Mesures de comparabilité pour la construction assistée de corpus comparables bilingues thématiques

Guiyao Ke

Mesures de comparabilité pour la construction assistée de corpus comparables bilingues thématiques

Thèse soutenue le 26 février 2014,
devant la commission d'examen composée de :

Mme. Pascale SEBILLOT
Professeur, INSA de Rennes, IRISA / Président

M. Eric GAUSSIER
Professeur, Université Joseph Fourier, LIG / Rapporteur

M. Emmanuel MORIN
Professeur, Université de Nantes, LINA / Rapporteur

M. Geoffrey WILLIAMS
Professeur, Université de Bretagne Sud, LICORN / Examinateur

M. Pierre-françois MARTEAU
Professeur, Université de Bretagne Sud, IRISA / Directeur

Remerciements

Je présente mes sincères remerciements à mon directeur de thèse Pr. Pierre-François MARTEAU pour avoir dirigé et encadré cette thèse ainsi que pour ses efforts, son aide, ses conseils et ses encouragements.

Je remercie tout spécialement les rapporteurs Pr. Emmanuel MORIN et Pr. Eric GAUSSIER, ainsi que les examinateurs Pr. Pascale Sébillot, présidente du jury de soutenance, et Pr. Geoffrey WILIAMS, pour m'avoir fait l'honneur d'assister à mon jury de thèse, pour leur nombreuses questions et suggestions constructives et pour le temps qu'ils ont consacré à l'évaluation de mon travail.

Je remercie aussi tous les différents collaborateurs du projet ANR METRICC, surtout ceux qui ont contribué à ce travail, tout particulièrement Dr. Gildas MENIER et Dr. Bo LI.

Je remercie également tous les membres du laboratoire IRISA (ex. VALORIA), particulièrement les doctorants, les docteurs et permanents de l'équipe de football, en particulier Abdulkader BENCHI, Djamel BENFERHAT, Thibaut LE NAOUR, Jean-François KAMP, etc.

Je remercie enfin Pr. Quansheng LIU du laboratoire LMBA-UBS et Pr. Thierry BAUTIER de l'Ecole Supérieure du Professorat et de l'Education de Bretagne pour leur soutien.

Je dédie ce travail à mes parents, à tous les membres de ma famille et toutes les personnes qui ont contribué, de près ou de loin, à travers leur soutien moral ou matériel, à ce que celui-ci aboutisse.

Table des matières

Table des figures

Liste des tableaux

Première partie

INTRODUCTION

1 Introduction

Sommaire

1.1 Contexte du projet METRICC

Cette thèse est issue du projet ANR METRICC (MEmoire de Traduction, Recherche d'Information et Corpus Comparables). Le projet METRICC aborde la problématique des corpus comparables d'une façon complète et originale. Plusieurs défis fondamentaux pour le domaine sont abordés. Ceux-ci s'expriment sous la forme des questions suivantes :

1. Comment construire des corpus comparables de la manière la plus efficace possible ? Comment évaluer la comparabilité et donc l'adéquation aux besoins du corpus ? Quels sont les indices permettant de valider un corpus avant d'effectuer les extractions de ressources ? Comment utiliser une telle mesure de comparabilité au moment même de la constitution du corpus (crawling) pour éviter les trop grandes dérives ?
2. Lorsque l'on dispose d'un corpus comparable adéquat, comment extraire les ressources bilingues nécessaires de la manière la plus efficace possible ?
3. Comment aider le traducteur travaillant sur un document donné ayant une thématique précise pour laquelle il est difficile de trouver un lexique bilingue pour la paire de langues voulue ?
4. Comment exploiter les corpus comparables pour enrichir les possibilités d'un système de recherche d'informations inter-langues ?

Pour répondre en partie à ces questions, plusieurs partenaires (laboratoires et entreprises) se sont associés dans le cadre de METRICC : le Laboratoire d'Informatique de Grenoble (LIG), le Laboratoire Informatique de Nantes-Atlantique (LINA), l'Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), les entreprises Lingua et Machina, Sinequa et Syllabs. L'objectif de cette thèse concerne plutôt le premier défi, celui de la construction des corpus comparables thématiques à partir de l'exploitation du WEB.

1.2 Motivations

Cette thèse a pour ambition de proposer des outils dédiés à la construction assistée de corpus comparables de "bonne qualité". Nous devons préciser en premier lieu ce que l'on entend par corpus comparables, leurs intérêts et retombées attendues, et les enjeux d'une assistance outillée à la construction de telles ressources qui en découlent.

1.2.1 Corpus comparables et corpus parallèles

La définition des corpus parallèles est précise et non ambiguë. Nous reprenons la définition proposée par [14] : "a parallel corpus contains texts and their translations into one or more languages" (il s'agit donc d'un ensemble de textes accompagné de leurs traductions dans une ou plusieurs langues). Les corpus parallèles sont importants dans le domaine de la traduction automatique ou assistée, de l'extraction des terminologies ou des dictionnaires bilingues, de la recherche d'informations multilingues, etc. Cependant, ils sont coûteux à développer et souvent difficiles à transposer d'un domaine de spécialité à l'autre.

Pour répondre (en partie) à ces inconvénients, la notion de corpus comparables a été proposée initialement dans les années 90 par [7], puis précisée ou adaptée au cours des années comme dans [39], [96]. Ces définitions se résument ainsi : un corpus comparable devrait couvrir un même thème ou un thème similaire ou partager certaines caractéristiques importantes (telles que le style, la période, etc.). Malheureusement, ces définitions sont trop vagues pour être exploitées en pratique. La définition la plus opérationnelle que nous avons identifiée est celle proposée par [30] : *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 $\mathcal{L}_1$*. A partir de cette définition, [77] a conçu une mesure de comparabilité que l'on pourrait qualifier de comparabilité "traductionnelle" quantitative, qui est relativement bien adaptée à une tâche d'aide à la traduction.

Les corpus comparables, au même titre que les corpus parallèles, peuvent servir dans plusieurs domaines d'application : extraction des terminologies ou des lexiques bilingues, fouille de données textuelles bilingues, traduction automatique ou assistée, assistant pédagogique pour l'apprentissage des langues, etc. Par rapport aux corpus parallèles, les corpus comparables ont trois avantages : premièrement, ils constituent des ressources riches et larges : en volume et en

période couverte ; deuxièmement, il n'y a plus de contrainte sur la traduction limitée au texte original : les corpus comparables fournissent des ressources linguistiques originales et thématiques. Enfin, ils sont moins coûteux à développer que les corpus parallèles. La contre-partie est que l'exploitation des corpus comparables est plus difficile comparativement aux corpus parallèles car les données qu'ils regroupent sont beaucoup plus hétérogènes et "bruitées". Ce sont ces raisons qui rendent la construction des corpus comparables de "bonne qualité" très attractive.

1.2.2 Motivation pour la constitution de corpus comparables

De nombreux travaux ont été réalisés pour construire des corpus comparables comme dans [137], [95], [135], [102], [149]. Nous allons les détailler dans notre chapitre état de l'art. Selon ces recherches, nous pouvons considérer que la constitution des corpus comparables passe par deux étapes principales : d'abord, la création d'un corpus initial de documents (le corpus de la langue source et de la langue cible) ; ensuite, l'utilisation de certaines techniques pour établir l'alignement des documents similaires entre la langue source et la langue cible afin de produire les corpus comparables définitifs. Cependant, une difficulté subsiste : comment faire un compromis entre deux facteurs importants : la qualité et la taille des corpus ? La croissance rapide des sources d'informations sur Internet fournit une belle occasion pour la construction des corpus comparables, en particulier via l'exploitation des publications quotidiennes issues des agences de presse en différentes langues, ou des ressources multilingues de qualité telles que Wikipédia.

En résumé, nos principales motivations reposent sur trois constats et questions associées :

— Les corpus comparables thématiques offrent de grands avantages par rapport aux corpus parallèles dès lors que leur qualité et leur taille sont suffisantes. Jusqu'où peut-on aller en matière d'assistance automatisée pour leur construction ?
— Il n'existe pas de cadre partagé et définitif d'évaluation de la notion de comparabilité partagée. La proposition de Li et Gaussier, indépendante du cadre applicatif constitue-t-elle une opportunité pour hiérarchiser ces mesures de comparabilité ?
— Il faut également pouvoir contrôler lors de leur construction, à la fois la volumétrie et la qualité des corpus comparables. Peut-on optimiser à la fois la qualité d'alignement des documents (ou clusters) et également la volumétrie dans la construction des corpus comparables thématiques ?

1.2.3 Corpus comparables thématiques versus corpus comparables généraux

Tout d'abord, nous avons besoin d'introduire les définitions suivantes :
— Un thème est un sous ensemble de documents caractéristique d'un vocabulaire partagé. Il se rapport à une idée, un sujet développé dans un discours, un écrit, un ouvrage

(défintion du Centre National de Ressources Textuelles et Lexicales (CNRS)). En fonction des traitements différents selon l'auditoire, il y a des genres différents. Alors, un genre est un format de production qui possède des caractéristiques de formatage et de choix lexicogrammatique typés [133], par exemple un article de recherche, un article de vulgarisation, un article de presse. Un domaine regroupe donc l'ensemble de termes spécialisés.

— Un événement est une série d'actions qui se passe à un moment donné, par exemple : une invasion en temps de guerre, les vendanges, etc. Il peut être assimilé à un thème ou un sous-thème.

— Un corpus comparable thématique : c'est un ensemble de documents multilingues qui traitent d'un même thème. En particulier, les termes (discriminants) caractérisant le domaine sont en général fréquents dans le corpus et peu ambigus.

— La notion de comparabilité thématique pour rendre cette notion opérationnelle s'exprime ainsi : deux corpus en langues $\mathcal{L}_1$ et $\mathcal{L}_2$ sont dits thématiquement comparables si :

  - d'une part 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 $\mathcal{L}_1$

  - d'autre part les termes des sous-parties des vocabulaires concernés doivent être tels que le ratio entre leur fréquence d'occurrence et leur nombre de traduction soit le plus grand possible (les termes fréquents et faiblement ambigus)

Dans ce contexte, la qualité (d'alignement des documents comparables) est-elle plus importante que la taille de corpus comparables ?

Il existe certains travaux qui démontrent que si la taille des corpus comparables est suffisante, la qualité a moins d'importance. Par exemple, [93] montre, dans le cadre de l'extraction de lexiques bilingues à partir de corpus comparables spécialisés, que si la qualité était prépondérante à la taille du corpus pour l'alignement de termes complexes cela n'était pas le cas pour l'alignement des termes simples. Cependant, le débat reste ouvert : en premier lieu, comment justifier que la qualité des corpus scientifiques est meilleure que les corpus mixtes (scientifiques + grand public) ? En second lieu, est-ce qu'une centaine de mots et quelques centaines de documents sont suffisants pour établir cette conclusion car la différence de résultats en matière d'extraction des terminologies entre les deux types de corpus reste petite ? Par ailleurs, dans [106], l'auteur montre que : "un corpus comparable correctement constitué est *au moins* aussi efficace qu'un corpus comparable moins bien constitué mais plus volumineux" ; et que : "les fréquences de cooccurrences des termes sont instables, même dans le cas de corpus fortement comparables, mais que ce phénomène est *aggravé* dans le cas de corpus moins comparables". Nous observons donc ici qu'une "bonne" qualité améliore potentiellement la performance de l'extraction terminologique.

Par contre, il existe plusieurs travaux comme [134], [87], [78], qui ont tendance à montrer que la qualité d'alignement des corpus comparables est plus importante que leur volume. No-

tamment, dans [107], les auteurs montrent que la qualité des corpus comparables (deux corpus comparables construits par leurs soins : un corpus construit à partir d'un alignement basé sur la similarité des concepts présents dans les documents et la date de publication, un autre corpus construit à partir d'un alignement basé sur les similarités de thème et des concepts avec des dates de publication différentes pour traiter des événements de longue durée) amélioré significativement les performances de l'extraction des traductions des mots et de la recherche d'information multilingue à partir des requêtes traduites.

Tout cela justifie notre motivation pour la construction de corpus comparables thématiques ayant une forte cohérence thématique tout en maintenant la qualité d'alignement et également en prenant en compte l'effet de la volumétrie de corpus.

1.3 Principales contributions de cette thèse

La plus grande contribution de cette thèse est le développement d'une nouvelle approche pour la constitution des corpus comparables thématiques de "bonne qualité" pouvant être facilement adaptable aux exigences en fournissant des corpus de niveaux variables de comparabilité. Nous explicitons cette approche en présentant 3 parties contributives :

1. La première porte sur le développement des mesures de comparabilité et leur évaluation
2. La deuxième porte sur les problématiques de clustering et de classification multilingue, et l'alignement des clusters comparables
3. La troisième développe une approche pour la constitution assistée de corpus comparables thématiques à partir de ressources hétérogènes.

Dans la première partie contributive, l'objectif est de fournir des mesures de comparabilité quantitatives pour mesurer la comparabilité entre deux documents ou même entre deux corpus de langues différentes.

Dans la deuxième partie contributive, le but est de fournir une approche efficace pour aligner deux espaces linguistiques différents, par exemple, un espace anglais et un espace français, par une approche de clustering ou de catégorisation qui fusionne des similarités *natives* dans chacun des espaces linguistiques avec des similarités *induites* par la mesure de comparabilité utilisée (nous appelons cette approche de clustering "SCF-clustering" et cette approche de classification "SCF-classification"). Nous montrons expérimentalement que cette fusion exploitant les mesures de comparabilité que nous avons développées dans la première partie améliorent la qualité du clustering ou de la catégorisation ainsi que l'alignement des clusters.

Enfin, dans la troisième partie contributive, nous généralisons ces résultats en proposant une approche semi-supervisée qui intègre les deux contributions précédentes, pour construire finalement des corpus comparables thématiques sur mesure et de qualité contrôlable.

1.4 Plan de thèse

Outre le chapitre d'introduction, cette thèse est constituée de cinq chapitres, comprenant un chapitre d'état de l'art, trois chapitres de contributions et un chapitre de conclusions et perspectives plus une annexe. L'organisation est la suivante :

Dans le ***Chapitre 2***, nous présentons un état de l'art décomposé en quatre sections : les corpus parallèles et les corpus comparables, les différentes approches pour la constitution des corpus comparables, les mesures de comparabilité, et les différentes types de clustering et de classification. Nous commentons les avantages et les inconvénients des corpus parallèles et l'intérêt des corpus comparables, nous analysons les différentes approches pour la constitution des corpus comparables, nous étudions les différentes mesures de comparabilité et nous présentons les différents types de clustering et de classification et la raison pour laquelle nous avons choisi telle méthode plutôt qu'une autre.

Le ***Chapitre 3*** est consacré aux mesures de comparabilité développées à partir de la mesure de comparabilité de référence [77]. Nous présentons d'abord la mesure de comparabilité de référence et proposons ensuite deux variantes. Les corpus de test, les dictionnaires bilingues utilisés et le prototype d'évaluation sont décrits. Les différentes étapes pour évaluer ces mesures de comparabilité sont également détaillées. Et enfin, nous commentons les avantages et les inconvénients des variantes par rapport à la mesure de comparabilité de référence.

Dans le ***Chapitre 4***, nous développons une nouvelle approche de clustering, de classification et d'alignement des clusters comparables. Cette nouvelle méthode combine les similarités *natives* avec la mesure de comparabilité pour concevoir une nouvelle mesure de similarité à caractère multilingue. Nous illustrons ensuite les différentes expérimentations effectuées sur deux types de corpus collectés sur le WEB : les Flux RSS (un corpus de test) de presses généralistes et Wikipédia (trois corpus de test). Pour chaque expérience, nous détaillons les résultats obtenus sur une classification de type k plus proches voisins (k-PPV) et deux types de clustering : K-médoides et HAC (Clustering hiérarchique ascendant), avec les pondérations tf et tf-idf. Enfin, nous analysons les résultats obtenus dans ces expérimentations.

Le ***Chapitre 5*** est dédié à l'intégration des deux contributions précédentes : les mesures de comparabilité et le modèle de mélange des similarités *natives* et les similarités *induites* par la comparabilité, pour développer une assistance à la constitution des corpus comparables de qualité. Nous détaillons en premier lieu les différentes étapes de cette approche. Nous présentons ensuite les corpus et le dictionnaire bilingue utilisés pour effectuer une fouille de textes comparables. Nous illustrons nos expérimentations en testant certains paramètres importants comme le nombre de clusters, le seuil de comparabilité ou de similarité. Enfin, nous commentons nos résultats en explicitant les clusters alignés obtenus.

Dans le ***Chapitre 6***, nous concluons cette thèse en listant les résultats principaux obtenus par rapport à la problématique posée et nous discutons les différentes voies possibles pour améliorer l'approche proposée et ses possibilités d'extension. ï»¿

Deuxième partie

ETAT DE L'ART

2

Etat de l'art

2.1 Introduction

Nous recensons dans ce chapitre les connaissances récentes sur les corpus notamment les corpus multilingues (bilingues), et tout particulièrement les corpus parallèles et les corpus comparables en section 2.2.

Dans la mesure où nous envisageons construire des corpus comparables à partir de clustering de données bilingues "brutes", i.e. collectées à partir de sources hétérogènes non dédiées, nous présentons dans la section 2.3, les différents types de classification et de clustering exploitables pour effectuer la classification et le clustering de documents bilingues, ainsi que les différentes techniques utilisées pour ajuster le nombre de clusters (K).

Enfin, la section 2.4 conclut ce chapitre.

2.2 Corpus : corpus parallèles et corpus comparables

La définition du terme "corpus" évolue beaucoup au fil du temps. Nous recensons dans les années 60 les définitions suivantes. Dans le Trésor de la Langue Française Informatisé (TLFI) [1], la définition proposée est : "Recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc." Dans le Larousse, la définition est un peu plus précise : "Recueil de documents relatifs à une discipline, réunis en vue de leur conservation. Ensemble fini d'énoncés écrits ou enregistrés, constitué en vue de leur analyse linguistique". Cependant, ces deux notions restent difficilement exploitables dans une optique traitement automatique des langues naturelles (TALN) ou fouille de textes car très générales et imprécises.
Pour cette raison, [125] a proposé deux définitions plus directement exploitables : l'une est adaptée à la notion de corpus au sens large et l'autre à la notion de corpus pour les traitements informatiques. Pour un corpus général : il s'agit d'un ensemble de morceaux de langue qui sont sélectionnés et classés selon des critères linguistiques explicites, afin d'être utilisés comme un *échantillon* de la langue. Pour un corpus en informatique : c'est un corpus qui est codé de manière standardisée et homogène pour la tâche ouverte de recherche d'information.

Pour rendre plus représentative la notion d'*échantillon* de la langue, [51] a proposé une autre définition : "un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue".
[34] a organisé ces différentes définitions de corpus selon trois facteurs :

1. La nature : le corpus est composé de données langagières.
2. La structure : les données du corpus sont sélectionnées, mises en forme et enrichies. Leur sélection se fait selon des critères de choix, de façon à ce que le corpus soit représentatif des objectifs visés. Les critères appliqués sont d'ordres linguistiques ou extralinguistiques. Le corpus ainsi constitué est ensuite mis en forme (normalisation) et enrichi (documentation, méta données, etc.).
3. La finalité : le corpus est représentatif d'un langage, d'un sous-langage ou de certains phénomènes linguistiques étudiés.

Enfin, dans [4], une définition encore plus générale a été proposée : "un corpus est un ensemble de ressources linguistiques originales d'une certaine taille et d'une certaine structure, collecté et traité pour une ou plusieurs applications informatisées".
Selon [4] et [14], il existe plusieurs types de corpus qui se déclinent selon les critères suivants :

1. Evolutions temporelles : corpus diachroniques et corpus synchroniques.
2. Niveaux de traitement : corpus annotés et corpus non-annotés.
3. Structures : corpus de structure équilibrée et corpus de structure aléatoire.
4. Usage : corpus généraux et corpus spécialisés.
5. Façons d'expression : corpus parlants et corpus de textes.
6. Nombre de langues : corpus monolingues et corpus multilingues (bilingues).

Depuis les années 90, la linguistique de corpus s'est bien développée, surtout dans l'aspect multilingues des corpus en raison de l'internationalisation. Le besoin d'échanges multilingues et les traitements automatiques associés deviennent de plus en plus importants. Les corpus multilingues sont des corpus qui contiennent au moins deux langues différentes et s'ils ne comportent que deux langues différentes, ils sont appelés corpus bilingues. Il y a en principe deux types de corpus multilingues : les corpus parallèles et les corpus comparables. Nous présentons succinctement ces deux corpus dans les sous-sections suivantes.

2.2.1 Corpus parallèles

[86] considère que les corpus parallèles sont des corpus qui contiennent des textes originaux et des textes de traduction dans au moins deux langues différentes. Dans [14], on trouve également une définition similaire d'un corpus parallèle : "c'est un ensemble de textes accompagnés de leurs traductions dans une ou plusieurs langues". Il s'agit donc d'un ensemble de

paires de textes tels que, deux à deux, dans chaque paire, ces textes sont des traductions l'un de l'autre. Ces corpus sont produits surtout par les grands organismes comme les Nations Unies, l'Union Européenne et autres organismes internationaux.
Les principaux corpus parallèles exploités à des fins expérimentales sont :

1. Le corpus Europarl [67] : ce corpus rassemble des textes du Parlement Européen dans 11 langues : il contient plus de 20 millions de mots par langue. Ce corpus constitue une référence importante pour le TALN. Dans nos expérimentations, nous l'avons également utilisé pour évaluer la qualité des mesures de comparabilité.
2. Le corpus Hansard [56] : ce corpus est issu des transcriptions des débats du parlement canadien de 1970 à 1988. Il contient plusieurs dizaines de millions de mots, et il est composé de textes anglais et de textes français.
3. Le corpus Hong-Kong Hansard : ce corpus a été créé par le Linguistic Data Consortium. Il rassemble des textes en anglais et en français issus des discussions et rapports du parlement de Hong Kong.
4. Le corpus de l'UBS (Union des banques suisses) [40] : ce corpus regroupe des rapports sur le développement de l'économie suisse dans quatre langues (anglais, français, allemand, italien).
5. Le corpus InterCorp [21] : ce corpus est riche, puisqu'il contient 31 langues et au moins quelques millions de mots pour les langues principales.
6. Le corpus ITU CRATER : ce corpus est constitué des rapports de l'Union internationale des télécommunications, contenant environ un million de mots pour trois langues (anglais, français et espagnol).
7. Le corpus TradooIT : ce corpus contient quelques centaines de millions de mots pour les trois langues : anglais, français et espagnol.
8. Le corpus JRC-Acquis [131] : ce corpus est issu de l'ensemble des lois applicables dans l'Union Européenne. Il couvre 22 langues et contient quelques dizaines de millions de mots par langue.

Les corpus parallèles ont une grande importance dans le domaine de la traduction automatique ou assistée, [98], [157], [17], [82], [68], [101], de l'extraction des terminologies [41], [12], [81], [74] ou de la construction des dictionnaires bilingues [64], et de la recherche d'information multilingue (CLIR) [9], etc. Malheureusement, ils sont coûteux à développer et souvent difficilement transposables d'un domaine de spécialité à l'autre [83]. A cause de ces limites, des recherches [80], [109], [155] ont tenté d'utiliser les URL, la structure des pages web et leur contenu pour extraire automatiquement des textes parallèles. Cependant, ces approches ne résolvent que partiellement les besoins : les contraintes liées aux domaines et aux langues d'intérêt subsistent. Ce sont les raisons pour lesquelles de nombreux travaux de recherche se sont tournés vers la constitution des corpus comparables, principale motivation de cette thèse.

2.2.2 Définitions des corpus comparables

La notion de corpus comparable a été initialement proposée par [7]. Les auteurs indiquent que ce sont des textes sans contrainte de traduction entre eux, mais "certainement" similaires. En 1996, [125] du EAGLES (Expert Advisory Group on Language Engineering Standards Guidelines) a proposé une autre définition pour les corpus comparables : ce sont des textes de différents types dans une seule langue ou des textes similaires dans au moins deux langues différentes. [86] considère également que les corpus comparables comportent la caractéristique bilingue et multilingue mais sans contrainte de traduction. Les textes de langues différentes sont donc indépendants et originaux dans les corpus comparables. D'autre part, [39], [96] ont proposé une autre définition : un corpus comparable est un corpus qui couvre un thème similaire et transmet des informations qui se chevauchent. [138] a également proposé une définition : les corpus comparables sont des corpus en deux ou plusieurs langues ayant une composition ou une structure similaire (ou quasi-similaire). [14] a complété plus tard cette définition : les corpus comparables sont composés de documents en plusieurs langues, sans lien de traduction entre eux, mais qui partagent certaines caractéristiques.
Pour conclure sur une définition des corpus comparables ci-dessus : nous considèrerons que ce sont des textes traitant d'un même sujet qui sont écrits dans plusieurs langues différentes, certainement similaires mais sans traduction mutuelle. Par exemple, un journal anglais et un journal français qui publient une nouvelle internationale sur un même événement (l'une en anglais et l'autre en français), mais dont les deux auteurs sont différents (absence de traduction de l'une vers l'autre), produisent une paire de documents comparables.
Toutes ces définitions restent malgré tout assez générales. La définition la plus "opérationnelle" est proposée par [30]. C'est une définition quantitative de la 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 $\mathcal{L}_1$*". Par ailleurs, [30] classe les corpus comparables en exploitant des critères qualitatifs qui contiennent le genre, l'auteur, la période, le média, etc. et des critères quantitatifs qui sont basés sur les mesures de fréquences de certains traits linguistiques. Le degré de comparabilité varie en fonction des différents critères choisis.

Par rapport aux corpus parallèles, les corpus comparables ont en principal trois avantages selon [142] : premièrement, ce sont des ressources riches et larges : en volume et en période temporelle couverte ; deuxièmement, il n'existe plus de contrainte sur la traduction limitée au texte original car les corpus comparables fournissent des ressources linguistiques originales et thématiques. Enfin, ils sont moins coûteux à développer que les corpus parallèles.

2.2.3 Applications des corpus comparables

Les corpus comparables peuvent être exploités dans plusieurs domaines [158] : l'extraction des lexiques bilingues [38], [37], [30], [94], [55], [79], [136] ou l'extraction des terminologies [39], [42], [143], [93], [72], [152], la fouille de données multilingues, la traduction automatique

ou assistée, l'apprentissage des langues [141], etc.

