnel du spécialiste. D'ailleurs, la structure même de l'algorithme des matrices de proximité montre que seules des décisions de très bas niveau sont confiées au système. Nous donnons, dans le prochain paragraphe, les grandes lignes de cet algorithme, adapté de (Tanaka et Prince 95).

3.2 Algorithme des matrices de proximité

3.2.1 Etat initial du dictionnaire

L'algorithme suppose un premier dictionnaire bilingue existant, indicé par 0, dans lequel, originellement, pour un terme *t* en entrée dans la langue source L_S nous avons $(e_1, e_2, \ldots e_p)$ équivalents préalablement fournis. Si le dictionnaire n'existe pas, la liste des équivalents est vide.

Chacun des e_i n'est pas une traduction "sûre" de *t,* ni une traduction équivalente aux autres qui sont proposées dans le dictionnaire. Pour cela nous avons choisi d'affecter un poids $w(t,e_i)$ à chaque branche de la relation, qui va bien entendu varier au cours de l'analyse de corpus. Au départ, nous avons choisi de considérer le rang de l'équivalent dans le dictionnaire comme étant un indice de son importance. Nous avons fait l'hypothèse que le lexicographe a naturellement fourni en premier la traduction la plus "sûre" ou la plus "fréquente" du terme. Mais il est tout à fait possible, voire même probable, qu'en fin de traitement, le système modifie les pondérations existantes, inverse les rangs, et introduise de nouveaux termes.

Par conséquent, on définit le poids initial d'un équivalent e_i de rang i , dans un ensemble de p équivalents du terme *t* comme étant :

$$w_0(t,e_i) = \frac{p+1-i}{\frac{p(p+1)}{2}} \quad \textbf{(1)}$$

qui est une valeur harmonisée de façon à avoir la somme des poids égale à 1 à l'initialisation. Si le dictionnaire n'existe pas, la liste des équivalents étant nulle, les poids sont eux aussi égaux à 0.

3.2.2 Recherche des co-occurrents

Les co-occurrents de *t* sont repérés dans un corpus C par une valeur de I, information mutuelle, supérieure à 0,1. Dans l'ensemble des co-occurrents, on extrait les termes de même catégorie grammaticale que *t* (partie réalisée par STK). Si un terme possède plusieurs catégories, on prendra celle(s) qui appartien(nen)t à l'intersection avec l'ensemble des catégories équivalentes à celle de *t* (p.e. *{NFP, NFS, NMP, NMS}* pour un nom) On ne cherche pas spécifiquement à désambiguïser finement ici.

Soit S l'ensemble des co-occurrents homogénéïsés. S= $(s_1, s_2, \ldots s_q)$.

3.2.3 Modification des pondérations des équivalents

A partir de ce point, on recherche si les s_j sont dans le dictionnaire bilingue. Si oui, on recherche leurs équivalents. Si le dictionnaire n'existe pas, il faut passer ici par le jugement d'un expert traducteur (ou par un dictionnaire non électronique) qui fournit une liste d'équivalents

des co-occurrents (chaque liste pouvant être éventuellement réduite à un élement, ou vide, mais il existe au moins une liste non vide d'équivalents) pour initialiser le système.

Si un équivalent d'un des s_j est aussi un des équivalents de t (i.e., c'est un terme e_k) alors son poids w(t, e_k) est augmenté proportionnellement de son poids vis-à-vis de s_j. Cela signifie que la probabilité pour que e_k traduise t dans le contexte du corpus C est plus grande que ne le laisse supposer son rang vis-à-vis de t. Deux cas sont possibles pour les équivalents originaux :

- l'équivalent n'est relié à aucun des co-occurrents. Il n'est pas pertinent vis-à-vis du contexte et du corpus
- l'équivalent est relié à un ou plusieurs co-occurrent , qui possèdent ou non d'autres équivalents. Il est un équivalent contextuel possible de t, dans le cadre du corpus C. Il faut alors définir le poids contextuel de chaque équivalent de t. Pour cela, nous adoptons la formule générique suivante de pondération dans le corpus C:

$$w(t,e_k) = \sum_{j=1}^{q} I(t,s_j) \times w(s_j,e_k) \quad \textbf{(2)}$$

Si e_k n'est relié à aucun des s_j, alors

$\forall j \in (1,q) \;\; w(s_j,e_k) = 0$ donc $w(t, e_k) = 0$.

3.2.4 Introduction de nouveaux équivalents

La formule sert aussi pour l'introduction de nouveaux équivalents, c'est-à-dire des termes n'existant pas dans la liste $(e_1,e_2,\ldots e_p)$ de départ.

En effet, on calcule ce poids contextuel pour tous les équivalents des s_j.

3.2.5 Matrices de proximité

On considère que le dictionnaire source est composé de la liste en sortie des études de fréquence (voir la section 2). Chacun des termes du dictionnaire source est associé par une fonction T de traduction à au moins un terme e qui le traduit, et qui appartient à la même (ou à une des mêmes) catégorie(s) grammaticale(s).

On initialise le dictionnaire sous forme d'une matrice dite d'information statique, MCM_S, dont l'indice en ligne i est un terme t_i en langue source L_S et l'indice colonne j est un terme e_j en langue cible L_C et dont les éléments sont :

$$MCM_s(i,j) = w_0(t_i,e_j), \quad \text{si } e_j = T(t_i)$$

$$MCM_s(i,j) = 0, \quad \text{sinon.}$$

où T(x) = y signifie que y est un équivalent de x, et w_0 est calculé comme dans la formule (1).

MCM_S est une **matrice de proximité initiale** des termes entre la langue source et la langue cible.

On appelle $MCMd_i$ la matrice obtenue en considérant les mêmes lignes et les mêmes colonnes que la matrice MCM_S mais telle que l'élément dans la ligne i et la colonne j correspond au poids contextuel (dynamique) de l'équivalent e_j dans le corpus de rang i.

$$MCMd_i(i,j) = w(t_i,e_j)$$

où w est le poids contextuel de e_j calculé d'après la formule (2).

C'est une matrice de proximité dans un corpus d'une certaine taille, et elle est appelée **matrice de proximité dynamique** de ce corpus. Elle dénote l'action pure de ce corpus en terme de réorganisation des équivalents.

La méthode incrémentale définie dans (Tanaka et Prince 95) est telle que nous considérons un ensemble de corpus tels que $Corpus_1 \subset Corpus_2 \subset \ldots \subset Corpus_m$

Donc le corpus de rang i , i > 1, donne une matrice de proximité résultante **calculée de la façon suivante :**

$$MCM_i = (MCMd_i - MCMd_{(i-1)}) \otimes MCM_S \oplus MCM_{i-1} \quad \textbf{(3)}$$

où la partie : $MCMd_i - MCMd_{(i-1)}$ dénote la variation de poids contextuel entre le corpus de rang i et son sous-ensemble de poids inférieur. Cette variation affecte la matrice de proximité initiale MCM_S.

La matrice résultante de rang i est dépendante de celle de rang $i-1$. On considère que la matrice de rang 0 est la matrice nulle.

Par conséquent,

$$MCM_1 = MCMd_1 \otimes MCM_s.$$

3.3 Propositions préalables pour un dictionnaire quadri-lingue

L'algorithme fourni ci-dessus a été testé dans (Tanaka et Prince 95). Il avait aussi pour but initial de rechercher la relation entre la taille d'un corpus et son impact sur le dictionnaire. En expérimentant sur six corpus incrémentaux, nous nous sommes aperçus que les résultats se stabilisaient aux alentours d'une taille de 25 Mo pour les corpus d'origine (corpus journalistiques de 100 Mo en langue japonaise, extraits du qutodien ASAHI sur des thèmes économiques et sociaux). Cela ne signifie pas que 25 Mo soit forcément un seuil *in abstracto*, mais que, manifestement, la stabilisation des résultats à partir de corpus

incrémentaux est probable. Par ailleurs, nous avions fait plusieurs tentatives avant de proposer la formule de récurrence (formule (3)) et certaines de ces tentatives ont été mentionnées dans (Tanaka et Prince 95).