1. Extraction des lexiques bilingues ou des terminologies

   Les auteurs Fung et McKeown dans [38] ont proposé une méthode basée sur l'analyse du contexte lexical et une dépendance lexicale basée sur une observation simple : un mot et sa traduction ont tendance à se présenter dans un même contexte lexical. Cette méthode devient la méthode standard dans le domaine de l'extraction des lexiques bilingues. Dans [37], les auteurs ont proposé une méthode "DKvec" pour extraire des lexiques bilingues anglais/japonais et anglais/chinois issues de corpus parallèles bruités (lorsque certaines phrases d'un texte ne sont pas traduites dans un autre texte où les frontières de phrases ne sont pas claires.) et de corpus comparables. Les précisions obtenues sont apparemment bonnes. Plus récemment, dans [30], les auteurs ont proposé une extension de la méthode standard afin de diminuer la dépendance de la couverture du dictionnaire bilingue. Cette extension est basée sur l'intuition que les mots partageant le même sens partageront les mêmes contextes. Dans [94], les auteurs ont vérifié que la représentativité (la qualité) des corpus comparables est plus importante que leur volumétrie en testant sur une tâche d'extraction des termes bilingues français/-japonais. Dans [55], les auteurs ont proposé une méthode basée sur la notion de termes du domaine : ce sont les termes les plus contextuellement pertinents et importants du domaine traité. Cette méthode, permettant de détecter et traiter les termes de contexte du domaine au lieu des termes de contexte général, ne nécessite pas de dictionnaire bilingue de grande taille. Les auteurs ont proposé dans [79] une méthode basée sur le clustering, avec une nouvelle approche intégrant la comparabilité. Celle-ci exploite une notion d'homogénéité du corpus, la plupart du vocabulaire du corpus original étant préservée. Dans [136], les auteurs ont proposé une approche sur le graphe de la similarité de relation de co-occurrence (directe ou indirecte) des termes sous une hypothèse : un mot et sa traduction ont tendance à avoir une relation de co-occurrence similaire (directe ou indirecte) avec tous les grains inter-lingues (un grain est une paire de traduction). Une relation directe est qu'un terme a une relation de co-occurrence avec un autre terme et une relation indirecte est qu'un terme n'a pas de cette relation directe avec un autre terme dans le graphe mais ils peuvent être inter-connectés via un terme intermédiaire. Cette approche permet de capturer les relations directes et indirectes de co-occurrence pour tous les grains afin de construire un graphe de similarité de relation de co-occurrence. Après la construction de ce graphe (un nœud est un terme et un arc est un lien de similarité), une technique de propagation d'étiquettes (les noms des termes et les similarités dans leur contexte) basée sur graphe [161] est appliquée pour transmettre les étiquettes d'un nœud étiqueté vers un nœud non étiqueté afin d'obtenir la distribution des étiquettes de chaque nœud. A partir de ces distributions, les grains sont finalement extraits. Les approches pour l'extraction des terminologies sont semblables à celles développées pour l'extraction des lexiques bilingues. La plupart des chercheurs

ont utilisé les corpus comparables pour acquérir de nouveaux mots et des paires de traduction candidates, propres à la terminologie du domaine spécialisé traité. Leur idée est également basée sur l'hypothèse qu'un terme dans une langue et le terme lui correspondant dans une autre langue ont un contexte similaire. Dans [143], les auteurs ont utilisé les similarités de contextes de document pour obtenir des paires de documents alignés, et pour chaque paire de documents alignés, les similarités de translittérations (basée sur les séquences de caractères, les couplages de sous-chaînes de caractères, la monotonie de l'alignement, etc.) sont calculées pour effectuer l'extraction des entités nommées. Par ailleurs, dans [39], [42], [93], les auteurs ont utilisé les informations de contexte pour effectuer l'extraction. Dans [72], les auteurs ont utilisé une mesure hybride non-supervisée qui combine des traits statistiques, lexicaux, linguistiques, contextuels et temporels en exploitant l'algorithme EM (espérance-maximisation) [31] (permettant de trouver le maximum de vraisemblance) pour extraire des terminologies bilingues. Dans [152], l'auteur a utilisé les corpus comparables pour extraire la collocation de deux thèmes "culture" et "cultiver" en anglais, français et italien.

2. Fouille de données multilingues
   Afin d'essayer de résoudre les problèmes liés à la volumétrie ou aux contraintes temporelles associés aux corpus parallèles, [11], [153], [95], [97] ont extrait des textes parallèles dans des corpus comparables en se basant sur l'alignement des phrases et des paragraphes.
3. Traduction automatique ou assistée
   [96] a utilisé des phrases parallèles extraites de corpus comparables de journaux pour améliorer la performance d'un système de traduction automatique et obtenu des performances satisfaisantes. Par contre, [121] a directement utilisé des corpus comparables afin de trouver des traductions équivalentes pour des expressions.

Cependant, la plupart de ces recherches soit est limitée par le volume des corpus comparables disponibles, soit ne détaille pas le processus de constitution de corpus comparables, soit la qualité d'alignement n'est pas toujours bonne. [134], [87], [78] ont vérifié que la qualité d'alignement des corpus comparables est plus importante que leur volume. Dans la section suivante, nous présentons les approches principales développées pour la constitution de corpus comparables.

2.2.4 Constitution des corpus comparables

La croissance rapide des sources d'informations sur Internet fournit une réelle opportunité pour la construction des corpus comparables. En particulier les pages de nouvelles issues des agences de presse disponibles en différentes langues, ou encore Wikipédia sont des ressources multilingues volumineuses, riches, exploitables, accessibles et en général libres de droit.

Avec l'augmentation des besoins en matière de corpus comparables, la qualité de ces derniers est devenue critique. Le point central de la construction des corpus comparables est l'ali-

gnement des documents ou clusters de documents entre langue source et langue cible. Plus les documents alignés sont similaires ou comparables, meilleur est l'alignement, et plus le corpus comparable produit est exploitable.

Beaucoup de recherches ont été menées pour construire des corpus comparables. Au début, des approches assez rudimentaires ont été exploitées. Par exemple, [123] a simplement utilisé la date de publication et la similarité de thésaurus (en considérant les documents comme la caractéristique d'indexation et les termes comme les éléments de recherche) pour construire la relation d'alignement entre des textes italiens et des textes allemands. Sur cette base, [15] a intégré un indicateur dans la construction des corpus comparables anglais (publiés par AP : Associated Press) et allemand (publiés par l'agence SDA suisse). Cet indicateur est créé par le mot qui a la fréquence moyenne dans tous les textes anglais parmi tous les mots. Cet indicateur est ensuite traduit par le dictionnaire bilingue anglais-allemand et utilisé comme une requête dans le corpus allemand. Les similarités obtenues et les dates sont utilisées pour organiser les corpus comparables. Par ailleurs, cette approche a permis de construire des corpus comparables français-allemand issus de l'agence SDA en utilisant les types des nouvelles, les terminologies, les valeurs numériques, etc. Par ailleurs, [108] a proposé une approche pour fouiller les corpus comparables en exploitant l'hypothèse suivante : si le contenu de pages Web existantes en différentes langues sont comparables, celles-ci possèdent une structure similaire, comme les titres, les paragraphes, etc. Nous pouvons constater qu'initialement la construction des corpus comparables est relativement empirique et hétérogène. Les approches proposées ne tiennent pas beaucoup compte de la qualité de l'alignement des textes obtenus en sortie.

Récemment, [137] a proposé une approche basée sur la corrélation des fréquences de mots d'un même thème exprimé en différentes langues dans des corpus comparables sous une hypothèse que les distributions de fréquences des mots thématiques en différentes langues sont souvent corrélées. Cette approche dépend uniquement des corpus comparables. [95] est le premier à utiliser un dictionnaire bilingue pour transformer les textes sources en textes en langue cible pour obtenir les 5 premières traductions (top-5) comme requête pour chercher dans les textes de la langue cible sur même période. En fonction des similarités obtenues, les K premiers documents de la langue cible (top-K) sont choisis en regroupant les paires de textes comparables de 1 à K. De même, [135] a utilisé la recherche d'information multilingue pour construire des corpus comparables anglais-suédois. Néanmoins, pour éviter la traduction du texte entier, seules les informations importantes sont extraites et traduites, puis recherchées dans le système de recherche d'information. Afin d'améliorer la qualité de l'alignement, les résultats de la recherche sont filtrés. [102] a fouillé des corpus comparables issus de Wikipédia en définissant un thème et les langues (la langue source et la langue cible) pour collecter les documents similaires à ce thème. Par ailleurs, [149] a proposé une approche d'alignement de documents basée sur les caractéristiques (TNC (titre et contenu), LIU (unité indépendante linguistique) et MTD (Distribution des termes monolingues)) et obtenu des résultats satisfaisants.

En résumé, on recense principalement trois types d'approches pour la constitution des corpus comparables :

1) L'approche basée sur la distribution de fréquences des mots
2) L'approche basée sur les caractéristiques (TNC (titre et contenu), LIU (unité indépendante linguistique) et MTD (distribution des termes monolingues))
3) L'approche basée sur la recherche d'information multilingue ("Cross-language information retrieval (CLIR)").
Nous allons entrer un peu plus dans le détail de ces approches dans la sous-section suivante.

2.2.4.1 Approche basée sur la distribution des fréquences des mots

[137] a proposé une approche qui ne dépend pas des ressources externes (comme les dictionnaires bilingues) pour fouiller des textes bilingues comparables. Cette approche est basée sur l'état de la distribution de la fréquence des termes sur une certaine période pour obtenir la relation entre les mots de langue source et les mots de langue cible sous l'hypothèse que les fréquences d'une paire constituée d'un mot et de sa traduction sont corrélées dans les textes comparables. Plus les fréquences des termes sont similaires, plus il est probable qu'ils décrivent le même sujet. Les similarités sont estimées via le coefficient de Pearson, comme indiqué dans l'équation suivante :

$$r(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{N}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)}} \tag{2.1}$$

Où : $x_i = \frac{c(x,s_i)}{\sum_{j=1}^{n} c(x,s_j)}, y_i = \frac{c(y,t_i)}{\sum_{j=1}^{n} c(y,t_j)}$, x est un mot de la langue source, y est un mot de la langue cible, n est le nombre de dates dans une période temporelle, $c(x,s_i)$ est la fréquence du mot x dans le document s de la langue source à la date i, $c(y,t_i)$ est la fréquence du mot y dans le document t de la langue cible à la date i.

Selon l'équation 2.1, nous pouvons ainsi obtenir la similarité de chaque paire de mots, puis calculer la similarité entre un document source et un document cible en faisant la somme de la similarité pondérée de chaque paire de mots comme indiqué dans l'équation suivante :

$$s(d_s,d_t) = \sum_{x \in d_s, y \in d_t} r(x,y) \times IDF(x) \times IDF(y) \times BM25(x,d_s) \times BM25(y,d_t) \tag{2.2}$$

Où : $IDF(x) = log\frac{n+1}{df(x)}$, $BM25(w,d) = \frac{k_1 c(w,d)}{c(w,d)+k_1(1-b+b\frac{|d|}{AvgDocLen})}$ d_s est un document source, d_t est un document cible, IDF(x) [127] est la fréquence inverse dans le document du mot x, IDF(y) est la fréquence inverse dans le document du mot y. df(x) est le nombre de documents qui contiennent le mots x. BM25(w,d) [110] est une mesure standard en recherche d'information, pour laquelle k_1 et b sont deux paramètres ajustables, |d| est le nombre de mots dans le document, c(w,d) est la fréquence du mot w dans le document d, AvgDocLen est le nombre moyen de mots dans les documents du corpus.

BM25 [110] et IDF [127] sont deux mesures souvent utilisées en fouille de textes. La

combinaison des ces deux mesures permet de diminuer les poids des mots les moins discriminants (les moins fréquents) et augmenter les poids des mots les plus discriminants (les plus fréquents). Pour construire des corpus comparables, les auteurs ont calculé les similarités entre chaque document source et tous les documents cibles, extrait les documents les plus similaires, et construit ainsi une paire de documents comparables.

L'approche basée sur la distribution des fréquences de mots pour construire des corpus comparables est adaptée à n'importe quelle langue et permet d'éviter la limitation liée aux ressources externes comme les dictionnaires bilingues (ce qui est un avantage, surtout pour les langues peu dotées en ressources numériques). Cependant, le calcul est très lourd (il faut calculer, sur une certaine fenêtre temporelle, les distributions des fréquences de mots) et donc l'approche est peu efficace, passe mal à l'échelle et n'est pas adaptée pour construire des corpus comparables de grande taille. De plus, cette approche dépend principalement des statistiques des fréquences de mots, elle ne peut donc pas garantir la qualité de l'alignement des corpus comparables produits.

2.2.4.2 Approche basée sur les caractéristiques

[149] propose une approche basée sur les caractéristiques pour aligner des documents comparables, comme indiqué dans la figure 2.1. Après avoir filtré les documents de la langue source et de la langue cible par exploitation d'une fenêtre temporelle, des champs "titre" et "contenu", les paires de documents alignés candidates sont obtenues. Ensuite, les trois caractéristiques : TNC (titre et contenu), LIU (unité indépendante linguistique) et MTD (distribution des termes monolingues) sont extraites de ces paires. Les valeurs de ces trois caractéristiques sont regroupées pour obtenir la similarité des paires de documents et en fonction de ces similarités, l'alignement des documents comparables est finalement établi. Nous détaillons ci-dessous les 4 étapes constitutives de cette approche.

— **ETAPE-1 : Création des paires de documents comparables candidates.**
Afin d'aligner les documents de la langue source avec les documents comparables correspondants dans la langue cible, le calcul des similarités entre chaque document de la langue source et tous les documents de la langue cible est nécessaire. Pour éliminer certains documents non pertinents, deux mesures de filtrage sont utilisées : un filtrage temporel et un filtrage qui porte sur les champs "titre" et "contenu". Pour le filtrage temporel, la date de publication, souvent présente, peut être exploitée. On suppose que si les instants de publication des documents sont proches, ils seront temporellement similaires. Nous pouvons donc définir une même période d'analyse pour la langue source et également pour la langue cible. Cela permet d'éliminer beaucoup de documents et ainsi diminuer la complexité de calcul. Un deuxième niveau de filtrage peut être effectué : c'est le filtrage sur les champs "titre" et "contenu". En utilisant le dictionnaire bilingue et en fonction d'un seuil portant sur le nombre de mots (le nombre de mots du titre seul par exemple), on traduit cette fenêtre de mots vers la langue cible et on ne considère

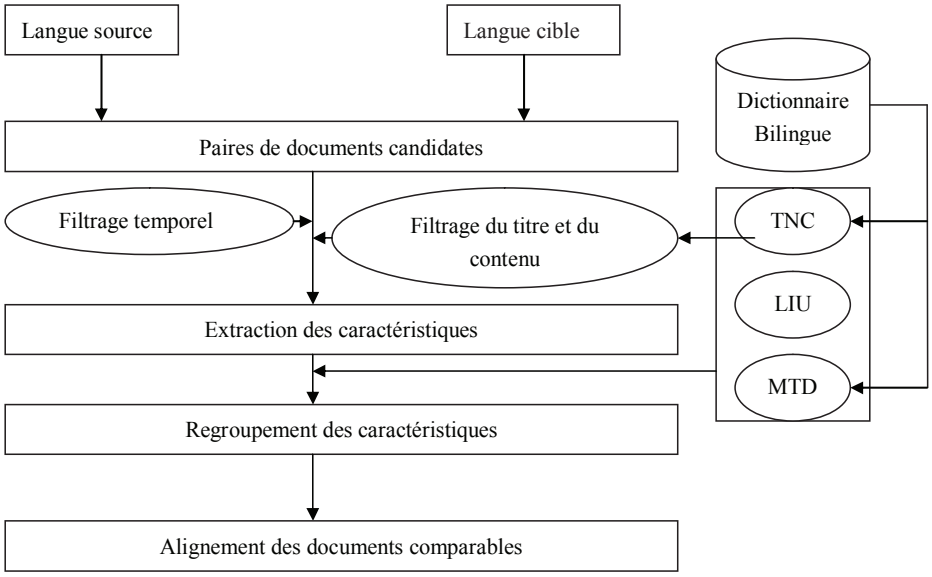

FIGURE 2.1 – Approche basée sur les caractéristiques TNC, LIU et MTD, pour l'alignement des documents

que les documents de la langue cible qui ont au moins un mot présent dans la traduction.

— **ETAPE-2 : Extraction des caractéristiques : TNC, LIU et MTD.**
Ce module extrait les caractéristiques des documents et les répartit en trois groupes : TNC (titre et contenu), LIU (unité indépendante linguistique) et MTD (Distribution des termes monolingues).

a. TNC : titre et contenu

Comme pour le filtrage des champs "titre" et "contenu", on effectue le décompte du nombre de traductions des mots dans le titre d'un document de la langue source qui sont présentes dans les documents de la langue cible, ainsi que le nombre de traductions des mots dans le titre d'un document de la langue cible qui sont présentes dans les documents de la langue source. Formellement :

$$TNC(d_s, d_t) = \sum_{w_i \in T_s} TR(w_i, c_t) + \sum_{w_j \in T_t} TR(w_j, c_s) \tag{2.3}$$

Où : c_t est le contenu du document d_t et c_s est le contenu du document d_s ; T_s est l'ensemble des mots dans le titre des documents d'une langue et T_t est l'ensemble des mots dans le titre des documents d'une autre langue. TR est une fonction indicatrice : si c contient la traduction de w, alors, TR(w,c)=1, sinon, TR(w,c)=0.

b. LIU : unité indépendante linguistique

Les LIU sont des unités qui ont les mêmes orthographes (translittération) dans

les différentes langues, comme les chiffres arabes, certains sigles, certains noms propres, etc. La valeur associée à la caractéristique LIU est le nombre d'unités communes entre un document de la langue source et un document de la langue cible.

$$LIU(d_s, d_t) = |d_s \cap d_t| \tag{2.4}$$

c. MTD : distribution des terminologies monolingues

MTD est une caractéristique associée à la distribution des terminologies monolingues (les mots et les groupes de mots). Elle exploite la distribution des mots dans les documents. Les terminologies sont moins nombreuses et moins ambiguës que les mots. D'autre part, le dictionnaire bilingue est utilisé pour traduire les terminologies. Dans le document cible, plus les terminologies contiennent de traductions, plus il est probable que les terminologies soient des traductions de l'une vers l'autre. De plus, lors du calcul des similarités, le coefficient de Pearson est remplacé par la transformée de Fourier discrète [3]. Son équation est la suivante :
Soit x et y deux mots dans deux documents différents, $x = \{x_1, x_2, ..., x_N\}$ est la distribution de fréquence du mot x dans tous les N documents, $y = \{y_1, y_2, ..., y_N\}$ est la distribution de fréquence du mot y dans tous les N documents.

$$X_k = \sum_{n=0}^{N-1} x_n \times e^{-i2\pi kn/N}, Y_k = \sum_{n=0}^{N-1} y_n \times e^{-i2\pi kn/N}, 0 \leq k < N \tag{2.5}$$

$$R(x,y) = \left(\sqrt{\sum_{i=0}^{m} (X_{k_i} - Y_{k_i})^2} \right)^{-1} \tag{2.6}$$

Où : x_n est la fréquence du mot x dans le n-ième document, y_n est la fréquence du mot y dans le n-ième document ; X_k et Y_k sont deux séquences périodiques N infinies après la transformée de Fourier discrète, et d'après leurs expérimentations, les auteurs ont choisi les $m = 7$ premières valeurs de X_k et Y_k, X_{k_i} est donc la i-ième valeur de X_k et Y_{k_i} est donc la i-ième valeur de Y_k.
La valeur de MTD est calculée en utilisant l'équation suivante :

$$S_{MTD} = \sum_{x \in t_s, y \in t_t} IDF(x) \times IDF(y) \times BM25(x, d_s) \times BM25(y, d_t) \times R(x,y) \times DicScore(x,y) \tag{2.7}$$

Où : t_s est l'ensemble des terminologies extraites des documents de la langue source et t_t l'ensemble des terminologies extraites des documents de la langue cible ; DicScore(x,y) est le nombre de paires de terminologies et leurs traductions dans le dictionnaire bilingue.

— **ETAPE-3 : Regroupement des caractéristiques.**

Pour construire le système non-supervisé d'alignement des documents comparables, les

trois valeurs de ces trois caractéristiques sont normalisées puis multipliées pour obtenir une valeur de similarité finale. L'équation de normalisation est la suivante :

$$norm(x) = \begin{cases} ln(x+T), & \text{x>(e-T)} \\ 1, & \text{sinon} \end{cases} \tag{2.8}$$

où : $(e-T)$ est un seuil pour déterminer si x est pris en compte ou pas, $e \approx 2,71828$ est le nombre d'Euler, T est un paramètre ajustable (pour les auteurs, la valeur de T choisie est 2,2, par suite, $(e-T) = 0,51828$).

— **ETAPE-4 : Alignement des documents comparables.**
Les paires de documents de la langue source et de la langue cible sont regroupées en choisissant les documents les plus similaires en fonction de la valeur de la similarité finale.

Cette approche combine les caractéristiques hétérogènes comme la distribution de fréquences de termes, la plage temporelle et le dictionnaire bilingue. Elle augmente ainsi l'efficacité et la qualité de la constitution des corpus comparables. Cependant, elle nécessite d'extraire l'ensemble des terminologies dans la langue source et dans la langue cible et elle est également influencée par la couverture du dictionnaire bilingue et la qualité d'extraction des terminologies.

2.2.4.3 Approche basée sur la recherche d'information multilingue

Le problème principal de la constitution des corpus comparables est d'aligner les documents de la langue source avec les documents comparables dans la langue cible. En raison des différentes langues, nous pouvons utiliser un système de recherche d'information multilingue pour extraire les documents de la langue cible qui sont similaires avec les documents de la langue source.

La recherche d'information multilingue vise à récupérer des informations écrites dans une langue différente de la langue de la requête de l'utilisateur. Elle permet aux utilisateurs d'accéder aux sources d'information existantes en plusieurs langues. En fonction des situations de la traduction, [100] divise les différentes approches selon quatre manières d'aligner les requêtes de la langue source avec les documents de la langue cible :

1) Sans traduction

2) Avec traduction de l'état de l'art

3) Avec traduction vers une langue pivot

4) Avec traduction en requêtes

1. Sans traduction

   Cette approche est basée sur les similarités orthographiques et phonétiques dans les différentes langues, sans faire appel à un mécanisme de traduction.

2. Avec traduction de l'état de l'art

   La traduction de l'état de l'art consiste à traduire tout l'état de l'art connu ou disponible de la langue cible vers la langue source et ensuite d'effectuer la recherche d'information monolingue afin de récupérer les documents de la langue source pertinents vis-à-vis de l'état de l'art. Cependant, comme la qualité de la traduction automatique n'est pas toujours satisfaisante, le travail de la traduction de l'état de l'art de la langue cible est coûteux. C'est pourquoi cette méthode est plutôt bien adaptée à la constitution des corpus comparables de petit volume.

3. Avec traduction vers une langue pivot

   Le cœur de la recherche d'information multilingue est la traduction entre la langue source et la langue cible. Beaucoup de méthodes de recherche d'information multilingue exploitent des mécanismes de traduction automatique, des dictionnaires bilingues et/ou des corpus. Pour résoudre le manque de ressources disponibles pour certaines langues (surtout les langues peu dotées en ressources numériques), une langue pivot peut donc être intégrée et la langue source et la langue cible sont traduites en langue pivot pour pouvoir effectuer une recherche d'information en langue pivot.

4. Avec traduction en requêtes

   La traduction en requêtes est la façon la plus utilisée en recherche d'information multilingue. La procédure consiste à traduire les questions posées dans la langue source sous la forme de requête et d'effectuer la recherche d'information monolingue. Cette approche transforme les requêtes de la langue source, mais elle reste homogène à une recherche d'information monolingue dans un certain sens. La procédure est schématisée dans la Figure 2.2.

   Conformément à la Figure 2.3, [135] a proposé une façon de construire des corpus comparables basée sur la traduction en requêtes. Les auteurs ont extrait des informations de la langue source qu'ils traitent comme une requête. Ils l'ont traduite en utilisant une technique de traduction vers une requête de la langue cible. Ensuite, ils ont effectué la recherche d'information dans la langue cible et obtenu des paires de documents comparables. Enfin, ils ont proposé un alignement entre les documents de la langue source et les documents de la langue cible. Ils ont montré ainsi la faisabilité de la construction des corpus comparables par des méthodes issues de la recherche d'information multilingue.

Toutes les études décrites précédemment présentent un même schéma d'organisation des processus de traitement mis en œuvre dans la production des corpus comparables :

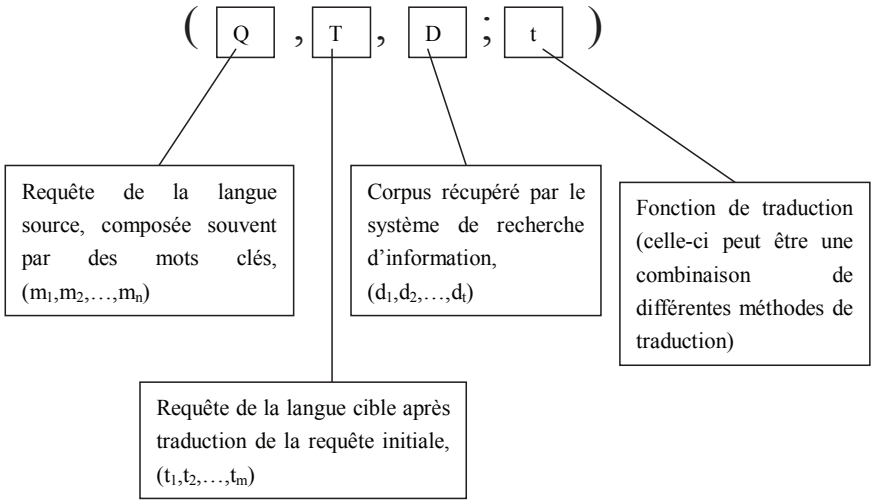

FIGURE 2.2 – Modèle de la traduction des requêtes

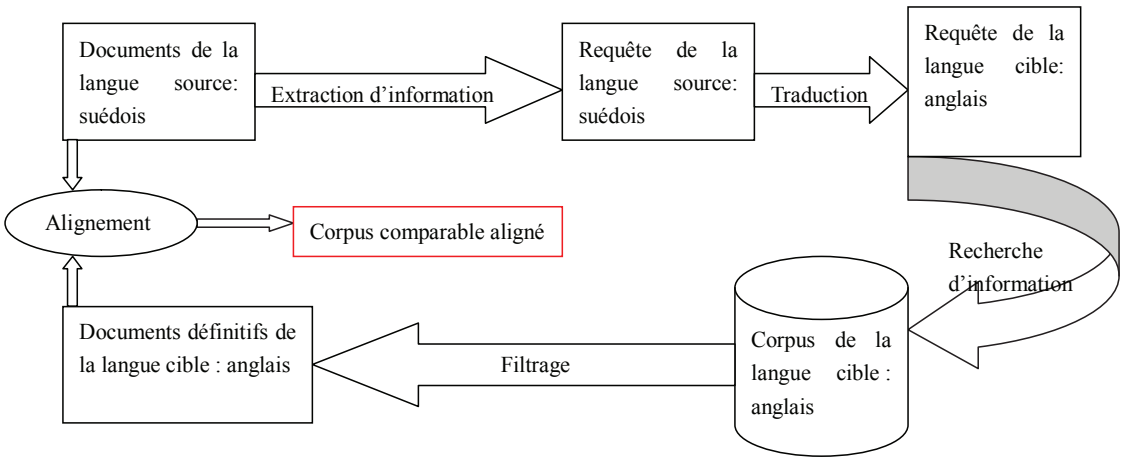

FIGURE 2.3 – Processus de constitution des corpus comparables suédois/anglais basé sur la recherche d'information multilingue [135]

1) constitution des corpus initiaux (contenant le corpus de la langue source et de la langue cible)

2) filtrage des documents les plus pertinents vis-à-vis des besoins

3) utilisation de techniques pour établir l'alignement des documents comparables entre la langue source et la langue cible pour obtenir les corpus comparables

Cependant, les corpus comparables existants actuellement possèdent quelques lacunes : soit la qualité de l'alignement n'est pas suffisamment élevée ; soit leur taille est trop petite ; soit les langues ne correspondent pas au besoin. Tout cela conduit à rechercher des approches novatrices ou consolidées afin de construire des corpus comparables de bonne qualité.

2.2.5 Mesures de comparabilité pour évaluer la qualité de comparabilité

Notre état de l'art montre qu'il y a peu de travaux sur les mesures de comparabilité proprement dites. On recense principalement les références suivantes : [113], [77], [103] et [132].

Dans [113], les auteurs déduisent un score de comparabilité globale de la similarité de toutes les paires de documents inter-linguistiques "cross-language documents" sous hypothèse que plus les documents sont similaires, plus leurs contextes de mots sont similaires. Soit C_1 un corpus dans la langue L_1 constitué de m documents $d_1^i (i = 1,2,\ldots,m)$ et soit C_2 un corpus dans la langue L_2 consitité de n documents $d_2^j (j = 1,2,\ldots,n)$. Tout d'abord, la similarité entre deux documents de langues différentes est calculée en utilisant l'outil *"Dokusare"*. Cet outil permet d'obtenir la similarité entre les deux documents de langues différentes et est décrit dans [114]. On peut ainsi obtenir une matrice DM de $(n * m)$, où chaque élément $DM_{ij} = S(d_1^i, d_2^j)$, qui correspond à l'élément sur la ligne i et la colonne j, est la similarité entre d_1^i et d_2^j. Les auteurs définissent ensuite un processus appelé EMD (basé sur la notion de flux au sens physique) pour estimer à partir de la matrice DM de similarité, un score global pour le corpus bilingue.

Dans [103], les auteurs ont proposé deux variantes de mesure de comparabilité basées sur le ratio entre deux fois la somme des liens inter-linguistiques (traductions) et la somme des tailles des deux vocabulaires dans les deux langues différentes. Les auteurs considèrent un corpus comparable C d'articles issus de Wikipédia, constitué par exemple par une partie portugaise C_p et une partie espagnole C_s. Pour chaque terme t_p dans le vocabulaire C_p^v de C_p, un coefficient de comparabilité peut être défini à partir de la recherche de son lien inter-langue (ou traduction) dans le vocabulaire C_s^v de C_s. Le vocabulaire associé à un corpus Wikipédia est constitué de l'ensemble des "liens internes" trouvés dans ce corpus. Ainsi, les deux parties du corpus, C_p et C_s, ont tendance à avoir un haut degré de comparabilité si nous trouvons de nombreux liens internes à C_p^v qui peuvent être traduits (par le moyen des liens inter-langues) dans de nombreux liens internes à C_s^v. Soit $Trans_bin(t_p, C_s^v)$ une fonction binaire qui renvoie 1 si la traduction du terme portugais t_p se trouve dans le vocabulaire espagnol C_s^v. La première variante est alors définie par :

$$Dice_{bin}(C_p, C_s) = \frac{2\sum_{t_p \in C_p^v} Trans_{bin}(t_p, C_s^v)}{|C_p^v| + |C_s^v|} \tag{2.9}$$

Pour éviter l'influence des liens internes communs (les liens présentés dans la plupart des articles), les auteurs ont proposé une autre variante en tenant compte de la pondération tf-idf tel que défini en Equation 2.10.

$$Dice_{tf\_idf}(C_p, C_s) = \frac{2\sum_{t_p \in C_p^v} Trans_{tf\_idf}(t_p, C_s^v)}{\sum_{t_p \in C_p^v} tf\_idf(t_p) + \sum_{t_s \in C_s^v} tf\_idf(t_s)} \tag{2.10}$$

Plus récemment, dans [132], les auteurs ont développé une autre mesure de comparabilité assez différente qui combine plusieurs métriques distinctes : une métrique basée sur l'alignement lexical, une métrique basée sur les mots-clés et des métriques basées sur la traduction automatique.

1. Métrique basée sur l'alignement lexical

   Tout d'abord, les auteurs ont automatiquement construit des dictionnaires bilingues (lorsque ces dictionnaires bilingues ne sont pas disponibles pour les langues peu dotées en ressources numériques, par exemple anglais/slovène ou anglais/lituanien) en utilisant l'alignement des mots à partir de corpus parallèles à grande échelle comme Europarl [67] et JRC-Acquis [131].

   Une fois ces dictionnaires conçus, les auteurs ont réalisé un alignement lexical en exploitant une approche d'alignement mots-pour-mots. Ils ont vérifié si chaque mot est présent dans les entrées du dictionnaire et si oui, la première traduction (la plus probable) est prise comme le mot d'alignement correspondant. Dans le cas où plusieurs traductions existent pour un mot, la deuxième traduction dont la probabilité est supérieure à 0,3 est également prise en compte. Enfin, les auteurs ont utilisé la mesure de similarité cosinus pour calculer le poids de comparabilité des paires de documents obtenus.

2. Métrique basée sur les mots-clés

   Partant de l'intuition que plus deux documents partagent des mots-clés, plus ils sont comparables, les auteurs ont effectué les étapes suivantes :

   D'abord, les auteurs ont traduit les textes non-anglais vers l'anglais en utilisant le dictionnaire bilingue. Ensuite, la pondération tf-idf est appliquée pour effectuer un tri en ordre décroissant et les 30 premiers mots sont gardés pour représenter le texte. Enfin, les auteurs ont utilisé la similarité cosinus pour calculer la valeur de comparabilité entre ces listes de mots-clés.

3. Métriques basées sur la traduction automatique

   Pour diminuer la perte de performance lorsque l'on ignore l'ordre des mots, la structure syntaxique et les entités nommées, les auteurs se sont tournés vers l'exploitation d'un système de traduction automatique (SMT). L'API de traduction automatique de Microsoft [1] a été utilisée pour traduire les langues peu dotées en ressources numériques comme le slovène et le lituanien en anglais et exploiter les caractéristiques suivantes pour la conception de la métrique de comparabilité.

   a. Caractéristique lexicale : la similarité lexicale W_L de chaque paire de documents est obtenue par la mesure de similarité cosinus sur cette caractéristique lexicale après la lemmatisation des mots non-vides.

   b. Caractéristique de structure : Elle est obtenue approximativement par le nombre de mots de contexte C_D (adjectifs, adverbes, noms, verbes et noms propres) et le nombre de phrases S_D dans chaque document avec l'intuition que si deux documents sont hautement comparables, leur nombre de mots de contexte et la taille des

1. http ://code.google.com/p/microsoft-translator-java-api/

documents doivent être similaires. La similarité de structure W_S est définie par :

$$W_S = 0.5 \times (C_{D1}/C_{D2}) + 0.5 \times (S_{D1}/S_{D2}) \tag{2.11}$$

En supposant que $C_{D1} \leq C_{D2}$ et $S_{D1} \leq S_{D2}$.

c. Caractéristique de mots-clés : les auteurs ont sélectionné les 20 premiers mots (par tri sur les poids tf-idf). La similarité "mots-clés" W_K de deux documents est également calculée par similarité cosinus.

d. Caractéristique des entités nommées : les auteurs ont extrait les entités nommées et ensuite utilisé la similarité cosinus pour calculer la similarité "entités nommées" W_N entre une paire de documents de langues différentes.

Enfin, ils ont combiné ces quatre valeurs de similarité selon l'Equation 2.12 pour obtenir la valeur de comparabilité globale :

$$SC = \alpha \times W_L + \beta \times W_S + \gamma \times W_K + \delta \times W_N \tag{2.12}$$

Où α, β, γ et $\delta \in [0,1]$ et $\alpha+\beta+\gamma+\delta = 1$. Dans leur expérience, les auteurs ont utilisé $\alpha = 0,5$, $\beta = 0,2$, $\gamma = 0,2$ et $\delta = 0,1$.

SC est ainsi une valeur comprise entre 0 et 1, et plus sa valeur est grande, plus la comparabilité est élevée.

Les mesures présentées ci-dessus sont relativement complexes à calculer. En dehors de ces travaux, à notre connaissance, il existe seulement un travail qui élabore et évalue une mesure de la comparabilité d'une manière systématique et quantitative. Cette mesure de comparabilité est proposée par Li et Gaussier dans [77]. Elle calcule de manière symétrique vis-à-vis des langues L_1 et L_2, le nombre des mots du vocabulaire source qui ont au moins une tradution présente dans le vocabulaire cible. La valeur définitive est obtenue par la somme de ces deux nombres, normalisée par la somme de la taille du vocabulaire source et la taille du vocabulaire cible. La mesure de comparabilité se présente formellement sous la forme :

$$C_{LG}(C_1,C_2) = \frac{\sum_{w_1 \in WC_1 \cap WD_1} \sigma(w_1) + \sum_{w_2 \in WC_2 \cap WD_2} \sigma(w_2)}{|WC_1 \cap WD_1| + |WC_2 \cap WD_2|} \tag{2.13}$$

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 lexicales 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.

Cette mesure de comparabilité est facile à calculer : dans les expérimentations proposées par les auteurs et que nous avons également reprises dans le chapitre 3, nous avons pu vérifier que la symétrie est très importante pour calculer la comparabilité.

Nous pouvons qualifier cette mesure de comparabilité "traductionnelle" dans la mesure où elle est bien adaptée à une tâche d'aide à la traduction, mais pas nécessairement adaptée à des tâches de classification ou de clustering de documents bilingues thématiques.

2.3 Clustering et classification de textes

Le clustering et la classification peuvent être vus comme deux techniques issues d'un apprentissage automatique.

La classification consiste à trouver un modèle (ou une fonction) pour décrire et identifier les classes de données ou les concepts afin d'être en mesure de prédire les classes pour des données non étiquetées. Le but de la classification est d'apprendre une fonction de classification ou un modèle de classification (souvent appelée classifieur) sur la base de données d'entraînement. Le classifieur entraîné peut ainsi proposer pour les données non classées une ou plusieures classes de rattachement. Il s'agit donc d'une technique supervisée à base d'apprentissage automatique. La classification est généralement dépendante des caractéristiques qui décrivent les données et il n'existe pas de classifieur générique pour tous types de données.

Le clustering consiste quant à lui à regrouper les données sans classe dans différents groupes, appelés clusters. L'objectif est de regrouper les données très similaires dans un même cluster et les données peu similaires dans des clusters différents. Le clustering est donc différent de la classification, car on ne connaît pas le nombre de clusters et on ne dispose pas de données pré-étiquetées pouvant servir à entraîner un modèle. Le clustering est uniquement basé sur quelques notions prédéfinies comme des distances ou similarités, des voisinages, etc. Le clustering relève d'une technique non-supervisée d'apprentissage automatique. Le nombre de clusters peut être fixé a priori ou déterminé par l'utilisation d'heuristique ad hoc.

2.3.1 Classification non supervisée : le clustering

Parmi les différentes techniques de clustering, on dénombre principalement quatre types d'approches :

1. Les méthodes basées sur le partitionnement des données comme l'algorithme des k-moyennes (k-means) : [36], [154] ou des k-médoides : [62], [63]
2. Les méthodes basées sur une approche hiérarchique comme dans [124], [29], [63], [47], [48], [61], [160]
3. Les méthodes basées sur une mesure de densité comme dans [57]
4. Les méthodes basées sur les graphes comme dans [69].

Nous allons préciser un peu quelques approches principales : les k-moyennes (k-means), les k-médoides, le clustering hiérarchique et le clustering multilingue.

2.3.1.1 k-moyennes

L'algorithme des k-moyennes fait partie d'un groupe d'algorithmes appelés méthodes de partitionnement. Le problème du clustering partitionné peut être formellement décrit comme suit : soit n données dans un espace métrique de dimension d, on détermine une partition des données en k groupes, ou clusters, de telle sorte que les données d'un cluster sont plus semblables entre eux qu'avec les données localisées dans des groupes différents. Rappelons qu'une partition divise un ensemble en plusieurs parties disjointes qui incluent tous les éléments de l'ensemble. La valeur de k peut être ou ne pas être spécifiée et un critère de clustering, généralement le critère d'erreur quadratique comme celui précisé par l'Equation 2.14, est alors adopté en général.

$$F = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{2.14}$$

Où

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{2.15}$$

et $r_{nk} = 1$ si la donnée n est affectée au cluster k, $r_{nk} = 0$ sinon.

La solution à ce problème est immédiate. On sélectionne un critère de clustering (une fonction de distance ou de similarité), puis pour chaque donnée, on sélectionne le cluster qui minimise le critère. L'algorithme des k-moyennes initialise k clusters en sélectionnant aléatoirement une donnée pour représenter chaque cluster. Chacune des données restantes est affectée à un cluster et le critère de clustering est utilisé pour recalculer la moyenne du cluster. Ces moyennes sont utilisées comme les nouveaux centres de clusters et chaque donnée est réaffectée au cluster le plus proche. On itère jusqu'à ce qu'il n'y ait plus de changement lorsque les clusters sont recalculés. L'algorithme est illustré comme suivant :

Algorithme des k-moyennes :

Données: Une série X de données $\{x_1, x_2, ..., x_n\}$
Résultat: Centres des clusters et leur contenu

```
1 Sélectionner aléatoirement k clusters;
2 Initialiser les centres de cluster avec ces k clusters;
3 tant que les centres de cluster changent faire
4     Partitionnement pour affecter ou réaffecter toutes les données au centre de cluster le
      plus proche;
5     Calcul des nouveaux centres comme la valeur moyenne des données contenues dans
      chaque cluster;
6 retourner Centres des clusters et leur contenu
```

Comme la convergence vers la meilleure solution est liée aux conditions initiales, il est usuel de répéter plusieurs fois cet algorithme et de sélectionner la meilleure solution.

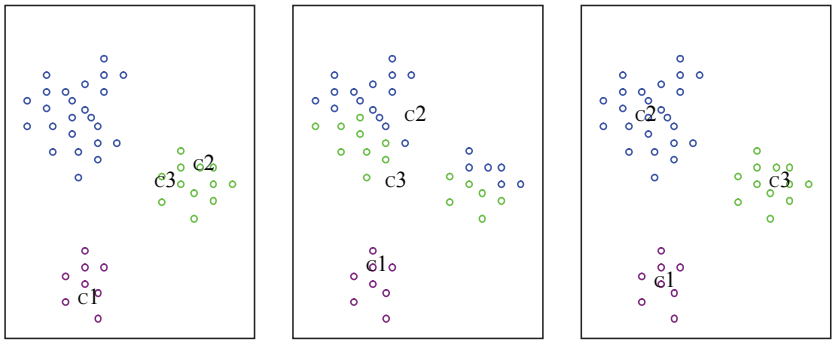

FIGURE 2.4 – Illustration de l'algorithme des k-moyennes : à gauche, les centres de cluster sont aléatoires ; au milieu, les centres de cluster commencent à converger ; à droite, les centres de cluster deviennent stables.

2.3.1.2 k-médoides

L'algorithme des k-médoides peut être considéré comme une variante de l'algorithme des k-moyennes.

Cependant, l'algorithme des k-médoides considère la donnée qui minimise la somme des distances entre elle-même et le reste des données dans un même cluster comme le médoide du cluster au lieu de prendre la moyenne du cluster comme dans l'algorithme des k-moyennes. De plus, le critère à optimiser change. L'algorithme des k-médoides essaie d'optimiser la fonction donnée en Equation 2.16. Pour les k-moyennes, les centres peuvent être n'importe quel élément d'un espace Euclidien continu, par contre, pour les k-médoides, nous ne pouvons choisir les centres que parmi les données à classer, qui, en général, appartiennent à un espace métrique non Euclidien. Par exemple, si nous traitons des individus identifiés par leur taille et leur poids, les k-moyennes sont applicables mais si nous traitons des individus identifiés par des variables sémantiques (caractéristiques de couleur ou de texture par exemple), nous ne pouvons pas utiliser la distance euclidienne pour calculer des distances moyennes.

$$F = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} V(x_n, \mu_k) \tag{2.16}$$

Où V représente une matrice de distance.

Algorithme des k-médoides :

Comme pour les k-moyennes, le résultat de cet algorithme dépend des conditions initiales. En général, on répète plusieurs fois l'exécution de cet algorithme pour retenir la meilleure solution au sens du critère de partitionnement choisi.

Le résultat produit par l'algorithme des k-médoides est similaire à celui produit par l'algorithme des k-moyennes, comme indiqué en Figure 2.4. L'algorithme des k-médoides (en complexité $O(N)^2$) est plus complexe que l'algorithme des k-moyennes (en complexité $O(N)$),

Données: Une série X de données $\{x_1, x_2, ..., x_n\}$
Résultat: Médoides des clusters et leur contenu
1 Sélectionner aléatoirement k données;
2 Initialiser les médoides des clusters avec ces k données;
3 **tant que** *les Médoides des clusters changent* **faire**
4 Partitionnement pour affecter ou réaffecter toutes les données au médoide de cluster le plus similaire;
5 Calcul des nouveaux médoides comme la somme minimale des similarités d'une donnée avec le reste des données dans chaque cluster;
6 retourner Médoides des clusters et leur contenu

mais il est plus général et plus robuste que l'algorithme des k-moyennes, car les k-médoides minimisent la somme de dissemblance au lieu de la somme de distances euclidiennes de sorte qu'ils sont adaptés au traitement de données non représentables dans un espace Euclidien.

2.3.1.3 Clustering hiérarchique

Les algorithmes de clustering hiérarchique peuvent être soit ascendants soit descendants. Tous les algorithmes de clustering hiérarchique ascendant sont initialisés en considérant que chaque donnée est un cluster distinct. Ces clusters sont successivement fusionnés sur la base d'une mesure de similarité jusqu'à ce qu'il n'y ait plus qu'un seul cluster restant ou qu'une condition de terminaison spécifiée soit satisfaite. Pour n données, n-1 fusions sont effectuées. Ces algorithmes hiérarchiques sont rigides en ce qu'une fois que la fusion a été effectuée, celle-ci ne peut pas être annulée. Bien qu'il y ait un faible coût de calcul, ces algorithmes rencontrent des problèmes lorsque des fusions erronées se produisent. Nous décrivons ci-dessous un algorithme simple et classique de clustering ascendant.

Dans le cadre du clustering hiérarchique, le graphique hiérarchique obtenu est appelé dendrogramme. La figure 2.5 montre un échantillon de dendrogramme qui pourrait être produit à partir d'un algorithme de clustering hiérarchique. Contrairement à l'algorithme des k-moyennes, le nombre de clusters (k) n'a pas besoin d'être spécifié dans le clustering hiérarchique. Après constitution de la hiérarchie, l'utilisateur peut spécifier le nombre de classes nécessaires, de 1 à n. Le niveau supérieur de la hiérarchie représente un cluster, ou k = 1. Pour extraire plus de clusters, il suffit de "couper" horizontalement à un niveau plus haut de similarité dans la hiérarchie.

Algorithme de clustering hiérarchique ascendant : Chaque donnée X est initialement utilisée pour créer un cluster contenant un seul élément. Ces clusters sont successivement fusionnés dans de nouveaux clusters, qui sont ajoutés à la série des clusters, C. Quand une paire de clusters est fusionnée, un nouveau lien est créé entre ces deux clusters et les clusters originaux sont supprimés de C. Ainsi, le nombre de clusters dans C diminue jusqu'à ce qu'il n'y ait plus qu'un seul cluster restant, contenant toutes les données de X. La hiérarchie des clusters est

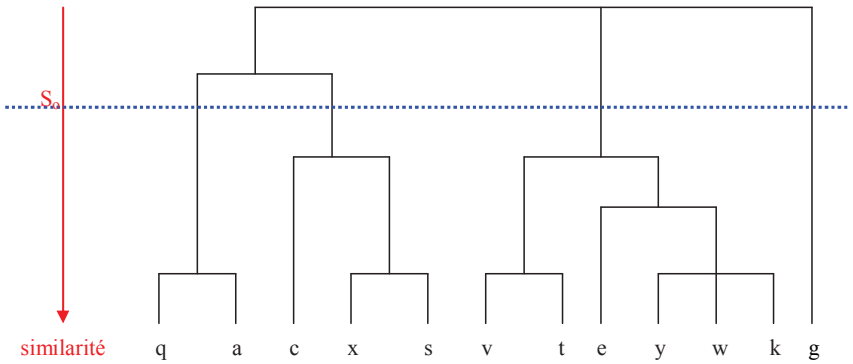

FIGURE 2.5 – Exemple de dendrogramme. Si on coupe horizontalement au niveau du seuil de similarité S_o, nous obtenons 3 clusters : {q,a,c,x,s}, {v,t,e,y,w,k} et {g}

Données: Une série X de données $\{x_1, x_2, ..., x_n\}$ et une fonction de distance $dist(c_1, c_2)$
Résultat: Dendrogramme
pour *i = 1 to n* **faire**
  $c_i = \{x_i\}$;
$C = \{c_1, c_2, ..., c_m\}$;
tant que $|C| > 1$ **faire**
  (c_a, c_b) = argmin $dist(c_i, c_j)$ pour tout c_i, c_j dans C;
  Enlever c_a et c_b de C;
  Créer un nouveau lien qui lie c_a et c_b;
  Ajouter $\{c_a, c_b, lelien\}$ à C;
retourner Dendrogramme

implicitement représentée dans la série imbriquée de C.

La fonction de distance ou de similarité utilisée peut déterminer plusieurs critères de fusion : "single-linkage", "maximun-linkage" et "average-linkage".

1. Pour le "single-linkage", la distance entre deux clusters de données est définie comme la distance la plus petite entre paires de données issues des deux clusters.
2. Pour le "maximun-linkage", également connu sous la terminologie "complete-linkage", la distance entre deux clusters est définie comme la plus grande distance parmi les distances entre paires de données issues des deux clusters.
3. Pour l"'average-linkage", la distance entre deux clusters est définie comme la distance moyenne évaluée sur toutes les distances entre paires de données issues des deux clusters.

En général, l"'average-linkage" est plus robuste et plus stable car il diminue l'effet produit pour les données extrêmes. C'est pour cela que nous avons choisi l"'average-linkage" en testant l'algorithme de clustering hiérarchique ascendant.

Le clustering hiérarchique descendant est inverse du clustering hiérarchique ascendant. Tout d'abord, on considère que toutes les données sont dans un même cluster. Ensuite, on les partitionne recursivement en sous-clusters jusqu'à ce qu'il n'y ait plus de partitionnement possible (chaque donnée est un cluster). Normalement, on partitionne en fonction d'un seuil de partitionnement qui définit la "densité minimale". Si la distance des deux données dans un cluster ou la distance moyenne de toutes les données dans un cluster est plus petite que ce seuil, on ne partitionne plus ce cluster. Pour le cluster le moins dense (la distance entre les données qu'il contient est la plus grande), si la distance des deux données dans ce cluster ou la distance moyenne de toutes les données dans ce cluster est plus petite que ce seuil, l'algorithme termine. L'algorithme du clustering hiérarchique descendant est plus complexe que celui du clustering hiérarchique ascendant car le partitionnement des données est en principe plus difficile que la fusion des données ou des clusters. Le clustering hiérarchique descendant est donc moins utilisé que le clustering hiérarchique ascendant.

Un inconvénient des algorithmes du clustering hiérarchique est qu'ils sont gourmands ("greedy") sans garantir l'obtention d'une solution globalement optimale ou même satisfaisante.

2.3.1.4 Clustering multilingue

Il existe également des méthodes de clustering adaptées au traitement des données multilingues. En général, il existe deux stratégies de clustering multilingue : la première est basée sur un clustering initial et une fusion finale, et la deuxième est basée sur une transformation des textes bilingues en monolingues (soit par une traduction totale, soit par une traduction ou transformation partielle) et un clustering final.

1. Clustering initial et fusion finale

   Dans [22], les auteurs ont proposé une approche de clustering multiligue en deux étapes :

   La première étape consiste à effectuer le clustering sur des textes bilingues (chinois et anglais).

   La deuxième étape consiste à traduire les termes chinois en anglais via un dictionnaire bilingue pour ensuite fusionner les clusters de langues différentes par le biais du calcul d'une similarité entre les clusters obtenus après traduction.

2. Transformation initiale et clustering final

   Dans [16] et [75], les auteurs ont utilisé respectivement un système de traduction automatique pour traduire des textes et ensuite effectuer le clustering bilingue.

   Dans [130], les auteurs ont utilisé "EUROVOC" (une liste de paires de thèmes multilingues) pour traduire des termes, et ensuite effectuer le clustering.

   Dans [85], les auteurs ont utilisé un dictionnaire bilingue pour effectuer une traduction afin d'effectuer un clustering monolingue.

Dans [92], les auteurs ont calculé les similarités des termes issus d'une même source (surtout les entités nommées et surtout pour les langues proches, par exemple, anglais et espagnol) comme les noms des personnes, les noms des organismes et les lieux pour faire le clustering.

Dans [150], les auteurs ont utilisé des corpus parallèles pour produire une traduction préalable à un clustering monolingue.

Dans [10], les auteurs ont proposé une approche qui utilise des méthodes traditionnelles comme dans [18]. La méthode traditionnelle est décrite comme suit : chaque mot est affecté à une classe afin de minimiser l'erreur d'un modèle n-gramme basé sur des classes. Cette approche utilise la méthode traditionnelle pour trouver des classes qui peuvent être utilisées dans la traduction. L'idée principale est de définir un nouveau modèle de langue avec les phrases où les mots sont étiquetés avec leurs traductions. Ces phrases sont ensuite utilisées pour trouver des clusters qui peuvent être utilisés lors de l'apprentissage du traducteur.

En général, ces approches sont très dépendantes de la qualité et/ou de la quantité de traductions et de ressources (les dictionnaires bilingues, les corpus) disponibles.

Par ailleurs, il exsite d'autres types de clustering spécifique comme le co-clustering ("bi-clustering, co-clustering, ou two-mode clustering") [49], [89], [32], [144], le clustering du graphe bipartite [159], etc. Le clustering double est un clustering qui effectue simultanément le clustering pour les lignes (par exemple les termes) et les colonnes (par exemple les documents) d'une matrice. Le clustering du graphe bipartite est basé sur le partitionnement d'un graphe bipartite (biclustering). Ce partitionnement est établi par la minimisation d'une somme normalisée des pondérations des arcs (une pondération est le nombre d'occurrence d'un terme dans un document) entre les paires de nœuds non alignées du graphe bipartite. La décomposition en valeurs singulières (SVD [6]) de la matrice des pondérations des arcs associés du graphe bipartite est utilisée pour faire cette minimisation. Ces approches sont exploitables simultanément pour effectuer un clustering des documents et des termes monolingues.

2.3.1.5 Problème du nombre de clusters K

Le problème de l'estimation du nombre de clusters K est un problème fréquent et difficile dans la tâche de clustering pour les algorithmes k-moyennes et k-médoides. Différentes approches existent pour résoudre ce problème. Cependant, le choix de K est souvent ambigu.

1. Détermination par la règle du "Pouce" (Rule of thumb) [35]

   C'est une mesure très simple. Le nombre de cluster K est directement estimé par $K \approx \sqrt{n/2}$ avec n le nombre de documents dans le corpus.

2. Détermination par la méthode du " Coude" (Elbow) [139]

   L'objectif de cette méthode est d'identifier le nombre de clusters K pour lequel, si on enlève un cluster, cela diminue considérablement la performance du modèle (grande

variance) mais si on ajoute un cluster, la performance du modèle change peu (faible variance).

3. Détermination par les données textuelles [20]

   Dans les données textuelles, une collection de documents est définie par une matrice documents-termes. Le nombre de clusters K peut être estimé approximativement par $(m \times n)/t$, où m est le nombre de documents, n est le nombre de termes et t est un nombre d'éléments non nuls dans la matrice (chaque ligne ou chaque colonne contient au moins un élément non nul).

4. Détermination par l'analyse de la matrice de noyau [53]

   La matrice de noyau est une matrice de similarité. Cette matrice est décomposée en valeurs et vecteurs propres. Les valeurs propres sont ensuite analysées pour obtenir la compacité de la distribution des données. Enfin, K est sélectionné selon la méthode du "Coude".

5. Détermination par "Silhouette" [112]

   Cette méthode est basée sur la matrice de distance. Pour chaque donnée sélectionnée i, on définit un indice $s(i) \in [-1,1]$ pour mesurer l'écart-type entre b(i), a(i) : où a(i) est la distance moyenne entre cette donnée i et les autres données dans le même cluster, b(i) est la distance moyenne entre cette donnée i et toutes les données du cluster voisin le plus proche. Alors, Si s(i) est proche de 1, c'est-à-dire la donnée i est plus proche de son cluster que du cluster voisin le plus proche, ce cluster est classé ; si s(i) est proche de -1, ce cluster est mal classé ; et si s(i) est proche de 0, ce cluster est ambigu. On l'applique pour chaque cluster et enfin prendre la moyenne de {s(i)}. Dans [112], les auteurs ont vérifié que si cette moyenne est plus grande que 0,5, le clustering est bon, par contre, si cette valeur est plus petite que 0,2, le clustering n'est pas très bon.

6. Détermination par le critère de Calinsky [19]

   Le critère de Calinsky cherche à maximiser la fonction suivante en fonction de K : $CH(K) = \frac{B/(K-1)}{W/(N-K)}$.
   Où : N est le nombre de données et k est le nombre de clusters. B désigne la variance globale inter-cluster :

$$B = \sum_{i=1}^{K} n_i \|m_i - \bar{m}\|^2$$

   où n_i est le nombre de données dans le cluster i, m_i est le centre du cluster i et $\bar{m}$ est la moyenne de toutes les données. W est la variance globale intra-cluster :

$$W = \sum_{j=1}^{K} \sum_{i=1}^{N} \|x_i - m_j\|^2$$

   où x_i est une donnée dans le cluster j, m_j est le centre du cluster j.