Les résultats appliqués sur un dictionnaire Japonais-Anglais nous ont montré que le corpus avait assez fortement modifié l'importance relative et introduit de nouveaux équivalents "locaux."

En ce qui concerne le dictionnaire quadrilingue qui nous intéresse, nous mettons au point la procédure suivante :

(1) sélection et lemmatisation des termes les plus fréquents en Français (domaine économique) à partir d'un corpus journalistique courant (section 2)
(2) création semi-manuelle d'une matrice de proximité initiale à partir de dictionnaires électroniques Français-Anglais, Anglais-Japonais, et de dictionnaires économiques manuels Français-Arabe (section 3)
(3) extraction dans des corpus journalistiques français de co-occurrents des termes
(4) sélection manuelle des co-occurrents synonymes en Français
(5) passage itératif de l'algorithme des matrices de proximité jusqu'à la stabilisation
(6) Étude des sorties pour la validation par les experts.

4. Conclusion

Nous avons exposé les différentes étapes nécessaires à la création et l'extension automatiques de dictionnaires terminologiques multi-lingues spécialisés à partir de corpus monolingues. Leur mise en place est maintenant amorcée par la création des outils nécessaires au traitement préalable des corpus sources en langue française.

Un prototype de traitement de corpus, STK, permet désormais le balisage, la segmentation et l'étiquetage de gros corpus, ainsi que l'extraction de couples ordonnés (forme fléchie, catégorie grammaticale), de manière à constituer le lexique spécialisé source. Différents problèmes inhérents à cette première phase ont été soulevés, de manière à pouvoir rendre compte de leur influence sur les phases suivantes. Les premiers résultats expérimentaux nous ont en outre amené à considérer des phénomènes inattendus tels l'influence du nombre sur le sens d'un mot en contexte.

Les algorithmes de recherche de co-occurrents et d'établissement des matrices de proximité n'ont pas encore été appliqués sur ces premiers résultats. Il ont cependant déjà été testés et validés dans le cadre d'une extension de dictionnaire Japonais-Anglais. Une stabilisation des matrices de proximités a pu être constatée lors de ces tests. Il nous semble intéressant d'étudier l'existence de tels seuils de stabilisation pour le dictionnaire quadrilingue, et leur éventuelle dépendance par rapport aux paramètres intervenant dans la première phase (qualité des différents niveaux d'analyse du corpus source).

Références

(Brill 92) E. Brill, A simple rule-based part-of-speech tagger. In : *Proceedings of the Third Conference of Applied Natural Language Processing*. ACL. Trento, 1992

(Church et al. 91) K. Church et alii., Using statistics in Lexical Analysis. In Zernik U. (ed.), *Lexical Acquisition Exploiting On-line Resources to Build a Lexicon*, Laurence Erlbaum Associates. Pp. 65-90. 1991

(Daille et al. 94) B. Daille et alii, Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the International Conference for Computational Linguistics*. Pp. 515-521. 1994.

(Ferrari 96) S. Ferrari, Métaphores : extraction automatique de lexiques métaphoriques à partir de corpus. In : *Actes du Colloque Étudiant de Linguistique Informatique de Montréal CLIM-96*, Montréal, Québec, Canada. 08-11 juin 1996.

(Tanaka et Prince 95) K. Tanaka et V. Prince, Dictionnaires bilingues incrémentaux utilisant un corpus monolingue. In *Actes des Journées Lexicomatiques et Dictionnairiques*, Lyon, septembre 1995. AUPELF-UREF. 1995

(Tanaka et Umemura 94) K. Tanaka et K. Umemura. Construction of a bilingual dictionary intermediated by a third language. In : *Proceedings of the International Conference for Computational Linguistics*. Pp. 392-397. 1994