7. Détermination par le coefficient gamma de Goodman et Kruskal [43], [44], [45], [46]
   On calcule les distances intra-cluster et inter-cluster. Si la distance intra-cluster est strictement plus petite que la distance inter-cluster, alors, cette paire de distances est dite concordante, sinon, elle est dite discordante. L'indice de concordance est définie comme $I(K) = \frac{S_+ - S_-}{S_+ + S_-} \in [-1, 1]$, où S_+ et S_- sont respectivement le nombre de paires concordantes et le nombre de paires discordantes. Le nombre de clusters K est choisi de manière à maximiser le I(K) ($\hat{K} = argmax_k(I(K))$).
8. Autres méthodes
   Il existe d'autres méthodes basées sur les modèles statistiques, comme le critère d'information bayésien [117], la statistique de "gap" [140], etc.

Malheureusement, toutes ces méthodes ne sont pas universelles car soit elles sont limitées par les données disponibles et leur type, soit elles sont limitées par les types de clustering exploitables. Alors, existe-il un type de clustering sans détermination du nombre de clusters K ?

La réponse est oui. Avec le clustering hiérarchique ascendant, c'est a priori le cas. Si l'on veut certain nombre de clusters, il suffit de couper l'arbre à une certaine profondeur, mais une question reste en suspens : quelle profondeur de l'arbre faut-il choisir ? Le problème de détermination du nombre de cluster K est transformé en un autre problème : comment choisir la profondeur convenable dans le dendogramme ? Le problème ne peut donc pas être dépassé.

La méthode que nous avons utilisée dans cette thèse est semblable à celle du critère de Calinsky et celle du coefficient gamma de Goodman et Kruskal car ces deux méthodes ont utilisé les notions de distances intra-cluster et inter-cluster pour déterminer le nombre de clusters K. Selon l'objectif du clustering : rendre les données dans un même cluster très similaires et les données entre les différents clusters assez éloignés, nous pensons que cette idée est simple, directe et naturelle pour déterminer le nombre de clusters K et nous nous en sommes inspiré pour concevoir une variante pour la détermination du nombre de clusters K dans nos expériences. Si on trace la courbe de toutes les valeurs des distances (ou similarités) intra-cluster et inter-clusters en faisant varier la valeur K du nombre de clusters, il est clair que les deux lignes ont un point d'intersection qui peut être considéré comme un "bon" compromis initial entre la qualité de clusters et la nombre de clusters, ce qui nous permet d'obtenir les clusters de qualité satisfaisante et en nombre pas trop petit pour notre tâche de construction de corpus comparables thématiques même si un raffinement itératif basé sur des principes de filtrage est nécessaire pour ajuster la qualité finale des corpus produits.

2.3.2 Classification supervisée : la catégorisation

Parmi les différents types de classification supervisée existants, nous pouvons dégager quelques grandes familles.

Nous trouvons tout d'abord les classifieurs probabilistes qui utilisent un ensemble d'entraînement, c'est-à-dire les données déjà classées, pour estimer les paramètres de la distribution de

probabilité des descripteurs (les mots pour les données textuelles) par rapport aux catégories. C'est dans cette famille que nous retrouvons le classifieur bayésien naïf.

Nous trouvons également des classifieurs exploitant un profil de fonction discriminante, par exemple les classifieurs linéaires. Dans ce contexte, le profil est un vecteur de descripteurs pondérés construit pour chaque catégorie. Ce vecteur est bien sûr construit à l'aide des données d'entraînement. Quand une nouvelle donnée doit être classée, son vecteur de descripteur est alors comparé à ce vecteur "type". Un avantage de cette approche est qu'elle produit un classifieur compréhensible par un humain, dans le sens où le profil de la catégorie peut être interprété assez facilement. Par contre, l'inconvénient principal de tous les classifieurs linéaires est que l'espace est divisé en régions à frontières linéaires, ce qui peut être restrictif, car tous les problèmes ne sont pas nécessairement linéairement séparables. Parmi les nombreux membres de cette famille, nous retrouvons Rocchio, Widrow-Hoff et EG [76]. Les séparateurs à vaste marge s'apparentent aux classifieurs linéaires, dans le sens où ils tentent de séparer l'espace en régions séparées par des hyperplans, mais certaines manipulations mathématiques les rendent adaptables à des problèmes non linéaires.

Il existe également une famille de classifieurs qui se base directement sur les exemples. Les nouvelles données à classer sont comparées directement aux données d'entraînement. L'algorithme des k plus proches voisins est le plus connu de cette famille. Il exploite également un ensemble d'entraînement associé à une fonction de similarité (ou distance). C'est sans doute le classifieur le plus simple à mettre en œuvre, mais pas le moins coûteux en temps de calcul.

Dans les pages qui suivent, trois de ces algorithmes seront exposés plus en détail. D'abord, le classifieur bayésien naïf qui, même s'il est généralement surclassé par d'autres algorithmes, est souvent utilisé comme point de référence en raison de sa simplicité. Ensuite, les séparateurs à vaste marge et l'algorithme des k plus proches voisins, qui représentent vraisemblablement à ce jour les deux meilleurs choix en matière de catégorisation de textes selon [60], [119], [156], [120].

2.3.2.1 Classifieur bayésien naïf

Comme son nom l'indique, ce classifieur exploite le théorème de Bayes [148], [122]. Ce théorème est donné en Equation 2.3.2.1. Il permet de calculer les probabilités a posteriori à partir des probabilités a priori et des probabilités conditionnelles.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.17}$$

où A et B sont deux événements.

L'idée principale du classifieur bayésien naïf est la suivante : pour une donnée à classer, on compare la probabilité à postériori pour chaque classe étant connue la donnée, et on associe la

classe la plus probable à cette donnée. Formellement :

1. Soit $x = \{a_1, a_2, ..., a_m\}$ une donnée à classer où chaque a_i est un attribut de la donnée x.
2. Soit $C = y_1, y_2, ..., y_n$ un ensemble de classes.
3. Calculer $P(y_1|x), P(y_2|x), ..., P(y_n|x)$.
4. Si $P(y_k|x) = max(P(y_1|x), P(y_2|x), ..., P(y_n|x))$, alors, x est affecté à la classe y_k.

Dans cette approche, le plus important est d'évaluer les probabilités a posteriori (étape 3). Ces probabilités peuvent-être estimées comme suit :

1. Déterminer un ensemble de données pour lesquelles la variable de classe est connue (corpus d'entraînement).
2. Calculer les probabilités conditionnelles pour chaque attribut connaissant chaque classe : $P(a_1|y_1), P(a_2|y_1), ..., P(a_m|y_1); P(a_1|y_2), P(a_2|y_2), ..., P(a_m|y_2), ..., P(a_1|y_n), P(a_2|y_n), ..., P(a_m|y_n)$.
3. Supposons que chaque attribut est indépendant, selon le théorème de Bayes, nous obtenons : $P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$. Comme le dénominateur est une constante, nous avons seulement besoin de maximiser le numérateur, et comme les attributs sont indépendants, le numérateur peut être calculé de la manière suivante : $P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i)$.

La procédure associée au classifieur bayésien naïf est décomposée en trois phases explicitées en Figure 2.6.

1. Phase de prétraitement

   L'objectif de cette phase est de déterminer les attributs des données et de classer manuellement un ensemble de données appelé corpus d'entraînement. L'entrée est le corpus entier et la sortie est l'ensemble des attributs et le corpus d'entraînement. C'est la seule phase qui nécessite une intervention manuelle. La bonne sélection des attributs et la bonne qualité du corpus d'entraînement peuvent considérablement influencer la qualité du classifieur.

2. Phase d'entraînement

   Le but de cette phase est d'entraîner le classifieur. On calcule la probabilité a priori de chaque classe présente dans le corpus d'entraînement ainsi que l'estimation de la probabilité de chaque attribut connaissant chaque classe. L'entrée est constituée de l'ensemble des attributs et du corpus d'entraînement, et la sortie est constituée du classifieur entraîné. Cette phase est automatique.

3. Phase d'exploitation

   L'objectif de cette phase est d'exploiter le classifieur entraîné pour classer les nouvelles données. L'entrée est constituée du classifieur entraîné et des nouvelles données à classer, la sortie est constituée de la classe prédite pour chaque nouvelle donnée. Cette phase est également automatique.

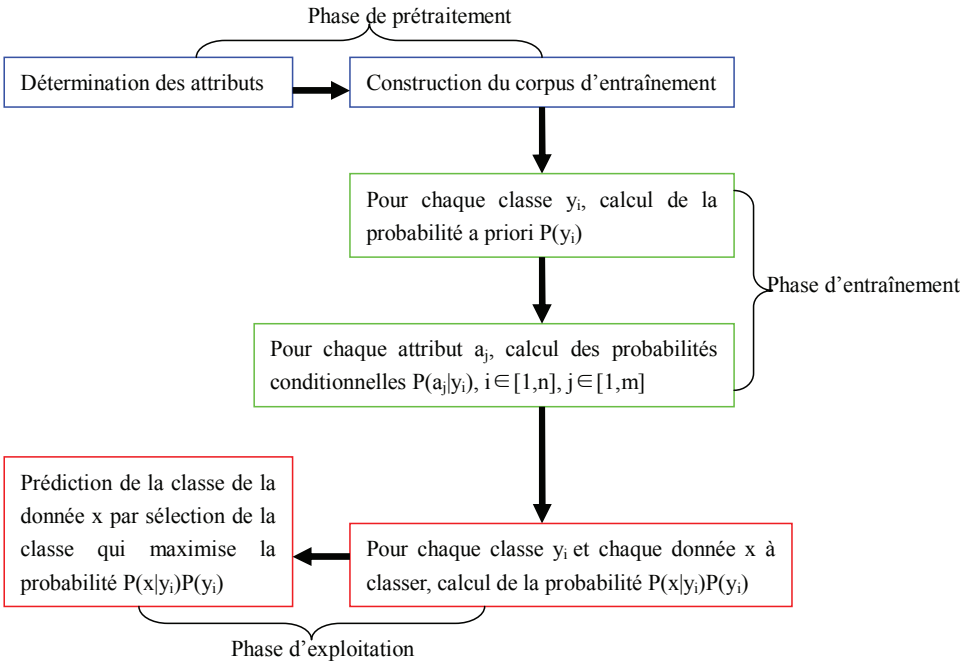

FIGURE 2.6 – Les phases du classifieur de Bayes Naïf

Par ailleurs, il existe des classifieurs bayésiens non Naïfs qui remettent en question l'hypothèse d'indépendance des attributs. Par contre, si les attributs sont dépendants, la conception de ce type de classifieur devient difficile et coûteuse [70]. D'autre part, certaines recherches ont montré que la relaxation de l'hypothèse d'indépendance des caractéristiques n'apporte que très peu d'amélioration comparativement au classifieur bayésien naïf [146] et [33].

2.3.2.2 Séparateurs à vaste marge

Les séparateurs à vaste marge ("Support Vector Machines" ou SVM) [147] relèvent d'une famille de classifieurs qui constitue l'état de l'art en matière de classification supervisée. Ils sont adaptés à la plupart des tâches de classification, notamment lorsque la taille de l'ensemble d'entraînement est petite, linéaire ou non linéaire et potentiellement en haute dimension. Dans la version initiale, l'objectif est de trouver une fonction f linéaire susceptible de séparer "au mieux" les données en deux classes différentes.

Par exemple, dans la Figure 2.7, est présenté un problème de séparation linéaire à deux classes C_1 et C_2. On cherche une fonction y = f(x) (appelée un hyperplan) pour les séparer. Cependant, il existe beaucoup d'hyperplans solutions, mais lequel est le meilleur ? Les SVM cherchent l'hyperplan qui maximise la marge séparant les deux classes, c'est-à-dire le "no man's land" entre les éléments des deux classes. Par exemple, dans la Figure 2.8, les points "ronds rouges" sur l'hyperplan H1 et les points "carrés bleus" sur l'hyperplan H2 sont les

vecteurs de support (”Support Vector”), qui définissent les frontières d'un hyperplan séparateur limite. Nous constatons que l'hyperplan H maximise la marge entre les vecteurs de support ”ronds rouges” et les vecteurs de support ”carrés bleus”.

Une propriété intéressante des SVM est que le meilleur hyperplan est déterminé uniquement par les vecteurs de support, ce qui diminue considérablement le calcul car seuls les vecteurs de supports sont exploités au sein des fonctions discriminantes, les autres données n'étant plus utilisées une fois l'apprentissage effectué. C'est une différence majeure par rapport à des algorithmes tels que k-PPV pour lesquels toutes les données d'apprentissage sont utilisées lors de la phase d'exploitation [156].

L'autre avantage du SVM est qu'il s'adapte facilement aux problèmes non linéairement séparables. Avant de procéder à la recherche de la meilleure séparation linéaire, les vecteurs d'entrée sont projetés dans un espace de dimension plus élevée. De cette façon, un séparateur linéaire trouvé par un SVM dans ce nouvel espace vectoriel devient un séparateur non linéaire dans l'espace original. Cette transformation des vecteurs se fait à l'aide d'une fonction noyau (”kernel”). Une fonction noyau doit respecter les conditions du théorème de Mercer [88]. Elle permet de transformer un produit scalaire dans un espace de grande dimension. Il existe plusieurs types de noyaux comme le noyau polynomial, le noyau gaussien, etc.

Selon [60], les SVM conviennent bien pour la classification de textes parce que, premièrement, la dimension élevée de l'espace de données les influence peu car ils sont relativement protégés contre le sur-apprentissage. Un autre aspect positif des SVM est que presque aucun ajustement manuel de paramètres n'est requis. Mis à part le choix du noyau, les méta paramètres peuvent être ajustés par validation croisée sur les données d'apprentissage.

Cependant, le SVM classique présente un inconvénient négligeable : le choix de fonction noyau n'est pas claire et les paramètres dans les fonctions noyaux sont généralement difficiles à interpréter. Par ailleurs, une faiblesse du SVM classique est qu'il sépare seulement en deux classes. Pour pallier cet inconvénient, le SVM multiclasse a été proposé et présenté dans [54] et [13] pour résoudre les problèmes multiclasses. L'idée est de réduire le problème multiclasse en plusieurs problèmes de classification à deux classes.

2.3.2.3 Algorithme des k plus proches voisins

L'algorithme des k plus proches voisins ou k-PPV (”k-nearest neighbors” ou kNN) [26] est une méthode d'apprentissage à base d'instances. Le classifieur k-PPV n'implique pas de phase d'entraînement en tant que telle (”lazy learners”). La seule opération préalable est le stockage des exemples d'entraînement. L'apprentissage est repoussé au moment où un nouveau document à classer arrive. De ce fait, la plus grosse part de l'effort requis en termes de temps de calcul est fournie au moment même de la classification. Lorsqu'un nouveau document à classer arrive, il est comparé aux documents d'entraînement à l'aide d'une mesure de similarité. Ses k plus proches voisins sont alors considérés : leur catégorie est observée et celle qui est majoritaire parmi les voisins est affectée au document à classer. C'est là une version de base de

l'algorithme qu'il est possible de raffiner. Souvent, il est possible de pondérer les voisins par la distance qui les sépare au texte à classer. Nous accordons plus de poids, lors de la prise de décision, aux documents les plus similaires.

La valeur de k est un des paramètres à déterminer lors de l'utilisation de ce type de classifieur. Souvent, une optimisation de ce paramètre est effectuée à l'aide de tests sur une portion des documents. Il est à noter que si nous ne pondérons pas les voisins, la valeur que nous choisissons pour k va être plus critique, plus déterminante par rapport à la performance du classifieur. Par contre, si nous décidons de pondérer les voisins, l'importance du paramètre k va être atténuée. Nous pouvons nous permettre de considérer un plus grand nombre de voisins, sachant que plus ils diffèrent du document à classer, moins ils ont d'impact sur la prise de décision. Cependant, il peut s'avérer nécessaire de limiter le nombre de voisins pour s'en tenir à un temps de calcul raisonnable.

Une des caractéristiques fondamentales de ce type de classifieur est l'utilisation d'une mesure de similarité entre les documents (par exemple la similarité cosinus). Les textes étant représentés sous forme vectorielle, donc comme des points dans un espace à n dimensions, nous pouvons au premier abord penser à déterminer les voisins les plus proches en calculant la distance euclidienne entre ces points.

Selon les différentes façons de prendre en compte le nombre de documents pris comme données d'entraînement, deux méthodes sont souvent utilisées dans la phase d'entraînement pour estimer les taux d'erreur de classification : "leave one out validation" et "K-fold cross validation".

1. L'idée de la "leave one out validation" est que pour chaque exemple d'entraînement, on apprend un classificateur avec toutes les données d'entraînement disponibles, sauf cet exemple, ensuite on teste la précision du classificateur sur cet exemple. La moyenne prise sur l'ensemble des précisions de chaque exemple donne l'espérance de la précision du classifieur.

2. L'idée de la "K-fold cross validation" est de diviser de façon aléatoire les données d'entraînement en K échantillons. Pour chaque échantillon, on apprend un classificateur avec toutes les données d'entraînement autres que celles de l'échantillon considéré et on teste la précision du classificateur sur l'échantillon. C'est la moyenne des K-précisions obtenues qui fournit l'espérance de la précision. On peut également produire un écart type sur cette précision. On choisit souvent K=10. La "leave one out validation' est donc un cas particulier de la "K-fold cross validation", pour laquelle $K = N$ où N est le nombre de documents considérés.

Une des caractéristiques de l'apprentissage à base d'instances est qu'il n'y a pas de construction d'une description explicite de la fonction à apprendre (dans notre cas, l'appartenance à une catégorie). L'avantage est que nous n'estimons pas qu'une seule fois la fonction pour tout l'espace, mais nous l'estimons plutôt localement et différemment pour chaque nouvelle instance. L'inconvénient est que bien que le coût d'entraînement du classifieur soit faible, car il ne réa-

lise que la mémorisation des exemples d'entraînement, le coût de classification de nouvelles instances peut être élevé, puisque c'est à ce moment que tout le calcul est effectué. Cependant, une bonne indexation des exemples aide beaucoup à pallier ce problème [90].

Parmi les différents types de classifieurs connus, nous rappelons que les séparateurs à vaste marge et l'algorithme des k plus proches voisins sont les deux meilleurs choix en matière de classification de textes selon [60], [119], [156], [120]. Par contre, k-PPV fonctionne mieux lorsque le corpus d'entraînement est grand (par exemple 300 documents d'entraînement par classe [120]), alors que SVM est plus adapté lorsque le corpus d'entraînement est relativement petit. Les deux méthodes fournissent en général des résultats meilleurs que les autres classifieurs (par exemple, le classifieur bayésien Naïf).

Par ailleurs, il existe d'autres types de classifieurs comme les arbres de décision [91], les réseaux de neurones [27], [99], la combinaison de classificateurs [71], [118], etc. Cependant, ceux-ci s'avèrent également moins performants que les SVM ou les k-PPV [120].

Dans nos expérimentations, nous avons choisi k-PPV comme algorithme de classification supervisé puisqu'il se comporte bien en matière de catégorisation. De plus, il est très facile à utiliser pour nos expérimentations et les résultats obtenus sont très interprétables.

2.3.2.4 Classification multilingue

Il existe en principal une série de méthodes spécifiquement développées pour la classification de données multilingues. Leur idée principale est commune : la tâche de classification bilingue est considérée d'un point de vue probabiliste. Etant donné un échantillon bilingue (s, t) composé d'une paire de phrases et d'un ensemble de classes C_1, C_2, C_n, la paire (s, t) sera affectée à classe C_i si celle-ci maximise un critère de probabilité a posteriori. Cette probabilité conditionnelle de classes peut être modélisée de différentes manières.

1. Algorithme de transducteur "inférence et décodage"

   Dans [105] et [66], les auteurs ont proposé des modèles de traduction dénommés "transducteurs à états finis stochastiques (SFSTs)". Les SFSTs sont des modèles de traduction qui peuvent être entraînés automatiquement à partir des paires d'échantillons bilingues. Ce sont des réseaux d'états finis qui acceptent les phrases d'une langue en entrée et produisent des phrases d'une langue en sortie. Chaque arc du réseau est associé au symbole d'entrée, une chaîne de caractères et une probabilité de transition associée à la paire d'entrées-sorties. Chaque fois qu'un symbole d'entrée est accepté, la chaîne de caractères correspondante est produite en sortie et un nouvel état est atteint. Une fois que la phrase d'entrée a été analysée complètement, une sortie finale peut être produite à partir du dernier état atteint.

2. Algorithme bayésien naïf de données bilingues

   L'idée principale de la classification bayésienne naïve de données bilingues [23], [25] est de considérer que les variables aléatoires s et t sont représentées par des vecteurs de caractéristiques indépendantes.

Ces algorithmes sont tous probabilistes et selon [24], l'agorithme bayésien naïf de données bilingues fonctionne mieux en général que l'algorithme de transducteur "inférence et décodage".

Par ailleurs, il existe assez peu d'autre type de méthode de classification multilingue. Un travail, qui effectué par [5], propose une mesure qui minimise les erreurs de classification monolingue dans chacune des langues tout en assurant la cohérence de la classification multilingue (en minimisant la divergence entre les classifieurs monolingues).

2.4 Conclusion

Dans ce chapitre, nous avons fait un tour rapide de l'état de l'art sur les corpus, tout particulièrement les corpus bilingues parallèles et comparables en présentant leurs applications principales. Nous avons étudié l'évolution des définitions des corpus comparables, les différentes approches pour la constitution des corpus comparables et les mesures de comparabilité quantitatives existantes. Nous avons également présenté les principales méthodes de clustering et de classification a priori exploitables dans une optique de construction des corpus thématiques comparables.

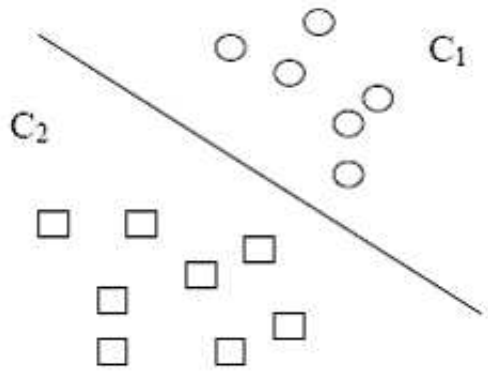

FIGURE 2.7 – Hyperplan pour diviser les deux classes

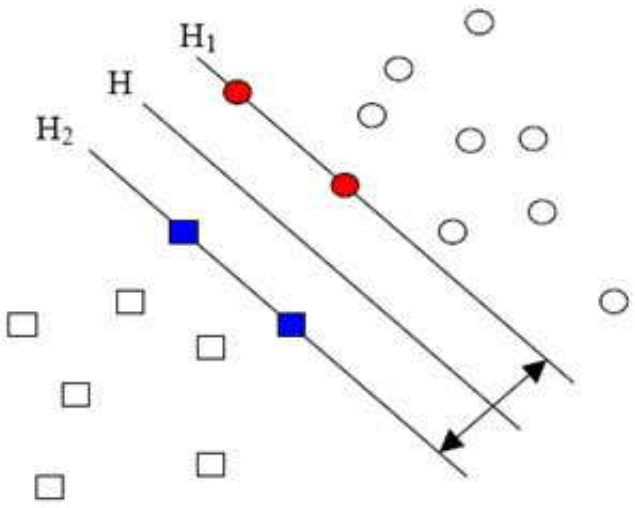

FIGURE 2.8 – Maximisation de la marge

Troisième partie

Contribution à l'élaboration de mesures de comparabilité quantitatives et à leur évaluation

Dans cette partie, nous abordons la construction de mesures de comparabilité quantitatives et leur évaluation empirique par dégradation progressive d'un corpus parallèle.

A partir de la mesure de Li et Gaussier, nous développons deux variantes "enrichies" par l'introduction de grandeurs d'ordre fréquentiel. Nous détaillons ensuite l'environnement d'évaluation dédié que nous exploitons pour étudier le comportement de ces mesures de comparabilité et tenter de les hiérachiser en fonction du contexte d'utilisation.

3

Mesures de comparabilité

Sommaire

3.1 Introduction

La notion de comparabilité entre documents est assez délicate à introduire : il est communément admis que deux documents de langues différentes sont comparables lorsque ces documents traitent de sujets analogues. Par extension, la notion de corpus comparable introduite par [39], [96] reste assez subjective : un corpus comparable est un corpus qui couvre un même thème ou un thème similaire et contient des informations qui se "recouvrent". Dans [30], les auteurs 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 $\mathcal{L}_1$*. [77] en ont dérivé une mesure qui s'appuie sur un dictionnaire de traduction parallèle. Ces auteurs ont d'autre part proposé d'évaluer

cette mesure en partant de documents parallèles (c'est-à-dire de paires de traductions) puis de dégrader progressivement ces paires de traductions en observant la variation produite sur la mesure de comparabilité proposée, 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 d'occurrences des traductions des entrées lexicales qui dépend d'une manière non explicite à la fois du dictionnaire de traduction et de la composition des corpus étudiés.

Dans ce chapitre nous proposons d'étudier et de comparer deux variantes autour de cette mesure de comparabilité en introduisant des informations quantitatives supplémentaires, celles-ci concernent 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 configurations expérimentales, en particulier, lors de tâches de classification ou de clustering de documents comparables thématiques. Ces nouvelles mesures sont présentées puis évaluées par rapport à la mesure développée par [77], en prenant en considération la couverture du dictionnaire de traduction exploité.

3.2 Variations autour d'une mesure quantitative de comparabilité

3.2.1 Mesure de comparabilité de Li et Gaussier (C_{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 comparabilité définie par [77] se présente formellement sous la forme :

$$C_{LG}(C_1,C_2) = \frac{\sum\limits_{w_1 \in WC_1 \cap WD_1} \sigma(w_1) + \sum\limits_{w_2 \in WC_2 \cap WD_2} \sigma(w_2)}{|WC_1 \cap WD_1| + |WC_2 \cap WD_2|} \tag{3.1}$$

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 lexicales 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.

Cette mesure de comparabilité est dite "traductionnelle" car elle est relativement bien adaptée à une tâche d'aide à la traduction.

3.2.2 Vers une définition quantitative de la comparabilité thématique

Comme précédemment mentionné, la mesure C_{LG} ne prend ni en compte le nombre d'occurrences des entrées lexicales dans les documents ni leurs nombres de traductions. Cependant, d'après les travaux effectués par [111], un thème est caractérisé par des mots clés fréquents

intra-thème et discriminants inter-thèmes. Cela nous amène à considérer, pour la construction d'une mesure quantitative de comparabilité "thématique", les fréquences d'occurrence des termes dans les documents et leur ambiguïté (estimée via le nombre de traductions possibles existantes dans le dictionnaire de traduction).

Nous proposons donc ci-après deux variantes de la mesure C_{LG} qui font intervenir explicitement ces deux grandeurs en conjecturant que leur prise en compte produira dans certaines situations expérimentales un effet positif. Nous développons certaines de ces situations dans le chapitre qui décrit notre deuxième contribution consacrée à des questions de classification et de clustering de documents bilingues thématiques.

3.2.2.1 Deux nouvelles mesures de comparabilité, variantes de la mesure C_{LG}

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.

Soit $A_{1|2}, A_1, A_{2|1}, A_2$ définis comme suit :

$$
\begin{aligned}
A_{1|2} &= \sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{W(w_1, C_1)}{\tau(w_1, WD_1)} \cdot \sigma(w_1) \right) \\
A_1 &= \sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{W(w_1, C_1)}{\tau(w_1, WD_1)} \right) \\
A_{2|1} &= \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{W(w_2, C_2)}{\tau(w_2, WD_2)} \cdot \sigma(w_2) \right) \\
A_2 &= \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{W(w_2, C_2)}{\tau(w_2, WD_2)} \right)
\end{aligned}
$$

où $W(w_i, C_i)$ est soit une pondération *tf* soit une pondération *tf-idf* ; $\tau(w_i, WD_i)$ est le nombre de traductions de l'entrée lexicale w_i du corpus C_i dans le dictionnaire de traduction WD_i. $\sigma(w_i)$ est défini comme précédemment ($\sigma(w_i) = 1$ 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.).

1. La première variante, C_{VA_1}, s'explique de la manière suivante :

$$
C_{VA_1} = \frac{1}{2} \cdot \left(\frac{A_{1|2}}{A_1} + \frac{A_{2|1}}{A_2} \right) \tag{3.2}
$$

2. La deuxième variante, C_{VA_2}, s'explique de la manière suivante :

$$C_{VA_2} = \frac{A_{1|2} + A_{2|1}}{A_1 + A_2} \tag{3.3}$$

Ces deux variantes sont très proches l'une de l'autre. Elle se distinguent essentiellement sur la manière de symétriser la mesure. Fondamentalement, la première variante s'apparente à une moyenne arithmétique tandis que la seconde variante se rapporte à une moyenne pondérée.

3.3 Protocole d'évaluation des mesures quantitatives de comparabilité

Nos expérimentations se sont focalisées sur les langues Anglaise et Française et suivent globalement le protocole proposé dans [77]. Ce protocole est construit sur le principe d'une dégradation progressive d'un corpus parallèle par remplacement déterministe des lignes par blocs de texte. Nous avons complété ce protocole en développant une approche non-déterministe pour le remplacement des blocs de texte afin d'évaluer l'impact de la procédure de remplacement des blocs de texte sur la qualité estimée des mesures.

3.3.1 Mesure d'évaluation et paramètres d'étude

3.3.1.1 Référence empirique étalon

La référence empirique est construite sur la base du pourcentage de dégradation du corpus parallèle Europarl. La dégradation progressive de ce corpus est abordée par remplacement de lignes au sein de blocs de texte de taille fixe (en nombre de lignes). Par exemple, si nous considérons 100 lignes par bloc de texte, pour chaque bloc de texte et pour chaque test, nous obtenons un vecteur de 101 valeurs (en partant de 0% de lignes remplacées pour aboutir à 100% de lignes remplacées). Les 101 valeurs correspondent à des pourcentages de dégradation du corpus parallèle. Nous obtenons ainsi une mesure de référence empirique, dite étalon, caractérisée par un vecteur (0%, 1%, 2%...100%) de $N = 101$ coordonnées.

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

Pour établir le degré d'adéquation ou d'inadéquation d'une mesure de comparabilité vis-à-vis de la référence empirique, nous utilisons le coefficient de corrélation de Pearson [104]. Celui-ci estime le degré de corrélation entre deux variables aléatoires x et y (ici, une mesure de comparabilité X et la référence empirique Y) de la manière suivante :

$$r_p = \frac{\sum_{n=1}^{N} (X_n - \bar{X}) \cdot (Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^{N} (X_n - \bar{X})^2} \sqrt{\sum_{n=1}^{N} (Y_n - \bar{Y})^2}} \tag{3.4}$$

Parmi d'autres estimateurs de corrélation, en particulier le coefficient de corrélation de Spearman [128] ou le coefficient de corrélation de Kendall [65], le coefficient de corrélation de Pearson est l'estimateur le plus classique, très utilisé en général 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.

3.3.1.3 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 de texte) par la quantité $TC_D = \frac{|V \cap D|}{|V|}$.

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

3.3.2 Prétraitements et principes d'évaluation

3.3.2.1 Dictionnaires bilingues utilisés

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

Le premier dictionnaire référencé sous l'intitulé *fullDicText* est un dictionnaire propriétaire utilisé dans les travaux de Li et Gaussier [77] 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 chez ELRA [1] 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.

3.3.2.2 Prétraitements

Nous disposons de deux corpus : un corpus parallèle français-anglais Europarl [2] corpus [67] et un corpus anglais TREC [3] Associated Press corpus : AP. Les mots non vides (voir les listes de mots-vides anglais et français en Annexe A) de ces corpus sont lemmatisés en exploitant le TreeTagger [115] [116] puis segmentés en phrases (une phrase par ligne). Les termes sont pondérés en fontion des deux modèles de pondération exploités : *tf* et *tf-idf*. A l'issue de ce prétraitement, nous disposons ainsi de trois documents contenant chacun plusieurs millions de

1. http ://catalogue.elra.info/
2. http ://www.statmt.org/europarl/
3. http ://trec.nist.gov/

lignes : un document parallèle français (EPF), un document parallèle anglais (EPE) (tous deux issus du corpus parallèle Europarl) et un document anglais Associated Press (AP).

3.3.2.3 Principes d'évaluation

En suivant les travaux de Li et Gaussier [77], nous partitionnons le corpus parallèle Europarl en sélectionnant un nombre variable de lignes pour chaque élément de partition : 1000 lignes, 10000 lignes, 100000 lignes et 1428000 lignes (142800 lignes correspond à l'intégralité du corpus Europarl). Chaque élément de la partition obtenue est ensuite divisée en 10 blocs de texte, chaque bloc de texte contenant le même nombre de lignes (100 lignes, 1000 lignes, 10000 lignes, et 142800 lignes). Nous évaluons ensuite les mesures de comparabilité au niveau des blocs de texte alignés.

Nous proposons deux séries d'expériences 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, et précisés en Figure 3.1. Nous faisons les remplacements seulement sur le côté anglais du corpus Europarl mais ne touchons à rien sur le côté français. L'évaluation des mesures de comparabilité consiste à quantifier la corrélation entre leur décroissance observée et la décroissance attendue au sens de la mesure *empirique* lorsque le degré de dégradation du corpus parallèle initial augmente.

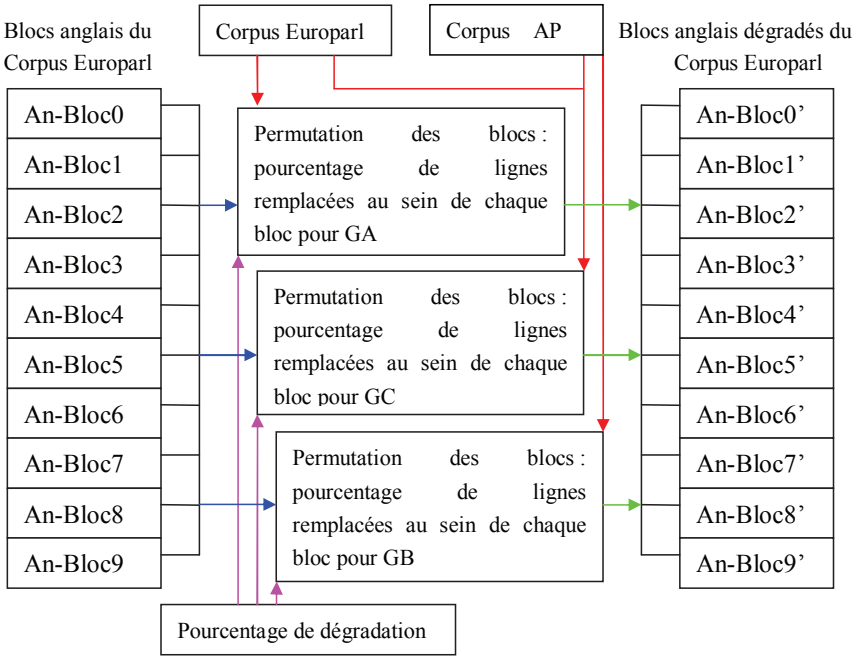

FIGURE 3.1 – Dégradation partitionnée et progressive du corpus Europarl pour les deux modes de remplacement (déterministe ou aléatoire).

3.3.2.4 Remplacement déterministe

Pour le premier test, nous construisons les corpus référencés par GA_d en remplaçant par permutation un certain nombre de lignes issues d'un bloc de texte (le nombre de lignes est fonction du pourcentage de dégradation du corpus parallèle $0\%, 1\% \cdots 100\%$) par le même nombre de lignes issues d'un autre bloc de texte. Toutes les permutations sont donc effectuées au sein du corpus Europal. La permutation des blocs de texte est prédéfinie, par exemple : le bloc de texte 1 est remplacé par le bloc de texte 6, le bloc de texte 2 est remplacé par le bloc de texte 7, le bloc de texte 3 est remplacé par le bloc de texte 8, etc. Comme le nombre de lignes de chaque bloc de texte est le même et si le nombre de lignes à remplacer est 100, nous remplaçons 100 lignes du bloc de texte 1 par les premières 100 lignes du bloc de texte 6, etc.

Pour le deuxième test, nous construisons les corpus référencés par GB_d, en remplaçant certaines lignes issues d'un bloc de texte (le nombre de lignes est fonction du pourcentage de dégradation du corpus parallèle souhaité) par le même nombre de lignes extraites du document *AP* par les lignes numérotées entre les deux valeurs suivantes : (*le numéro du bloc de texte * le nombre de lignes à remplacer*) et (*(le numéro du bloc de texte + 1) * le nombre de lignes à remplacer - 1*).

Pour le troisième test, nous construisons les corpus référencés par GC_d, en remplaçant toutes les lignes d'un bloc de texte par toutes les lignes d'un autre bloc de texte, c'est-à-dire par exemple, le bloc de texte 1 devient le bloc de texte 6 et le bloc de texte 2 devient le bloc de texte 7, etc. A ce stade, et dans chaque bloc de texte, 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* de la même manière que le deuxième test.

3.3.2.5 Remplacement aléatoire

Pour le premier test, nous construisons les corpus référencés par GA_a en remplaçant aléatoirement au sein de chaque bloc de texte 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 Europal.

Pour le deuxième test, nous construisons les corpus référencés par GB_a en remplaçant aléatoirement au sein de chaque bloc de texte 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* (tirage sans remise).

Pour le troisième test, nous construisons le corpus référencé par GC_a, en remplaçant au sein de chaque bloc de texte d'abord toutes les lignes d'un bloc de texte par le même nombre de lignes issues du complément du bloc de texte dans l'ensemble des lignes du corpus Europarl (sans remplacement). Ensuite, au sein de chaque bloc de texte, 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 $GB_{d|a}$.

3.4 Evaluations des mesures de comparabilité sur la base de série de corpus dégradés décrits précédement

Les mesures de comparabilité C_{LG}, C_{VA_1}, C_{VA_2} sont évaluées et comparées au sens de leur corrélation avec la mesure empirique attendue, en faisant varier les paramètres taille des blocs et le taux de couverture.

3.4.1 Influence de la taille des blocs de texte 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 de texte exprimée en nombre de lignes varie dans l'ensemble $\{10^2, 10^3, 10^4, 10^5\}$.

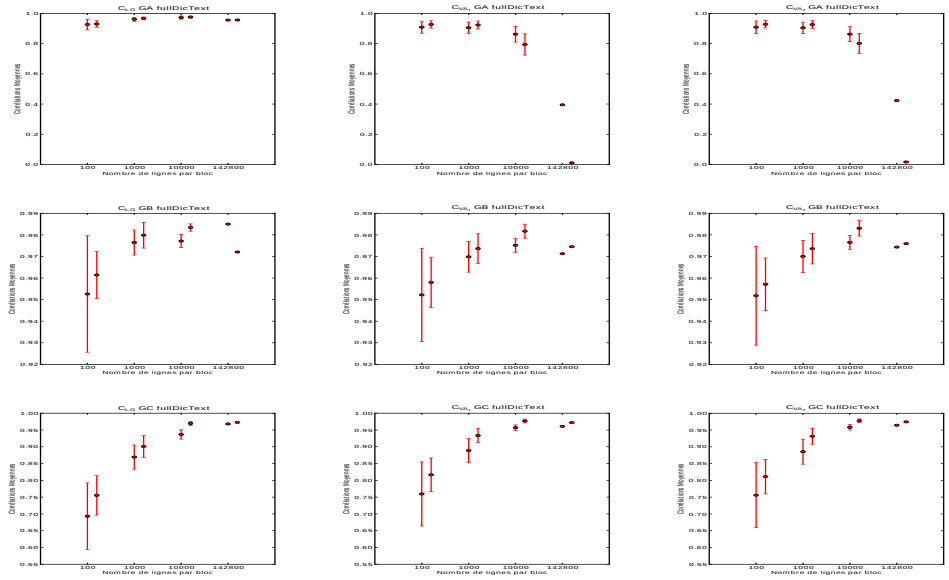

FIGURE 3.2 – Influence de la taille des blocs de texte de corpus sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire bilingue *fullDicText*. Les deux modes de remplacement sont représentés pour chaque taille de bloc de texte avec un léger décalage : déterministe à gauche et aléatoire à droite

Les figures 3.2 et 3.3 montrent, pour les deux modes de remplacement, que la mesure C_{LG} est mieux en adéquation avec la référence empirique au sens du coefficient de corrélation de

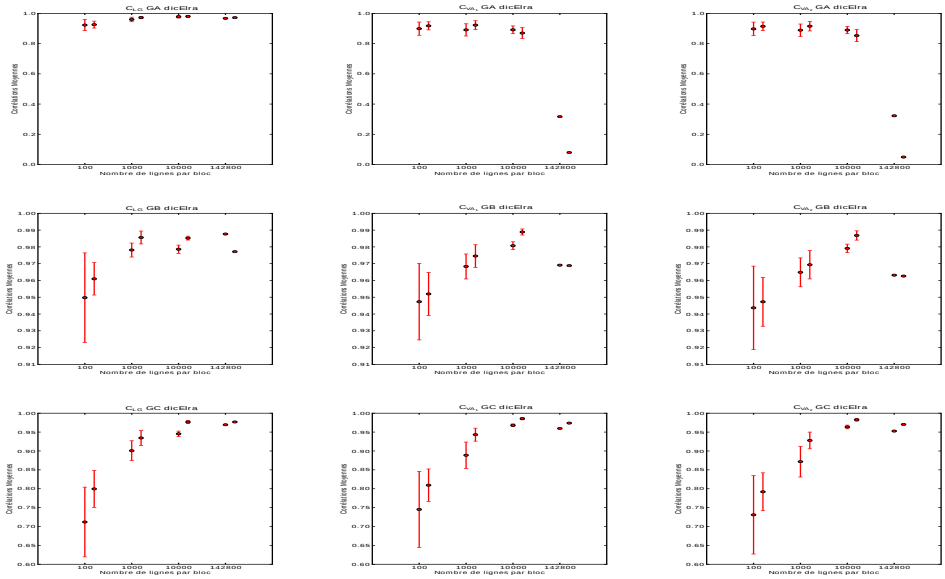

FIGURE 3.3 – Influence de la taille des blocs de texte de corpus sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire bilingue *dicElra*. Les deux modes de remplacement sont représentés pour chaque taille de bloc de texte avec un léger décalage : déterministe à gauche et aléatoire à droite

Pearson sur les expériences *GA* que ses variantes C_{VA_1} et C_{VA_2}, en particulier pour des tailles de blocs de texte 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 C_{VA_1} et C_{VA_2} semblent être légèrement plus robustes que C_{LG}, principalement pour des tailles de bloc de texte petites. Les deux dictionnaires bilingues utilisés conduisent à des résultats très voisins. Par contre, la procédure de remplacement aléatoire semble améliorer significativement, pour toutes les mesures et pour les deux dictionnaires, la corrélation avec la référence empirique étalon, tant en moyenne qu'en écart type.

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

Nous étudions ici l'influence des taux de couverture TC_D et TC_V (des dictionnaires et des vocabulaires respectivement en faisant varier la taille des blocs de texte) 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 C_{LG}, C_{VA_1} et C_{VA_2}. Les figures 3.4 et 3.5 présentent ces corrélations moyennes pour les deux dictionnaires *fullDicText* et *dicElra* et pour les deux modes de remplacement, aléatoire et déterministe.

Nous constatons sur les figures 3.4 et 3.5 une meilleure corrélation moyenne pour la mesure C_{LG} sur les corpus *GA*, tandis que les variantes C_{VA_1} et C_{VA_2} voient leurs corrélations s'effondrer

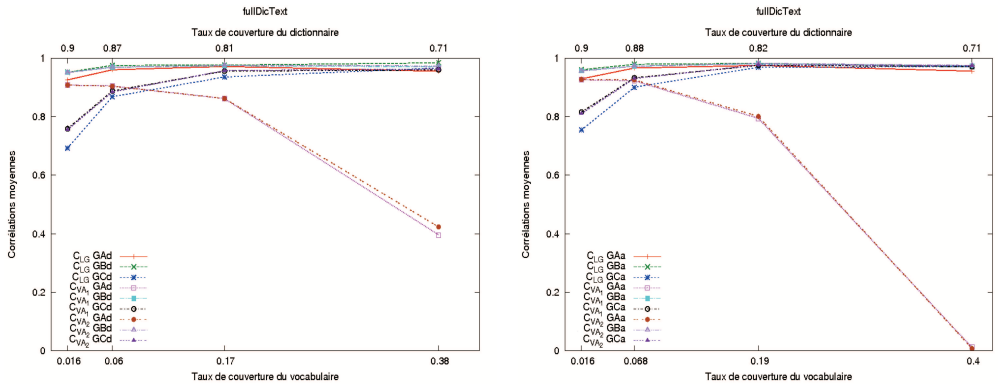

FIGURE 3.4 – Influence du taux de couverture TC_V sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire *fullDicText*, à gauche pour les corpus dégradés par remplacement déterministe, à droite pour les corpus dégradés par remplacement aléatoire

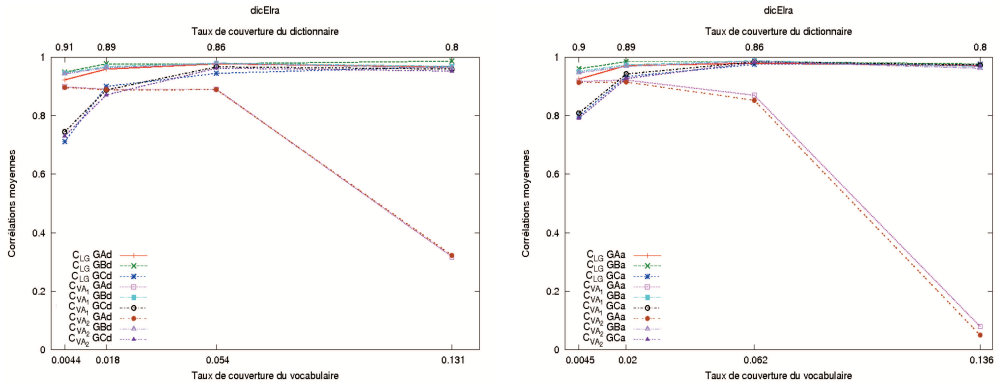

FIGURE 3.5 – Influence du taux de couverture TC_V sur les corrélations moyennes des mesures vis-à-vis de la référence empirique étalon pour le dictionnaire *dicElra*, à gauche pour les corpus dégradés par remplacement déterministe, à droite pour les corpus dégradés par remplacement aléatoire

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 empirique, comparativement à la mesure C_{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 *dicElra* ainsi que pour les deux modes de remplacement déterministe et aléatoire.

3.4.3 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 :

$$\Delta(i) = \frac{|\sigma_i + \sigma_{i+1} + 2 \cdot (m_i - \sigma_i/2 - (m_{i+1} + \sigma_{i+1}/2))|}{\sigma_i + \sigma_{i+1}} = \frac{2 \cdot |m_i - m_{i+1}|}{\sigma_i + \sigma_{i+1}} \tag{3.5}$$

où m_i et σ_i sont les moyennes et écarts types des valeurs de comparabilité associées aux niveaux (0%, 1%, $\cdots$100%) de dégradation du corpus Europarl indexés par $i \in \{1, \cdots, 101\}$. 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.

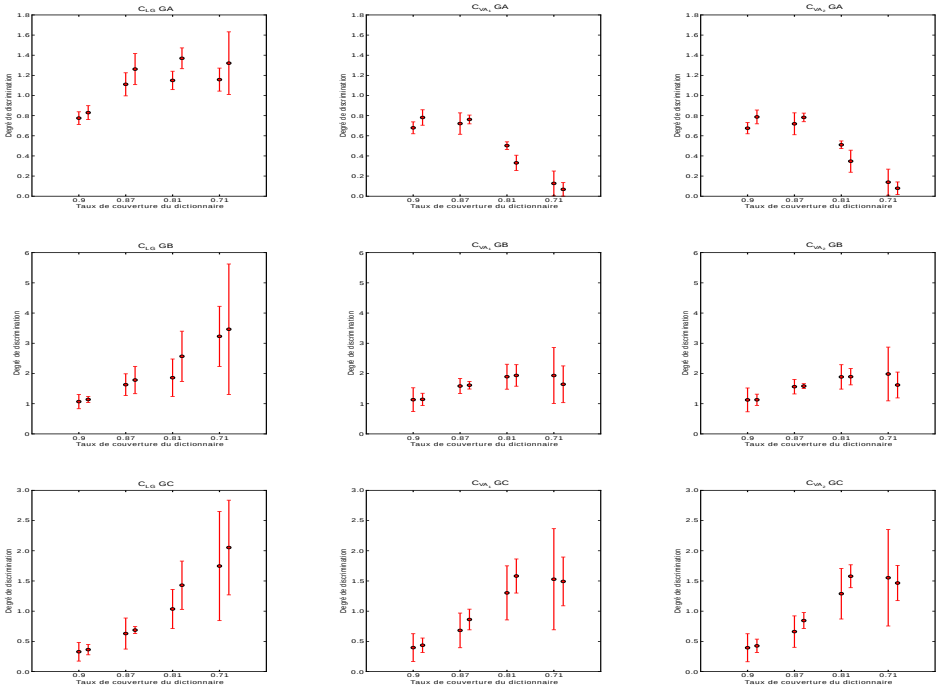

FIGURE 3.6 – Capacité des mesures de comparabilité à discriminer les degrés de dégradation du corpus Europarl : moyennes et écarts-types de $\Delta(.)$ en fonction des taux de couverture du dictionnaire TC_D *fullDicText* exploité sur les corpus produits par remplacements déterministe (décalages à gauche) et aléatoire (décalages à droite).

Les Figures 3.6 et 3.7 présentent pour les trois mesures C_{LG}, C_{VA_1} et C_{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

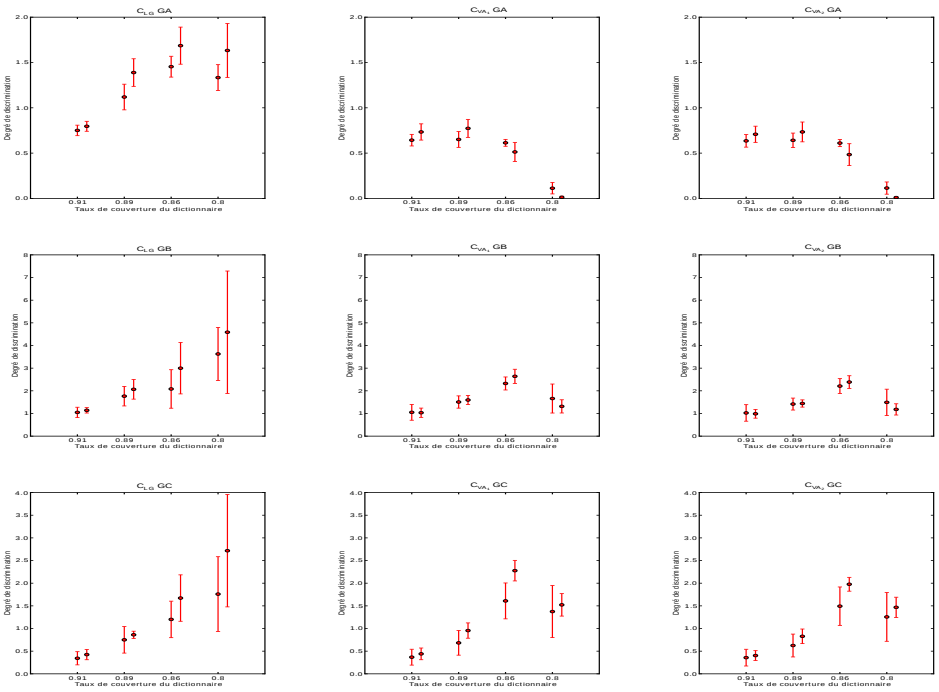

FIGURE 3.7 – Capacité des mesures de comparabilité à discriminer les degrés de dégradation du corpus Europarl : moyennes et écarts-types de $\Delta(.)$ en fonction des taux de couverture du dictionnaire TC_D *dicElra* exploité sur les corpus produits par remplacements déterministe (décalages à gauche) et aléatoire (décalages à droite).

en fonction du taux de couverture des dictionnaire *fullDicText* et *dicElra* respectivement. Ici également, nous constatons que les variantes C_{VA_1} et C_{VA_2} sont moins discriminantes que la mesure C_{LG} sur les corpus *GA* surtout pour les taux de couverture faibles. Sur les corpus *GB*, 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 dans la plupart des cas.

3.5 Conclusion

Les résultats obtenus montrent que la mesure C_{LG} et ses variantes C_{VA_1}, C_{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 proposé. Il ressort néanmoins que la mesure C_{LG} est bien mieux corrélée à la mesure étalon sur les corpus les plus proches du corpus

parallèle initial (Europarl) GA_d et GA_a, tandis que les variantes C_{VA_1} et C_{VA_2} sont légèrement plus robustes lorsque les mesures sont confrontées aux corpus GC_d et GC_a, 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 GB_d et GB_a les trois mesures atteignent des niveaux de corrélation similaires 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*, C_{VA_2} est légèrement mieux corrélée à la mesure étalon, tandis que pour le dictionnaire *dicElra*, c'est la variante C_{VA_1} qui semble mieux corrélée.

Les degrés de corrélation de ces mesures augmentent lorsque le nombre de lignes par bloc de texte augmente, en particulier pour le corpus *GC* (augmentation de plus de 20% entre la configuration 100 lignes par bloc de texte et la configuration 142800 lignes par bloc de texte). Par exemple, pour deux documents d'environ 100 lignes chacun, si la valeur de comparabilité est supérieure à $0,7$, les deux documents seront 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 taille (exprimé en nombre de phrases) afin de juger si les documents sont suffisamment comparables ou non pour une tâche considérée, par exemple, la tâche de la construction des corpus comparables.

Par ailleurs, la capacité 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 C_{LG} se comporte mieux sur les corpus *GA* tandis que les variantes C_{VA_1} et C_{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 C_{VA_1} et C_{VA_2}.

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 [77] 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 de remplacement déterministe.

Enfin, les résultats relatifs à l'évaluation de la mesure C_{LG} que nous avons obtenus confirment pleinement les résultats obtenus par Li et Gaussier [77] sur cette même mesure.

Quatrième partie

Contribution à la classification et au clustering de documents bilingues comparables thématiques

Dans cette partie, nous étudions l'impact des trois mesures de comparabilités (la mesure de référence de Li et Gaussier C_{LG} et ses deux variantes C_{VA_1}, C_{VA_2}) sur ce que nous convenons d'appeler la SCF-catégorisation et le SCF-clustering de données bilingues thématiques. L'idée principale repose sur l'exploitation du graphe bipartie liant deux espaces linguistiques associés à la mesure de comparabilité pour induire dans chacun des espaces une similarité dite induite par comparabilité : la fusion des similarités *natives* et *induites*. Nous développons ensuite un modèle de mélange dédié pour fusionner les similarités *natives* avec les similarités *induites* par mesure de comparabilié, pour identifier et aligner les clusters comparables en forte cohérence thématique. Nous proposons en premier lieu une expérimentation assez exhaustive sur un petit corpus collecté à partir de flux RSS, puis appliquons cette même approche à 3 corpus plus significatifs construits à partir de quelques catégories documentaires extraites du site de Wikipédia.

4

Clustering et catégorisation des données bilingues par fusion des similarités *natives* et des similarités *induites* par mesure de comparabilité

Sommaire

4.1 Introduction

Il n'existe aucune méthode directe pour aligner des clusters thématiques de documents comparables partitionnés en deux espaces linguistiques différents. Des travaux connectés à cette problématique existent, comme le biclustering, co-clustering, ou clustering bi-mode introduit dans [89] et [145]. Cependant, ces travaux concernent principalement le traitement dual des vecteurs lignes et colonnes (individus $\times$ variables) d'une matrice donnée. Dans [159], les auteurs ont proposé une méthode de clustering de documents basée sur le partionnement d'un graphe bipartite par la minimisation de la somme des poids des arcs (par exemple, le nombre d'occurrences d'un nœud "terme" dans un nœud "document") entre les paires de nœuds non appareillés normalisées dans ce graphe bipartite. Dans [8], les auteurs ont proposé une méthode basée sur la combinaison de l'analyse canonique des corrélations et de la rotation procustéenne pour aligner des mots et leurs traductions. L'idée principale est de représenter les mots de langues différentes sur un espace de référence commun pour trouver le mot (la traduction) de la langue cible le plus proche avec le mot de la langue source.

Récemment, dans [58], [59], les auteurs ont développé une approche supervisée qui apprend de manière supervisée des représentations inter-langues en s'appuyant sur des ensembles de documents d'apprentissage alignés. Ces auteurs ont exploité des mesures d'association de mots, conjointement à un dictionnaire bilingue pour éliminer les paires de documents alignés erronées. Par ailleurs, dans [79], Li et Gaussier ont proposé une solution pour regrouper les corpus bilingues en utilisant la mesure de comparabilité seule.

Les mesures de comparabilité définies pour les corpus bilingues comparables s'appliquent en effet lorsqu'il s'agit de traiter des documents monolingues partitionnés en deux espaces linguistiques distincts, pour autant qu'un dictionnaire bilingue reliant ces deux espaces linguistiques est disponible. Au niveau du document, nous sommes donc confrontés à une situation où des similarités monolingues *natives* existent en général dans chacun des espaces linguistiques, ces derniers étant potentiellement liés de manière forte par une mesure de comparabilité. Dans le cadre de la construction de corpus comparables thématique, cela conduit à aborder la classification et/ou le clustering de données bilingues en cohérence thématique forte. Nous ciblons en effet l'alignement des clusters fortement comparables de documents qui sont en outre thématiquement cohérent dans chaque espace linguistique, c'est à dire caractérisé par une grande similarité intra-langue.

Une telle situation se présente lors de la collecte de données thématiques multilingues sur le web par exemple. Avec le besoin pressant des ressources comparables, en particulier thématiques, les approches qui exploitent conjointement les similarités *natives* et la comparabilité inter-langue deviennent particulièrement utiles. La motivation de notre approche est de renforcer ou diminuer la similarité entre les documents du même cluster dans un espace linguistique par la comparabilité dans un autre espace linguistique.

La Figure 4.1, reprise de [84] présente deux ensembles de documents *EN* (anglais) et *FR* (français) munis respectivement des fonctions de similarité $S_{en}(.,.)$ et $S_{fr}(.,.)$, dites *natives* et

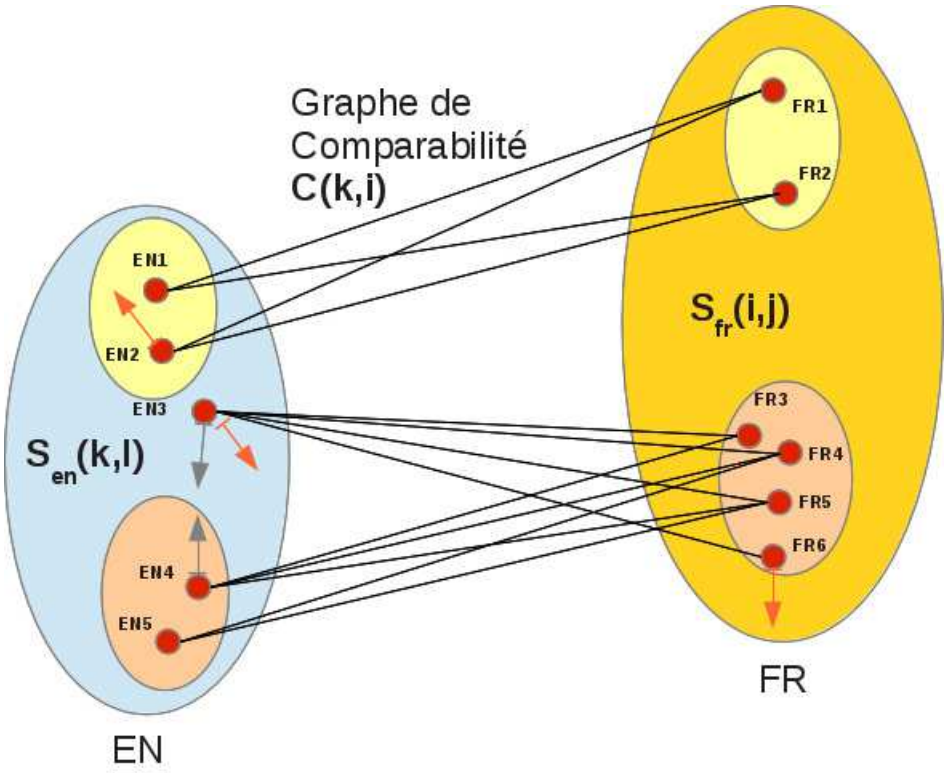

FIGURE 4.1 – Couplage de deux espaces linguistiques par graphe de comparabilité

mis en relation par un graphe de comparabilité défini par la fonction de comparabilité $C(.,.)$. Les arcs de ce graphe sont bidirectionnels et pondérés par une valeur de comparabilité comprise dans l'intervalle $[0,1]$. L'idée principale que nous reprenons ici est celle du renforcement de la similarité par la comparabilité [84] : autrement dit, si deux documents du corpus *EN* sont comparables à un sous-ensemble de documents du corpus *FR* fortement similaires, alors leur similarité devrait être renforcée (et réciproquement). *A contrario*, si deux documents du corpus *EN* sont comparables à un sous-ensemble de documents du corpus *FR* faiblement similaires, alors leur similarité devrait décroître (et réciproquement). Ainsi, sur la figure 4.1, du point de vue de la similarité appréhendée sous l'angle de la comparabilité, le document EN3 devrait s'éloigner du document EN2 pour se rapprocher des documents EN4 et EN5. De même, le document FR6 devrait s'éloigner des documents FR5, FR4 et FR3. L'utilité escomptée d'un tel renforcement/affaiblissement est une forme de filtrage du bruit inhérent aux modèles de représentation des documents caractérisés par un manque de connaissance (linguistique ou terminologique entre autres).

Dans ce chapitre, nous développons, étudions et évaluons l'impact des trois mesures de comparabilités (la mesure de référence de Li et Gaussier et ses deux variantes) sur ce que nous convenons d'appeler la SCF-catégorisation et le SCF-clustering de données bilingues théma-

tiques. A cette fin, nous développons une nouvelle approche dédiée pour combiner la comparabilité et les similarités *natives*, pour identifier et aligner les clusters comparables en forte cohérence thématique. Nous proposons en premier lieu une expérimentation assez exhaustive sur un petit corpus collecté à partir de flux RSS, puis appliquons cette même approche à 3 corpus plus significatifs construits à partir de quelques catégories Wikipédia. Enfin, nous développons les principes à la base d'une généralisation de cette approche dans l'objectif de produire une assistance à la construction à la demande de corpus comparables bilingues thématiquement cohérents.

4.2 Modèle de fusion des similarités *natives* et des similarités *induites* par la comparabilité

Avant de présenter notre approche, nous détaillons la mesure basée sur la comparabilité proposée par Li et Gaussier [79] pour mesurer la similarité entre deux clusters de documents bilingues. Cette mesure est inspirée de la notion d'homogénéité, autrement dit, étant donnés deux clusters C_1 et C_2, la partie anglaise C_1^e de C_1 et la partie française C_1^f de C_1 devraient être comparables à leurs homologues dans C_2 (respectivement de même pour la partie française C_2^f de C_2 et la partie anglaise C_2^e de C_2). Cela conduit à la mesure de similarité suivante s'appliquant sur deux clusters bilingues C_1 et C_2 :

$$sim(C_1,C_2) = \beta \cdot C_{LG}(C_1^e,C_2^f) + (1-\beta) \cdot C_{LG}(C_2^e,C_1^f) \tag{4.1}$$

Où β est la proportion des paires possibles de documents dans le corpus comparable courant (C_1^e,C_2^f) sur toutes les paires possibles de documents. Son équation est définie comme suit :

$$\beta = \frac{\#_d(C_1^e) \cdot \#_d(C_2^f)}{\#_d(C_1^e) \cdot \#_d(C_2^f) + \#_d(C_2^e) \cdot \#_d(C_1^f)}$$

Pour diminuer l'impact du déséquilibre sur les tailles (nombre de documents) des clusters, les auteurs ont proposé un paramètre ϕ pour pénaliser les corpus déséquilibrés. L'équation précédente devient alors :

$$sim_l(C_1,C_2) = sim(C_1,C_2) \cdot \phi(C_1 \cup C_2) \tag{4.2}$$

Où

$$\phi(C) = \frac{1}{1 + log(1 + \frac{|\#_d(C^e) - \#_d(C^f)|}{min(\#_d(C^e),\#_d(C^f))})}$$

, dont $\#_d(C)$ est le nombre de documents dans C.

L'approche que nous proposons est différente dans le sens où elle vise le clustering ou la classification conjointe des données dans deux espaces distincts au sein desquels existent des

mesures de similarité *native* [1]. Notre approche consiste à exploiter la mesure de comparabilité qui associe les deux espaces linguistiques pour fournir une nouvelle mesure de similarité dans chaque espace linguistique, résultat de la fusion des similarités *natives* et des similarités *induites* par la comparabilité. Notre approche ne dépend donc que de l'existence d'un dictionnaire bilingue et d'une mesure de comparabilité définissant un graphe bipartite de documents de langues différentes. Des données pré-alignées exploitables en tant que données d'apprentissage ne sont pas nécessaires.

4.2.1 Mesure de similarité induite par mesure de comparabilité

Dans [84], les auteurs ont proposé un algorithme, $Hit-ComSim$, pour construire itérativement la notion de similarité induite par un graphe bipartite de comparabilité. Malheureusement, cet algorithme ne passe pas bien à l'échelle du fait de sa complexité algorithmique en $O(N^4)$. Nous proposons ici une approche beaucoup plus simple et directe consistant à exploiter directement la matrice de comparabilité construite à partir de deux collections de documents bilingues.

Soient O_1 et O_2 deux collections finies de documents qui appartiennent à deux espaces linguistiques distincts dans lesquels deux mesures de similarité *natives* S_{O_1} et S_{O_2} ont été définies. Soit $C(.,.) : S_{O_1} \times S_{O_2} \rightarrow \mathcal{R}$ la matrice de comparabilité qui caractérise le graphe bi-partite associant les deux collections.

Nous définissons la mesure de similarité *induite* par la matrice de comparabilité C les mesures normalisées (dans $[0,1]$) suivantes respectivement notées $S_{O_1,C}$ et $S_{O_1,C}$:

$\forall (d_i,d_j) \in O_1^2$ et $\forall (d'_i,d'_j) \in O_2^2$

$$\begin{aligned} S_{O_1,C}(d_i,d_j) &= \frac{CC^T(i,j)}{\sqrt{CC^T(i,i)CC^T(j,j)}} \\ S_{O_2,C}(d'_i,d'_j) &= \frac{C^TC(i,j)}{\sqrt{C^TC(i,i)C^TC(j,j)}} \end{aligned} \tag{4.3}$$

L'interprétation des similarités *induites* est immédiate. Tout d'abord, si nous considérons chaque ligne i de la matrice C comme un vecteur de caractéristiques qui caractérise le document $d_i \in O_1$, pour tout $(d_i,d_j) \in O_1$, $CC^T(i,j)$ peut être interprété comme un produit scalaire entre les deux vecteurs de caractéristiques représentant d_i et d_j respectivement. $S_{O_1,C}(d_i,d_j)$ prend ainsi la forme d'une similarité cosinus entre les documents d_i et d_j construite uniquement sur la base de la mesure de comparabilité. De même, si nous considérons chaque colonne i de la matrice C comme un vecteur de caractéristiques qui caractérise le document $d'_i \in O_2$, $S_{O_2,C}(d'_i,d'_j)$ s'apparente à une similarité cosinus entre documents d'_i et $d'_j \in O_2$ construite uniquement sur la base de la mesure de comparabilité.

1. une similarité *native* doit être comprise comme une mesure quantitative de similarité intra-langue, comme la mesure de similarité cosinus associé à une pondération *tf-idf* par exemple

4.2.2 Fusion des similarités *natives* et similarités *induites*

Le modèle de mélange comparabilité/similarités que nous proposons s'exprime sous la forme d'une simple fusion linéaire des similarités *natives* et *induites* définies dans chaque espace linguistique. Formellement, nous utilisons un seul paramètre $\alpha \in [0,1]$ pour combiner linéairement les deux mesures comme suit :

$$\begin{aligned} S'_{O_1}(d_i,d_j) &= \alpha S_{O_1,C}(d_i,d_j) + (1-\alpha) S_{O_1}(d_i,d_j) \\ S'_{O_2}(d'_i,d'_j) &= \alpha S_{O_2,C}(d'_i,d'_j) + (1-\alpha) S_{O_2}(d'_i,d'_j) \end{aligned} \tag{4.4}$$

Puisque les similarités *induites* sont normalisées dans l'intervalle d'unité $[0,1]$, nous préconisons l'utilisation des similarités cosinus [2] (également normalisées dans l'intervalle unité) en tant que similarités *natives* dans les deux espaces linguistiques connectés via le graphe de comparabilité. Ainsi, les similarités mixtes définies par l'équation 4.4 restent consistantes pour toute les valeurs de $\alpha \in [0,1]$.

Cette fusion permet d'aligner deux espaces monolinguistiques (munis des similarités *natives*) par une mesure de comparabilité afin de pouvoir effectuer soit un clustering multilingue soit une classification multilingue. Afin de référencer ces deux types particuliers de clustering et de classification, nous les appellons SCF-clustering et SCF-classification ("SCF" signifie les similarités *natives* (S) et la mesure de comparabilité (C) fusionnées (F)).

4.3 Corpus de test développés et prétraitement des données collectées associé

Nous avons collecté un premier corpus de test, intitulé *RSS7* [3], à partir de 23 flux RSS listés dans la Table 4.1. Ces flux ont été indexés à l'aide de la plateforme Lucene [4] [50]. A partir de cette collection brute, nous avons construit 7 classes de documents extraits à l'aide de mots-clés listés dans la Table 4.2 qui caratérisent également les classes de document prises en compte. Ce corpus est constitué de 252 documents sélectionnés manuellement : il comprend 129 documents en anglais et 123 documents en français. Ces documents constituent sept paires de classes thématiques (une classe anglaise et une française). Chaque classe, contient au minimum une douzaine de documents vérifiés manuellement. La taille des classes est donnée en Tableau 4.2.

Pour tester le passage à l'échelle de notre modèle de mélange de similarités, nous avons collecté également un corpus beaucoup plus important à partir de quelques catégories pré-définies de Wikipédia en exploitant une plate-forme développée dans l'équipe. Nous avons ainsi collecté en totalité 154828 documents se répartissant dans 21 classes, et comprenant 87793 documents en anglais et 67035 documents en français.

2. associées au modèle vectoriel pondéré pour la représentation des contenus des documents
3. le corpus RSS7 est accessible à l'url : http ://people.irisa.fr/Pierre-Francois.Marteau/Corpora/RSS_7classes.zip
4. http ://lucene.apache.org/

Flux RSS	Langue
www.globaltimes.cn/...	EN
www.shanghaidaily.com/...	EN
v1.theglobeandmail.com...	EN
www.thetimes.co.uk/...	EN
rss.nytimes.com/...	EN
feeds.washingtonpost.com/...	EN
feeds.latimes.com/...	EN
www.chinadaily.com.cn/...	EN
feeds.bbci.co.uk/...	EN
www.france24.com/...	EN
rss.cnn.com/rss/...	EN
www.abc.net.au/...	EN
liberation.fr.feedsportal.com/...	FR
www.lavenir.net/rss.aspx...	FR
www.ledevoir.com/rss/...	FR
www.lessentiel.lu/...	FR
rss.feedsportal.com/...	FR
www.romandie.com/rss/flux.xml	FR
rss.lemonde.fr/...	FR
www.courrierinternational.com/...	FR
feeds.lefigaro.fr/...	FR
www.lapresse.ca/...	FR
www.lesoir.be/...	FR

TABLE 4.1 – Liste des flux RSS collectés pour la constitution du corpus RSS7. Tous ces flux sont issus des files d'agence de presse internationale diffusées par les grands quotidiens ou chaînes de télévision en anglais (EN) et en français (FR).

Classe anglaise	#doc	classe française	#doc
Mali	20	Mali	20
Syria	20	Syrie	20
Algeria	12	Algérie	16
Central African Republic	20	République Centre Africaine	20
Gay marriage	20	Mariage gay	20
Pope	20	Pape	20
David Beckham	17	David Beckham	7

TABLE 4.2 – Liste des classes avec leur taille en nombre de documents pour le corpus de test *RSS7*

Comme les complexités pour calculer à la fois la matrice de comparabilié et les matrices de similarité *native* et *induite* ($O(N^2)$ pour les deux premières matrice et $O(N^3)$ pour la dernière du fait des coûts de calcul nécessaire pour évaluer CC^T et C^TC), où N est le nombre de docu-

ments par langue), nous avons choisi de développer trois sous-corpus associés à des difficultés de classification ou de clustering différents. Nous développons ci-dessous les mécanismes de construction de ces trois corpus.

Wikipedia_A est le sous-corpus le plus simple à construire. Pour chaque classe et chaque langue, nous avons simplement choisi aléatoirement une centaine de documents au sein des catégories Wikipédia collectées initialement. Pour les catégories Wikipédia ayant moins de 100 documents nous avons conservé l'intégralité des documents contenus dans ces catégories. Ce corpus est celui qui intègre le plus de variabilité thématique.

Le sous-corpus *Wikipedia_B* est construit de manière à être beaucoup plus cohérent thématiquement que le corpus *Wikipedia_A*. En premier lieu, pour chaque langue et chaque classe, nous construisons la matrice de similarité *native* intra-classe. Dans chaque matrice obtenue, nous comptabilisons ensuite le nombre de valeurs de similarité supérieures à un seuil de similarité relativement élevé (nous avons choisi 0,5) et ordonnons sur la base de ce critère les lignes de la matrice par ordre décroissant. Nous sélectionnons ensuite dans cette matrice ordonnée les cent meilleures lignes qui correspondent aux documents thématiquement cohérents conservés pour la classe et la langue considérées et ne conservons que les lignes qui contiennent au moins 5 valeurs de similarité au dessus du seuil fixé de similarité (ici 0,5). Nous obtenons ainsi le corpus *Wikipedia_B*, lequel contient au plus une centaine de documents par classe et par langue. Ce corpus est celui qui intègre le plus de cohérence thématique.

Le sous-corpus *Wikipedia_C* est un corpus intermédiaire entre les corpus *Wikipedia_A* et *Wikipedia_B*. Il est basé initialement sur le corpus *Wikipedia_B* auquel sont ajoutés aléatoirement dans chaque classe (et pour chaque langue) 50% du nombre des documents contenus dans chaque classe. Ainsi par exemple, si la $Classe_i$ de la $Langue_k$ dans le corpus *Wikipedia_B* contient 100 documents, alors, à l'issue de la phase d'ajout, la $Classe_i$ de la $Langue_k$ dans le corpus *Wikipedia_C* comprendra 150 documents si la catégorie $Classe_i$ contient plus de 150 documents. Ce corpus intègre un niveau de cohérence thématique intermédiaire entre les corpus *Wikipedia_A* (très moyen) et *Wikipedia_B* (très élevé).

Les classes et leurs tailles en nombre de documents pour les trois corpus [5] *Wikipedia_A*, *Wikipedia_B* et *Wikipedia_C* sont données en Tableau 4.3.

Ces corpus sont finalement lemmatisés en utilisant *TreeTagger* [115] [116]. Par ailleurs, les modèles vectoriels avec pondération *tf* et *tf-idf* [127] sont exploités pour représenter les contenus des documents.

5. ces trois corpus sont accessibles à l'url : http ://people.irisa.fr/Pierre-Francois.Marteau/Corpora/Wikipedia_21classes.zip

Classe anglaise	#doc *W_A*	#doc *W_B*	#doc *W_C*	Classe française	#doc *W_A*	#doc *W_B*	#doc *W_C*
Astronomy	100	101	151	Astronomie	100	82	123
Biology	100	101	151	Biologie	100	77	115
Economy	100	96	144	Economie	100	101	151
Food	100	98	147	Nourriture	36	3	4
Football	100	101	151	Football	100	101	151
Genetics	100	55	82	Génétique	100	101	151
Geograpphy	100	93	139	Géographie	100	101	151
Computer	100	101	151	Ordinateur	100	101	151
Literature	100	100	150	Littérature	100	101	151
Mathematics	100	101	151	Mathématique	100	42	63
Medicine	100	101	151	Médecine	100	87	130
Movie	100	101	151	Film	100	101	151
Music	100	101	151	Musique	100	101	151
Skating	100	101	151	Patinage	100	101	151
Heritage	100	101	151	Patrimoine	100	101	151
Politics	100	101	151	Politique	100	101	151
Religion	100	100	150	Religion	100	89	133
Rugby	100	101	151	Rugby	100	101	151
Health	100	101	151	Santé	100	42	63
Sculpture	100	101	151	Sculpture	100	101	151
Tennis	100	101	151	Tennis	100	101	151

TABLE 4.3 – Liste des classes avec leur taille (en nombre de documents) pour les trois corpus : *Wikipedia_A* : *W_A*, *Wikipedia_B* : *W_B*, *Wikipedia_C* : *W_C*

4.3.1 Dictionnaire bilingue

Pour évaluer la comparabilité entre une paire de documents anglais/français, nous utilisons le dictionnaire bilingue précédemment exploité référencé sous la rubrique *dicElra* et disponible chez ELRA sous la référence ELRA-M0033.

4.3.2 Protocole d'évaluation

La performance du classificateur basée sur la règle du plus proche voisin $(1 - PPV)$ [26] est évaluée en utilisant la mesure du taux d'erreur de classification.

La performance des algorithmes de clustering testés sont également évalués en comparant l'étiquette de classe obtenue pour chaque document avec celle connue dans les corpus. Les mesures de précision (*AC*) et d'information mutuelle normalisée (*NMI*) sont utilisées pour évaluer la performance de la tâche de classification [151]. En tant que mesure intrinsèque d'évaluation pour estimer la qualité des tâches de clustering dans chaque espace linguistique, nous utilisons également la mesure de Davies-Bouldin (*DB*) [28] qui évalue le quotient de similarités moyennes intra et inter clusters.

La mesure de précision *AC* (*accuracy*) mesure la fraction des documents qui sont correctement étiquettés, en supposant une correspondance un-vers-un entre les vraies classes et les classes prédites. Soit p une permutation possible de l'ensemble des indices des clusters et des *vraies* classes. La précision est donc définie comme :

$$AC = \frac{1}{N} MAX_p \sum_{i=1\cdots K} n_{i,p(i)} \tag{4.5}$$

Où $n_{i,p(i)}$ désigne le nombre de documents partagés par la classe i supposée connue et le cluster indicé par $p(i)$, K est le nombre de classes et de clusters, et N est le nombre total de documents.

La mesure *NMI* entre un clustering de référence C et un clustering estimé $\tilde{C}$ est définie comme suit :

$$NMI(\tilde{C}, C) = \frac{I(\tilde{C}, C)}{(H(\tilde{C}, \tilde{C}) + H(C, C))/2} \tag{4.6}$$

où

$$I(\tilde{C}, C) = \sum_k \sum_j P(\tilde{c}_k \cap c_j) \log \frac{P(\tilde{c}_k \cap c_j)}{P(\tilde{c}_k) P(c_j)}$$

est l'information mutuelle entre les deux clusterings,
et

$$\begin{aligned} H(\tilde{C}) &= -\sum_k P(\tilde{c}_k) \log P(\tilde{c}_k) \\ H(C) &= -\sum_k P(c_k) \log P(c_k) \end{aligned}$$

sont les entropies associées aux deux clusterings C et $\tilde{C}$.

Enfin, la mesure de Davies-Boulding *DB* est une mesure d'évaluation intrinsèque des données, qui est définie comme suit :

$$DB = \frac{1}{K} \sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{4.7}$$

où K est le nombre de clusters, c_k est le médoide du cluster, σ_k est la distance moyenne des éléments du cluster k au médoide c_k, et $d(c_i, c_j)$ est la distance entre les médoides c_i et c_j. Plus basse est la valeur de *DB*, meilleure est la qualité du clustering puisque cela correspond à des distances intra-cluster faibles (similarités intra-cluster élevées) et des distances inter-clusters élevées (similarités inter-cluster faibles).

Enfin, la qualité de l'alignement des clusters est évaluée par calcul de la matrice de comparabilité moyenne entre les clusters et l'analyse du graphe bipartite des clusters alignés obtenu en faisant varier un seuil d'élagage (pruning) : seuls les liens de comparabilité supérieurs au seuil de comparabilité fixé sont conservés.

4.4 Expérimentations sur le corpus *RSS7*

Le corpus *RSS7*, est un petit corpus, mais très cohérent thématiquement. Il est idéal pour évaluer les effets significatifs liés à la prise en compte conjointe des similarités *natives* et *induites*.

4.4.1 Impact du modèle de mélange des similarités *natives* et des similarités *induites* par mesure de comparabilité sur la classification 1-PPV

Nous étudions tout d'abord l'impact de la fusion des similarités *natives* avec les similarités *induites* sur le taux d'erreur de la classification 1-PPV, en faisant varier le paramètre $\alpha \in [0,1]$ pour les trois mesures de comparabilité C_{LG}, C_{VA_1} et C_{VA_2} avec les pondérations *tf-idf* et *tf*.

La Figure 4.2 montre que la fusion des similarités *natives* avec les similarités *induites* par mesure de comparabilité a un impact significatif pour l'ensemble des trois mesures de comparabilité, en particulier pour les deux variantes C_{VA_1} et C_{VA_2} en abaissant d'environ 5% le taux d'erreur pour la classification des documents en anglais et en français en prenant en compte la pondération *tf-idf* et d'environ 2,5% du taux d'erreur pour la classification des documents en langue anglaise et en langue française en prenant en compte la pondération *tf*. La mesure C_{LG} améliore la précision de la classification, mais le choix d'une valeur correcte pour α est beaucoup plus limité (par exemple, la baisse du taux d'erreur est d'environ 4% pour la classification des documents en anglais pour $\alpha \in [0,7,0,8]$ avec la pondération *tf-idf* et d'environ 1,5% pour la classification des documents en langue anglaise pour $\alpha \in [0,8,0,9]$ avec la pondération *tf*). Cependant, la mesure C_{LG} ne convient pas pour la classification des documents en français, et elle est moins stable que les deux variantes en général. Nous observons également que la pondération *tf* est nettement préférable à la pondération *tf-idf* car elle permet de diminuer de manière plus signficative pour une plage plus large de valeurs pour α.

4.4.2 Evaluation du modèle de mélange des similarités *natives* et des similarités *induites* par la comparabilité sur le clustering k-médoides avec les pondérations *tf-idf* et *tf*

Nous étudions ici l'impact de la fusion des similarités *natives* et des similarités *induites* par mesure de comparabilité sur le clustering k-médoides [62] [63] en considérant les trois mesures de comparabilité étudiées. Nous exploitons ici les trois mesures d'évaluation *AC*, *NMI* et *DB* avec les pondérations *tf-idf* et *tf*.

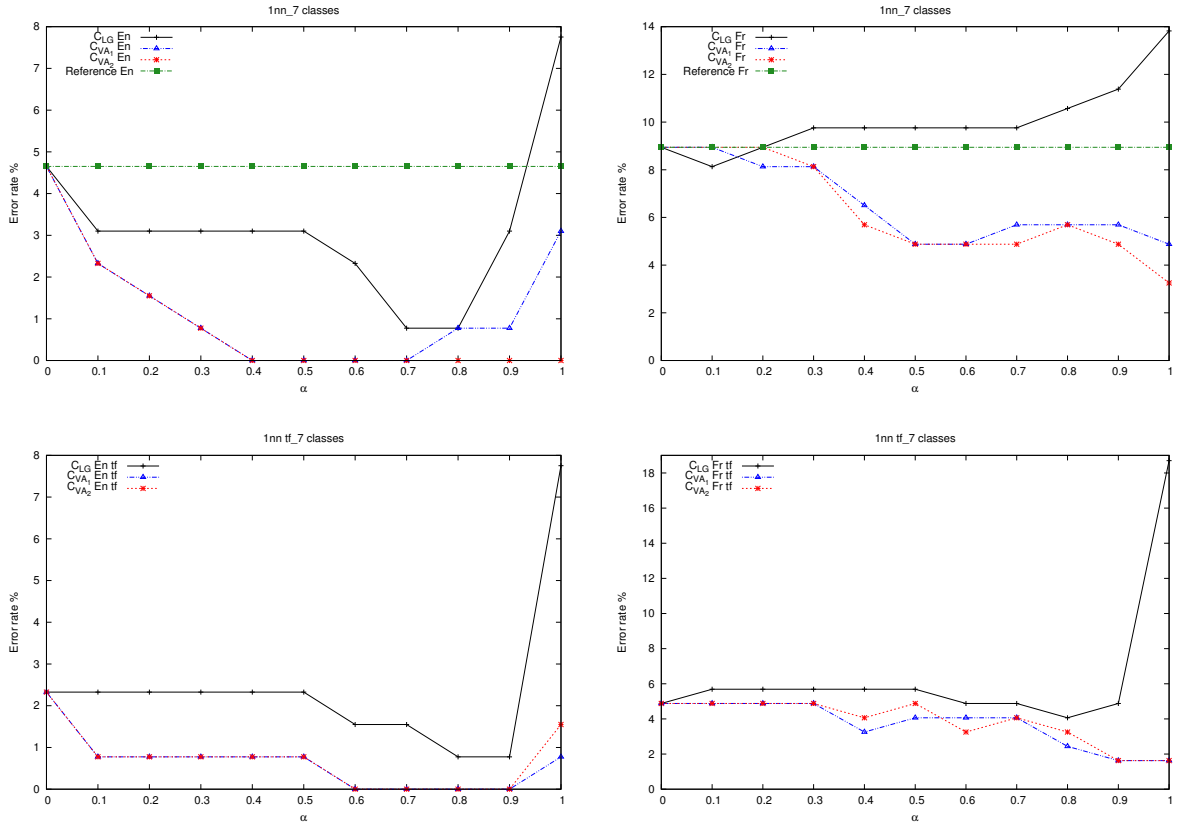

FIGURE 4.2 – Evaluation de l'impact de la fusion des similarités *natives* et des similarités *induites* par mesure comparabilité sur le taux d'erreur d'une classification 1-PPV, pour les trois mesures de comparabilité testées : à gauche, la classification des documents anglais ; à droite, la classification des documents français. Le modèle vectoriel est exploité avec pondération *tf-idf* en haut, et avec pondération *tf* en bas.

Les Figures 4.3 et 4.4 montrent que les deux critères *AC* et *NMI* peuvent être améliorés jusqu'à 15% dans le cadre du clustering des documents en langue anglaise et jusqu'à 30% dans le cadre du clustering des documents en langue française pour les variantes C_{VA_1} and C_{VA_2} avec la pondération *tf-idf*. Les deux critères *AC* et *NMI* peuvent également être améliorés jusqu'à environ 15% sur le clustering des documents en langue anglaise et également sur le clustering des documents en langue française pour C_{VA_1} et C_{VA_2} avec la pondération *tf*. Pour C_{LG} par contre, l'amélioration est limitée à environ 3-5% en sélectionnant la meilleure valeur possible pour α en fonction du choix des pondérations. Ces expérimentations nous orientent également sur le choix d'une bonne valeur de α. Par exemple, un α égal ou un peu plus grand que 0,5 sera un bon choix car autour de 0,5, nous avons obtenu les meilleurs valeurs de *AC* et *NMI*. Nous avons constaté encore une fois que la pondération *tf* est bien meilleure que la pondération *tf-idf*.

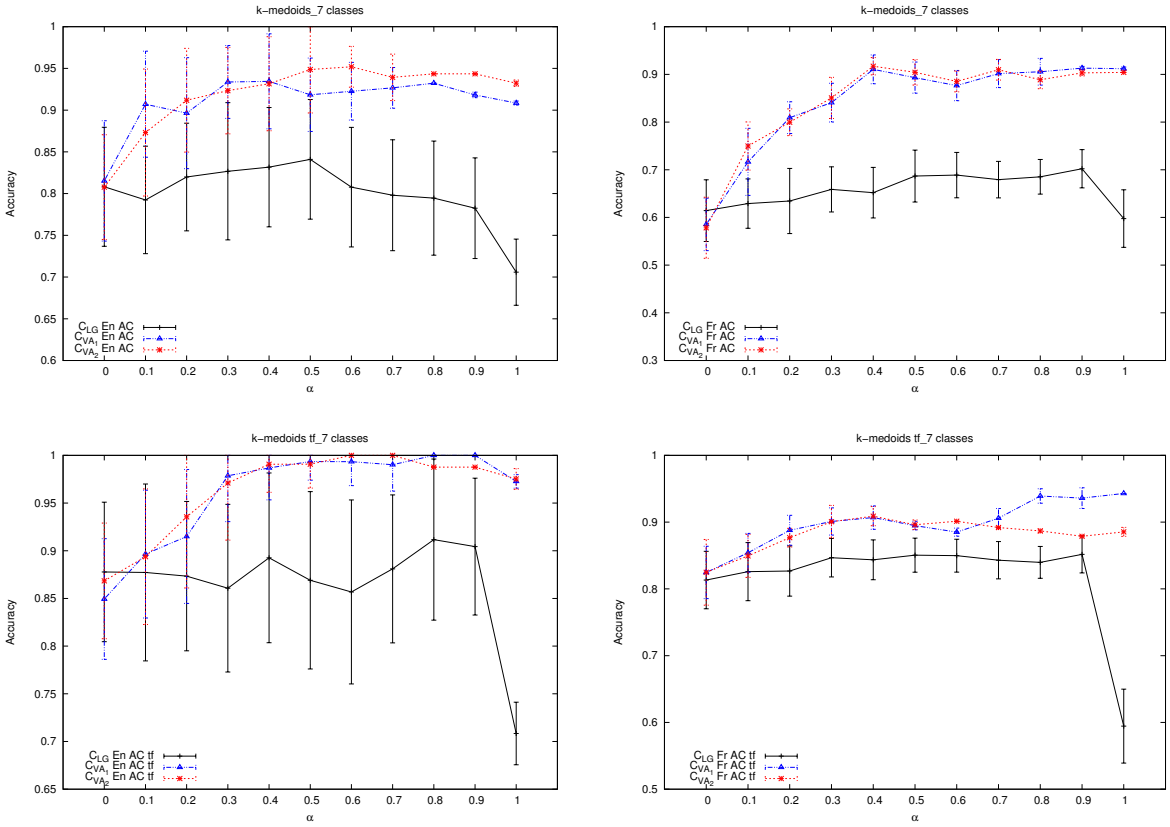

FIGURE 4.3 – Evaluation de la fusion des similarités *natives* et des similarités *induites* par mesure de comparabilité sur le clustering k-médoides au sens de la mesure *AC*. Le modèle vectoriel est exploité avec pondération *tf-idf* en haut, et avec pondération *tf* en bas.

La Figure 4.5 présente les variations de la valeur *DB* lorsque le paramètre α varie pour les trois mesures de comparabilité testées. Nous montrons ici que pour C_{VA_1} et C_{VA_2}, cette valeur diminue régulièrement lorsque α augmente, mais pour la mesure C_{LG}, cette valeur augmente toujours en général. La fusion de la comparabilité et de similarités a donc un impact significatif sur les trois mesures, mais cet impact est plutôt négatif pour la mesure C_{LG} et plutôt positif pour les deux variantes. Pour ce test également, les variantes C_{VA_1} et C_{VA_2} semblent donc mieux adaptées à une tâche de SCF-clustering de documents bilingues thématiques, comparativement à la mesure C_{LG}. Ici également, la pondération *tf* conduit à un meilleur DB index que la pondération *tf-idf*.

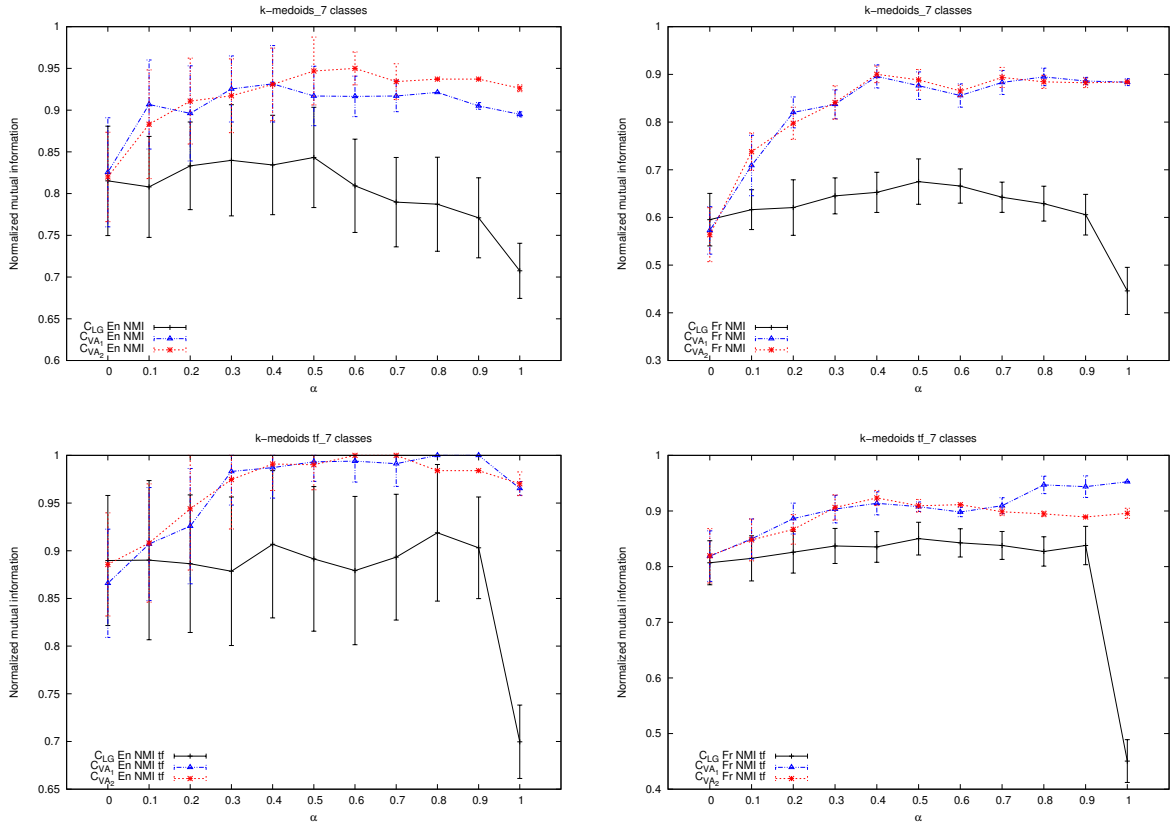

FIGURE 4.4 – Evaluation de la fusion des similarités *natives* et des similarités *induites* par mesure de comparabilité sur le clustering k-médoides au sens de la mesure *NMI*. Le modèle vectoriel est exploité avec pondération *tf-idf* en haut, et avec pondération *tf* en bas.

4.4.3 Impact de la fusion des similarités *natives* et des similarités *induites* par mesure de comparabilité sur un clustering hiérarchique ascendant avec les pondérations *tf-idf* et *tf*

Nous étudions ici l'impact de la fusion des similarités *natives* et des similarités *induites* par mesure de comparabilité sur un clustering hiérarchique ascendant (HAC) [126] [2] en exploitant le critère d'agglomération basé sur la similarité moyenne ("average-linkage") pour les trois mesures de comparabilité. Nous avons utilisé ici encore les critères *AC* et *NMI* avec les pondérations *tf-idf* et *tf* pour évaluer la qualité des clusters obtenus.

Les Figures 4.6 et 4.7 montrent qu'avec la pondération *tf-idf*, les critères *AC* et *NMI* peuvent être améliorés jusqu'à 15% pour le clustering des documents en langue anglaise et en langue française pour la mesure C_{VA_2}. L'amélioration est légèrement moindre pour la mesure C_{VA_1}. Pour la mesure C_{LG}, cependant, l'amélioration est très faible pour la meilleure valeur de α possible, et la fusion de similarités *natives* et *induites* dégrade en général la qualité du clustering

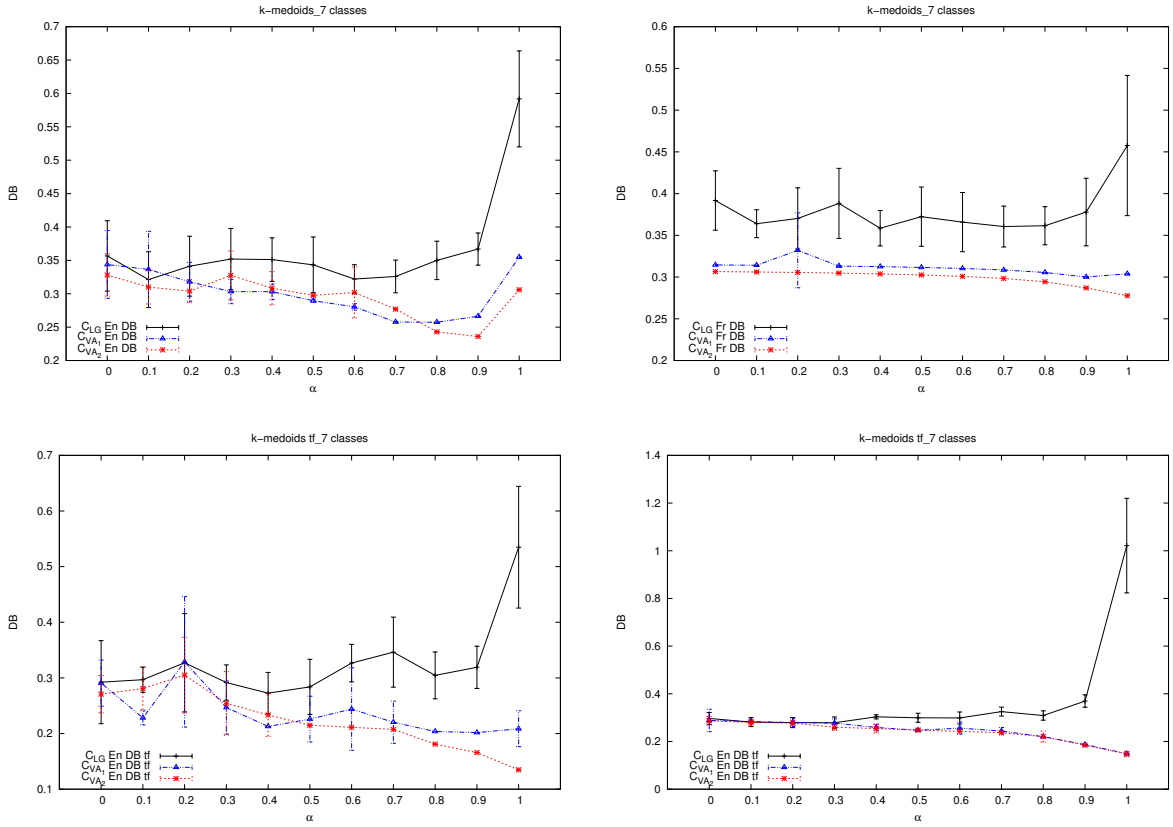

FIGURE 4.5 – Evaluation de la fusion des similarités *natives* et des similarités *induites* par mesure comparabilité sur le clustering k-médoides au sens de la mesure *DB*. Le modèle vectoriel est exploité avec pondération *tf-idf* en haut, et avec pondération *tf* en bas.

obtenu. Avec la pondération *tf*, les résultats sont légèrement meilleurs. Cela confirme que la pondération *tf* est préférable à la pondération *tf-idf* dans ces expérimentations. Nous avons donc choisi d'utiliser la pondération *tf* pour aligner les clusters comparables ci-après.

4.4.4 Alignement des clusters comparables par le modèle de mélange de la comparabilités (pour la variante C_{VA_2}) avec les similarités natives, en considérant un modèle vectoriel avec pondération *tf*

Pour quantifier l'utilité de notre modèle de mélange "similarités - comparabilité" dans le contexte de l'alignement des clusters comparables, nous avons calculé la comparabilité moyenne entre chaque paire de clusters bilingues, en tenant compte de tous les documents contenus dans chacun des clusters. La méthode de clustering utilisée est la méthode k-médoide et la mesure de comparabilité exploitée est C_{VA_2}.

Les Figures 4.8 et 4.9 montrent pour la mesure C_{VA_2}, l'alignement des clusters comparables

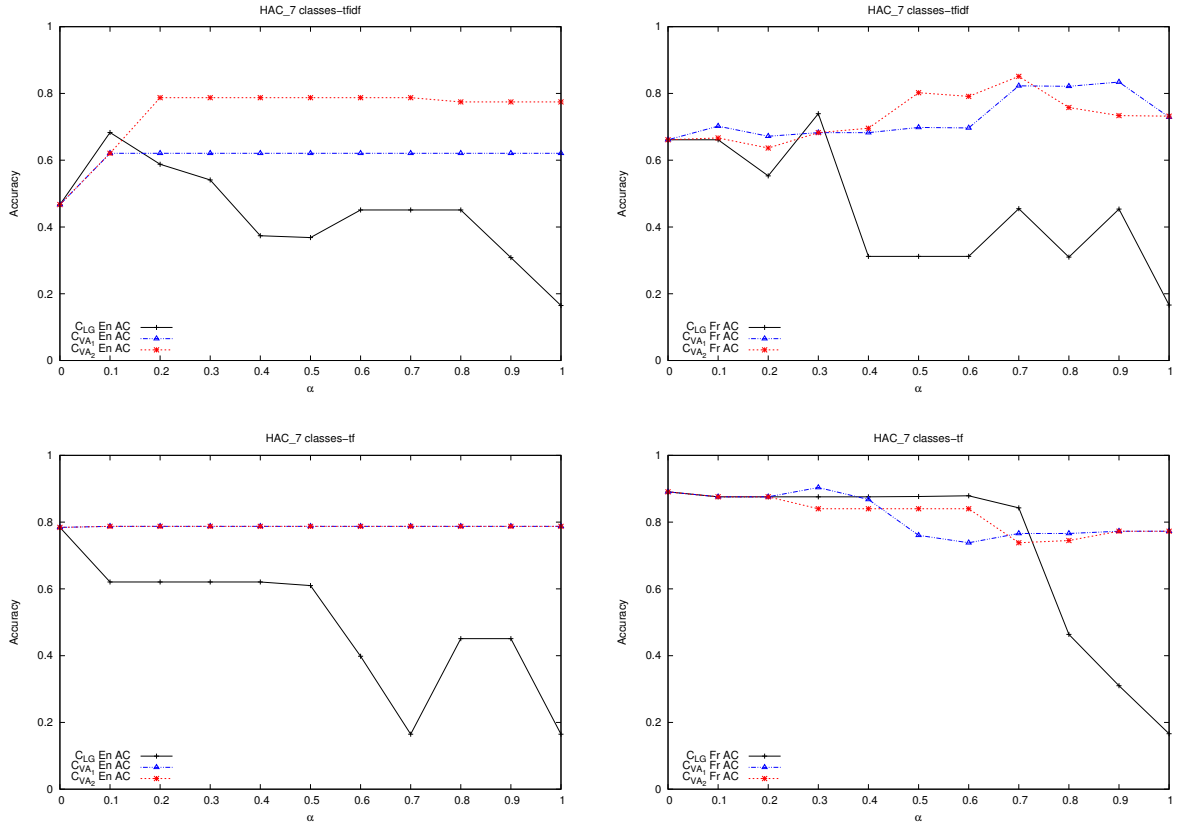

FIGURE 4.6 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par mesure de comparabilité sur un clustering hiérarchique ascendant en utilisant la mesure *AC*. Le modèle vectoriel est exploité avec pondération *tf-idf* en haut, et avec pondération *tf* en bas.

obtenu en combinant les similarités et les comparabilités avec $\alpha = 0,8$. La Figure 4.8 donne la matrice des comparabilités moyennes d'inter-clusters, tandis que la Figure 4.9 présente le graphe bipartite dérivé de la matrice précédente en gardant seulement pour chaque cluster (nœud), les deux meilleurs liens de comparabilité. Ces résultats montrent que l'approche de fusion de comparabilité/similarités associée à la mesure de comparabilité C_{VA_2} a parfaitement réalisé l'alignement des clusters comparables.

4.5 Expérimentations sur les corpus Wikipédia

Nous testons ci-après les trois corpus issus de Wikipédia : *Wikipedia_A*, *Wikipedia_B* et *Wikipedia_C* de manière analogue afin de consolider les résultats obtenus précédemment.

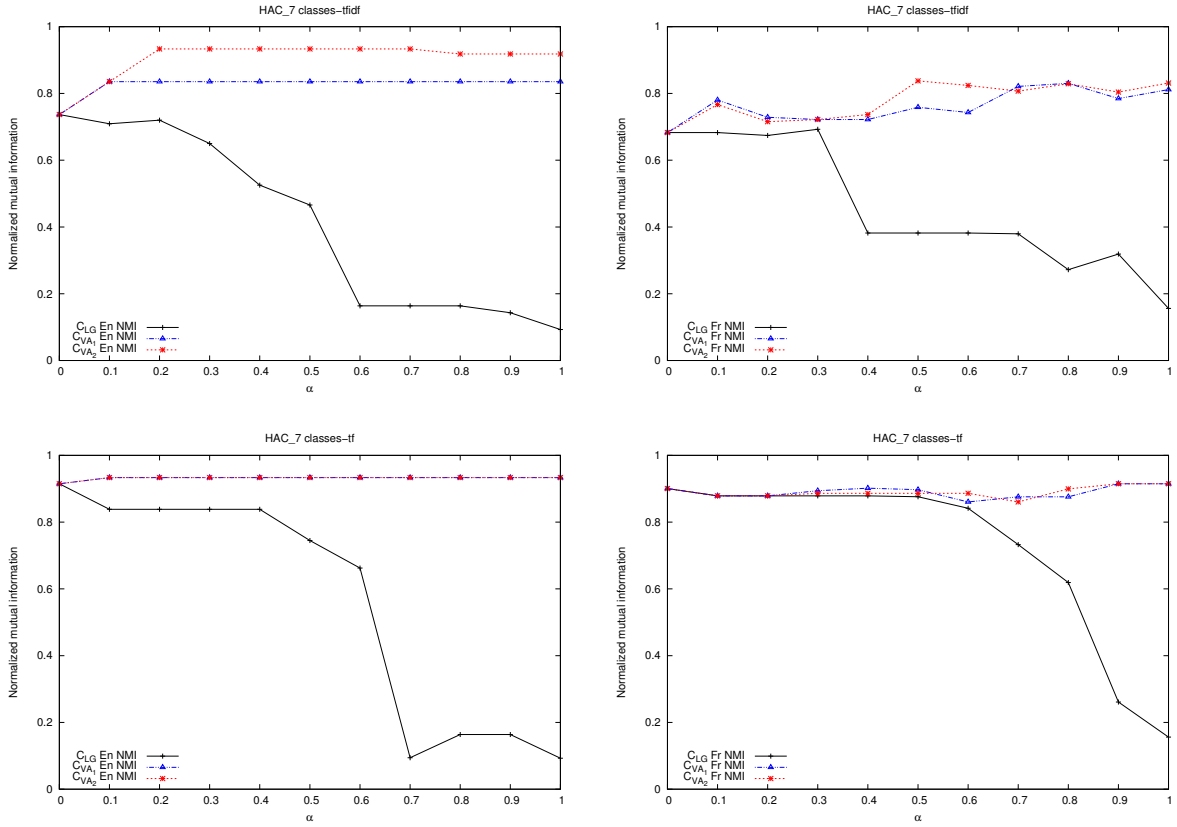

FIGURE 4.7 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par mesure de comparabilité sur un clustering hiérarchique ascendant au sens de la mesure *NMI*. Le modèle vectoriel est exploité avec pondération *tf-idf* en haut, et avec pondération *tf* en bas.

Comp (α =0,8)	**FRC1: David Beckham**	**FRC2: Mali**	**FRC3: Algérie**	**FRC4: République centrafricaine**	**FRC5: Mariage gay**	**FRC6: Pape**	**FRC7: Syrie**
ENC1: Syria	0,157	**0,278**	0,211	0,257	0,234	0,171	**0,384**
ENC2: Algeria	0,164	**0,346**	**0,453**	0,243	0,224	0,169	0,272
ENC3: Central African Republic	0,154	**0,275**	0,165	**0,366**	0,221	0,180	0,261
ENC4: Pope	0,168	0,207	0,169	0,247	**0,269**	**0,383**	0,214
ENC5: Gay marriage	0,200	0,221	0,163	**0,249**	**0,374**	0,217	0,239
ENC6: Mali	0,178	**0,379**	0,295	**0,302**	0,229	0,183	0,282
ENC7: David Beckham	**0,247**	0,180	0,149	0,168	**0,193**	0,156	0,173

FIGURE 4.8 – Comparabilités inter-clusters par modèle de fusion de la comparabilité et les similarités pour la variante C_{VA_2} et une valeur de α=0,8, avec la pondération *tf*

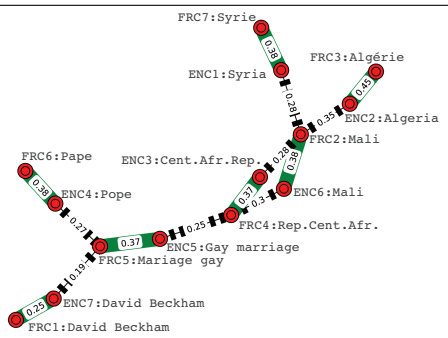

FIGURE 4.9 – Alignement des clusters par fusion de la comparabilité et les similarités avec le graphe de α=0,8 pour la variante C_{VA_2}, avec la pondération *tf*

4.5.1 Expériences sur le sous-corpus *Wikipedia_A*

4.5.1.1 Impact du modèle de fusion similarités/comparabilité sur la classification $1 - PPV$ avec les pondérations *tf-idf* et *tf*

Nous étudions tout d'abord l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de la classification $1 - PPV$ avec les pondérations *tf-idf* et *tf*, en faisant varier le paramètre $\alpha \in [0, 1]$.

Les Figures 4.10 et 4.11 montrent qu'avec la pondération *tf-idf*, la fusion des similarités *natives* avec les similarités *induites* par la comparabilité a un impact positif sur les deux mesures de comparabilité C_{VA_1} et C_{VA_2} en abaissant d'environ 1,5% le taux d'erreur avec C_{VA_2} et environ 1% avec C_{VA_1} pour les deux langues, par contre, avec la pondération *tf*, l'amélioration est d'environ 6% sur le taux d'erreur avec C_{VA_1} et C_{VA_2} pour les deux langues. Cependant, les taux d'erreur augmentent légèrement sur la mesure C_{LG}, même pour la meilleure valeur possible α. Pour cette expérimenation, les deux heuristiques de pondération *tf* et *tf-idf* conduisent à des résultats très proches.

4.5.1.2 Impact du modèle de fusion similarités/comparabilité sur le clustering k-médoides avec les pondérations *tf-idf* et *tf*

Nous étudions ici l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides [62] [63] pour les trois mesures de comparabilité avec les pondérations *tf-idf* et *tf*. Nous avons utilisé pour l'évaluation des résultats de clustering, les trois mesures *AC*, *NMI* et *DB*.

Les Figures 4.12 et 4.13 montrent qu'avec la pondération *tf-idf*, les deux critères *AC* et *NMI* peuvent être améliorés de près de 7% dans le cadre du clustering des documents en langue anglaise et en langue française pour C_{VA_2}, légèrement moins pour C_{VA_1}. Avec la pondération *tf*, l'amélioration pour ces trois mesures est faible pour *AC*, mais importante pour la mesure *NMI* qui peut être améliorée de près de 3% dans le cadre du clustering des documents en anglais et

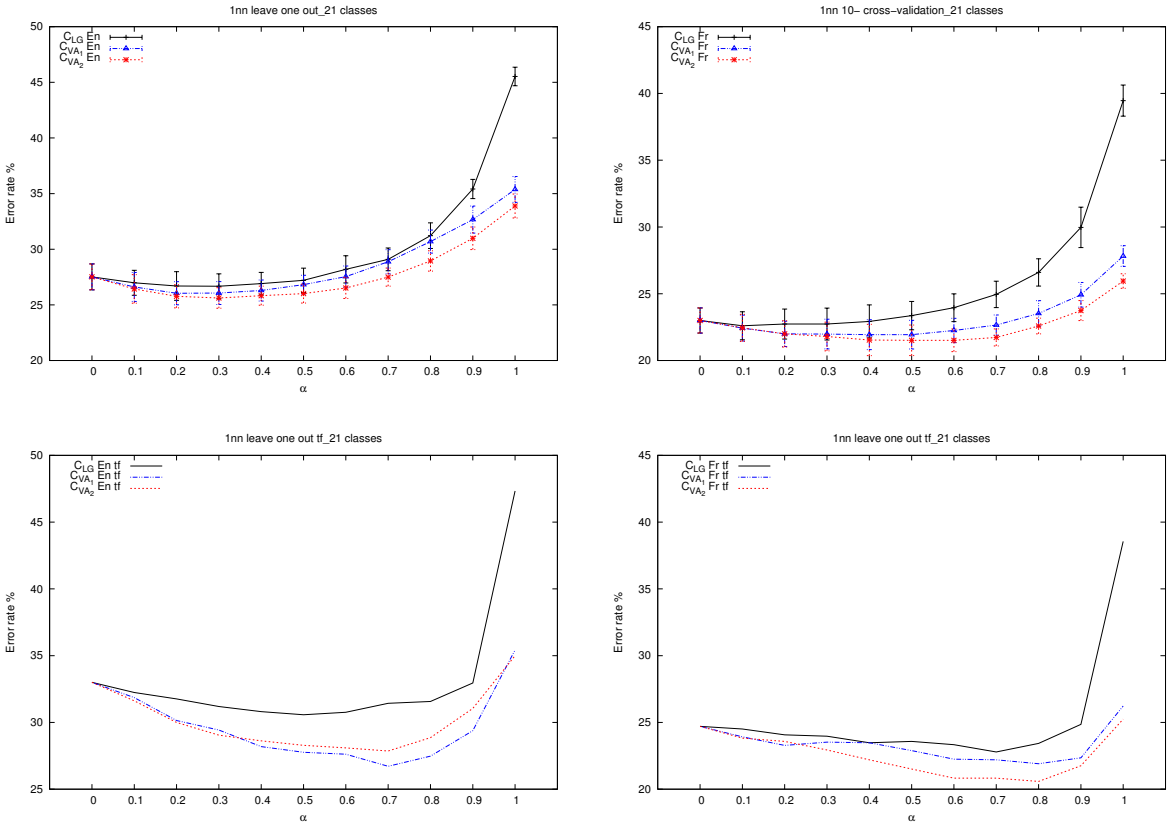

FIGURE 4.10 – Impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de "leave one out" de la classification $1 - PPV$ sur *Wikipedia_A*

6% en français pour C_{VA_1} et C_{VA_2}. Cependant, pour C_{LG}, l'amélioration est très limitée même en sélectionnant la meilleure valeur possible pour α. Ces expérimentations nous donnent aussi une indication pour le choix d'une bonne valeur de α. Par exemple, un α égal ou légèrement plus grand que 0,5 sera un bon choix. Dans cette expérimentation, nous avons constaté que les mesures *AC* et *NMI* avec la pondération *tf* sont toujours au-dessus des mesures *AC* et *NMI* obtenues avec la pondération *tf-idf*.

La Figure 4.14 présente les variations de la valeur DB lorsque le paramètre α varie pour les trois mesures de comparabilité testées. Nous montrons qu'avec la pondération *tf-idf*, pour C_{VA_1} et C_{VA_2}, cette valeur diminue régulièrement lorsque α augmente surtout pour la langue française. Ce phénomène est moins évident pour la langue anglaise. Avec la pondération *tf*, pour C_{VA_2}, cette valeur diminue régulièrement dans la plupart des cas lorsque α augmente surtout pour la langue française et c'est moins évident pour C_{VA_1}. Pour la mesure C_{LG}, l'effet est beaucoup moins évident pour les deux pondérations. La fusion de la comparabilité et des similarités a un impact évident sur les trois mesures, mais cet impact est positif pour les variantes C_{VA_1} et

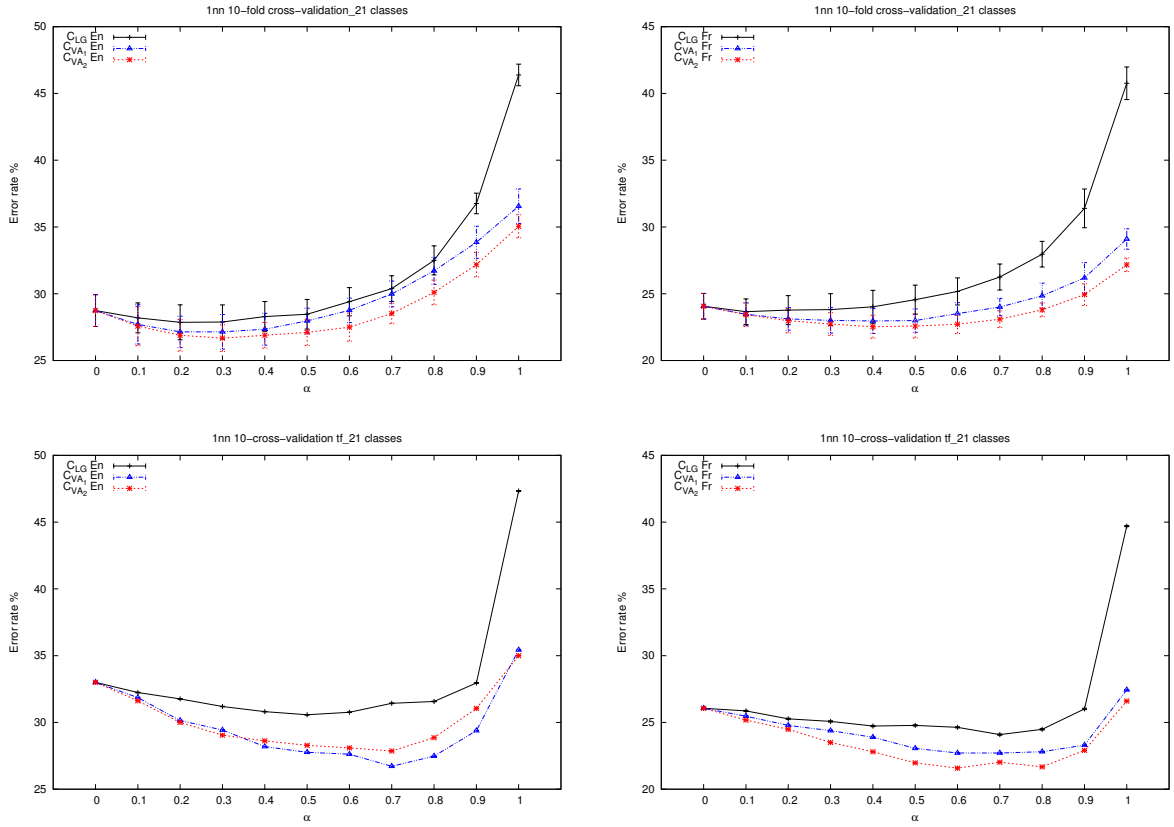

FIGURE 4.11 – Impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de "10 cross-validation" de la classification $1 - PPV$ sur *Wikipedia_A*

C_{VA_2} et plutôt négatif pour la mesure C_{LG}. Dans cette expérimentation, nous constatons que les deux pondérations n'amènent pas beaucoup de différence.

Selon les expérimentations précédentes, nous pouvons conclure ici que la pondération *tf* donne, dans la plupart des cas, plus d'effets positifs que la pondération *tf-idf* et dans les autres cas, les deux pondérations semblent très voisines. Cette situtation se présente pour les deux corpus différents : le premier issu de Flux RSS, le second issu de Wikipédia. Nous allons donc utiliser uniquement la pondération *tf* dans la suite de notre exposé.

4.5.2 Expériences sur le sous-corpus *Wikipedia_B*

4.5.2.1 Impact du modèle de fusion similarités/comparabilité sur la classification $1 - PPV$ avec la pondération *tf*

Nous étudions d'abord l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de la classification, en faisant varier le paramètre

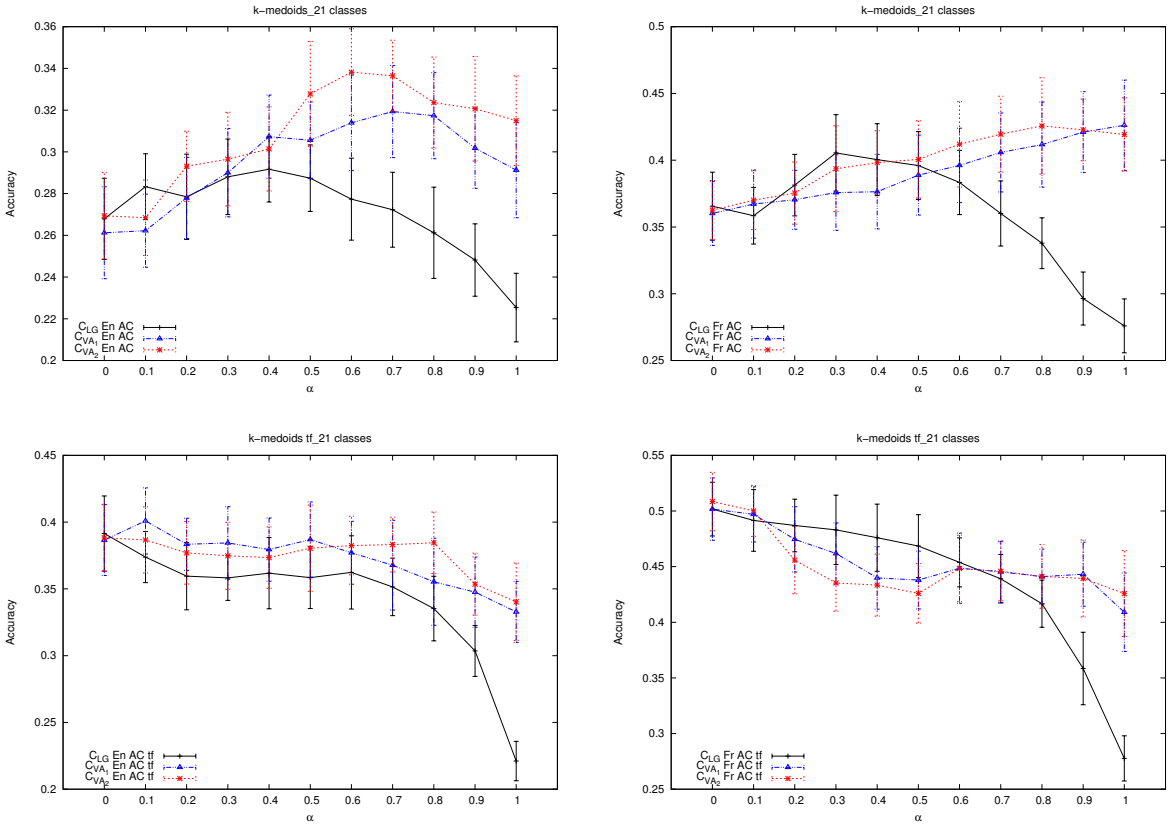

FIGURE 4.12 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *AC* avec les pondérations *tf-idf* et *tf* sur *Wikipedia_A*

$\alpha \in [0, 1]$.

Les Figures 4.15 et 4.16 montrent que la fusion des similarités *natives* avec les similarités *induites* par la comparabilité a un impact faible (légèrement positif sur la langue française) sur ce corpus pour les trois mesures de comparabilité qui se comportent de la même manière. La raison est que ce corpus est très propre, avec des catégories bien discriminées. Le taux d'erreur est donc très faible de sorte que notre approche apporte peu d'améliorations.

4.5.2.2 Impact du modèle de fusion similarités/comparabilité sur le clustering k-médoides avec la pondération *tf*

Nous étudions ici l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides [62] [63] pour les trois mesures de comparabilité avec la pondération *tf*. Nous avons utilisé les trois mesures *AC*, *NMI* et *DB* pour l'évaluation des résultats du clustering.

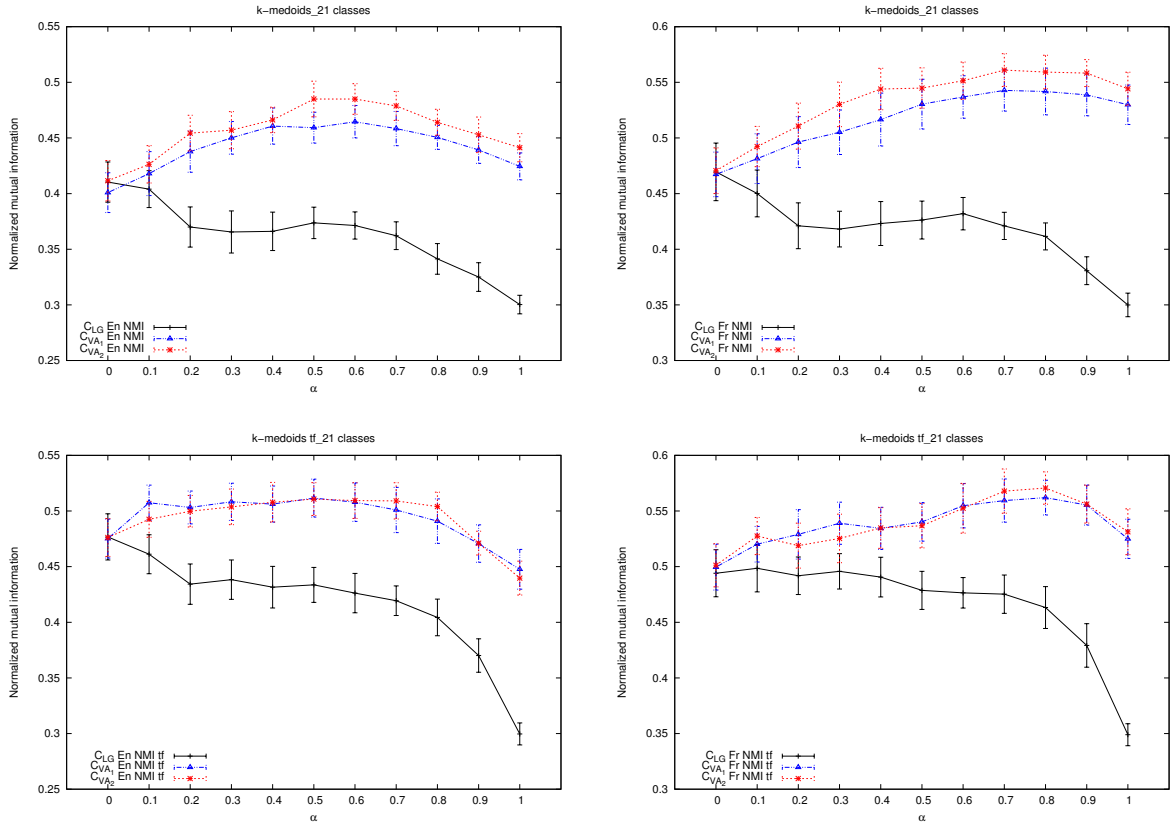

FIGURE 4.13 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *NMI* avec les pondérations *tf-idf* et *tf* sur *Wikipedia_A*

La Figure 4.17 et la Figure 4.18 montrent que l'amélioration pour ces trois mesures est peu visible pour la langue anglaise, mais très significative pour la langue française pour les deux variantes (amélioration jusqu'à 10% pour la mesure *AC* et 7% pour la mesure *NMI*). Cependant, nous observons plutôt l'inverse pour C_{LG}.

La Figure 4.19 présente les variations de la valeur DB lorsque le paramètre α varie pour les trois mesures de comparabilité testées. Nous observons que pour C_{VA_1} et C_{VA_2}, cette valeur diminue régulièrement dans la plupart des cas lorsque α augmente pour la langue française. Il ne semble pas y avoir d'amélioration pour la langue anglaise. Par contre, pour la mesure C_{LG}, l'effet est plutôt négatif.

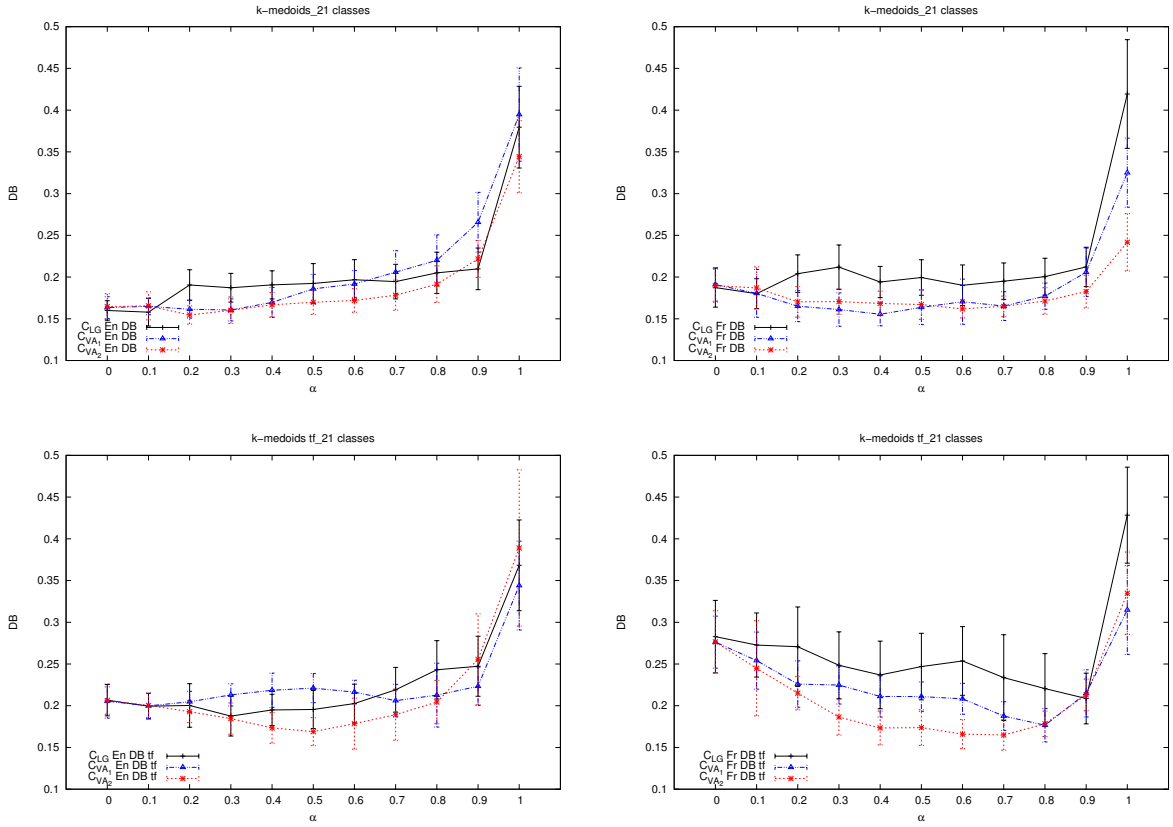

FIGURE 4.14 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure DB avec les pondérations *tf-idf* et *tf* sur *Wikipedia_A*

4.5.3 Expériences sur le sous-corpus *Wikipedia_C*

4.5.3.1 Impact du modèle de fusion similarités/comparabilité sur la classification $1 - PPV$ avec la pondération *tf*

Nous étudions ici l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de la classification, en faisant varier le paramètre $\alpha \in [0,1]$.

Les Figures 4.20 et 4.21 montrent que la fusion des similarités *natives* avec les similarités *induites* par la comparabilité a un effet significatif sur l'ensemble des trois mesures de comparabilité, en particulier pour les deux variantes C_{VA_1} et C_{VA_2} en abaissant d'environ 3% le taux d'erreur pour la classification des documents en anglais et de 1,5% pour les documents en français. La mesure C_{LG} améliore la précision de la classification, mais le choix d'une meilleure valeur de α est nécessaire (par exemple, $\alpha \in [0,7,0,8]$). Cependant, l'amélioration pour la mesurer C_{LG} est faible pour les deux langues, et elle est également moins stable que les deux

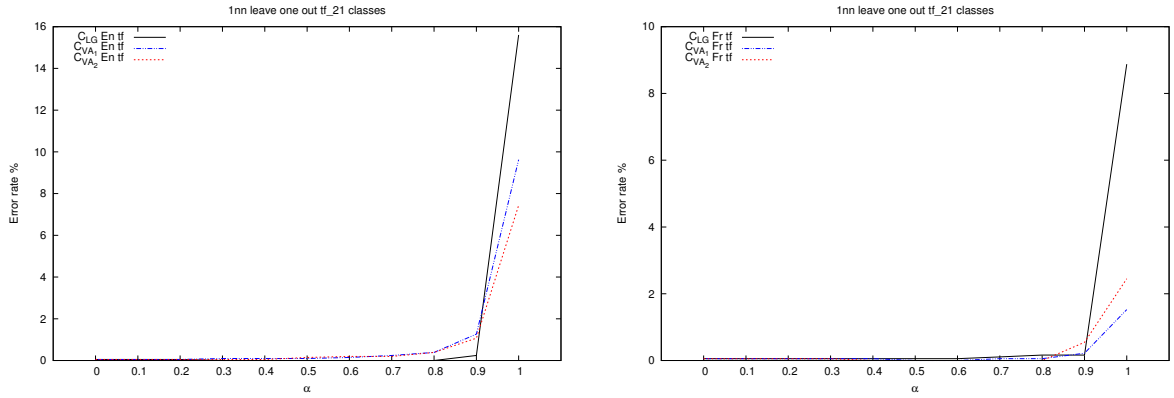

FIGURE 4.15 – Impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de "leave one out" de la classification $1 - PPV$ avec la pondération *tf* sur *Wikipedia_B*

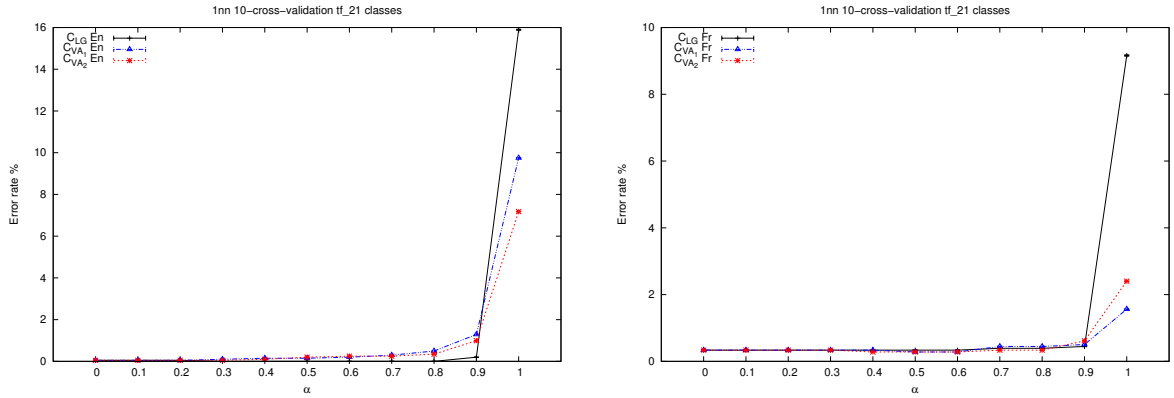

FIGURE 4.16 – Impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de "10 cross-validation" de la classification $1 - PPV$ avec la pondération *tf* sur *Wikipedia_B*

variantes.

4.5.3.2 Impact du modèle de fusion similarités/comparabilité sur le clustering k-médoides avec la pondération *tf*

Nous étudions ici l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides [62] [63] pour les trois mesures de comparabilité avec la pondération *tf*. Nous avons utilisé les mesures *AC*, *NMI* et *DB* pour l'évaluation des résultats du clustering.

Sur les Figures 4.22 et 4.23, nous observons que les deux mesures *AC* et *NMI* peuvent être

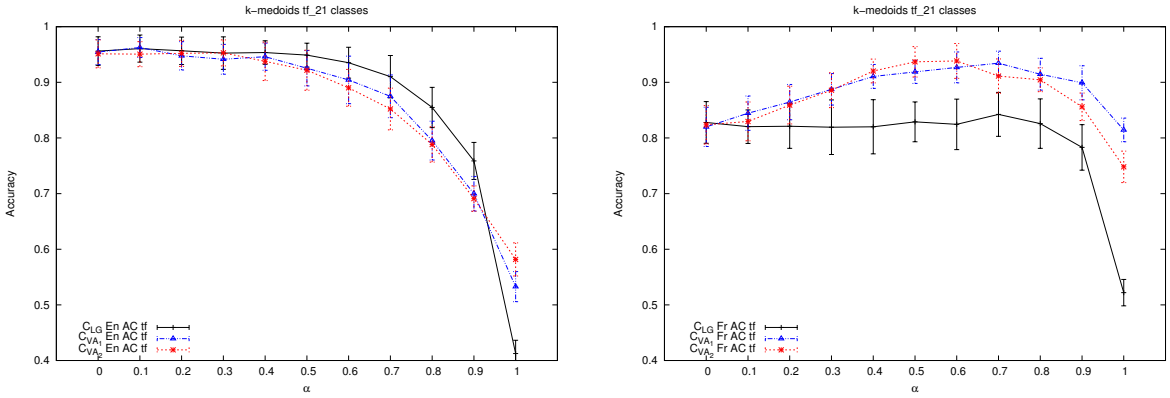

FIGURE 4.17 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *AC* avec la pondération *tf* sur *Wikipedia_B*

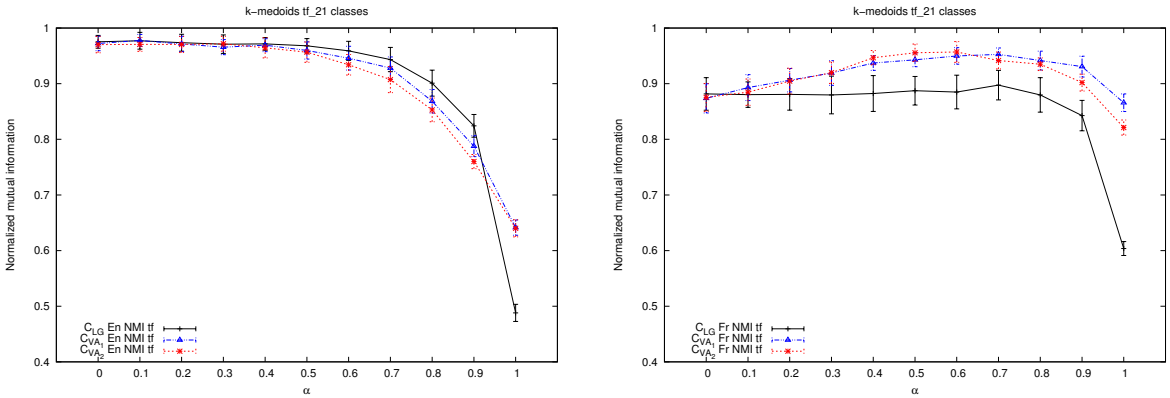

FIGURE 4.18 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *NMI* avec la pondération *tf* sur *Wikipedia_B*

améliorées jusqu'à 15% dans le cadre du clustering des documents en français et jusqu'à 3% dans le cadre du clustering des documents en anglais à la fois pour C_{VA_1} and C_{VA_2}. Cependant, C_{LG} n'apporte que peu d'améliorations pour les deux langues.

La Figure 4.24 donne les variations de la mesure de DB lorsque le paramètre α varie pour les trois mesures de comparabilité testées. Nous observons ici également que pour C_{VA_1} et C_{VA_2}, cette valeur diminue en choisissant bien un α, surtout pour la langue française, cependant, pour la mesure C_{LG}, cette valeur augmente en général. La fusion de la comparabilité et de similitudes a un impact évident sur les trois mesures, mais cet impact est très différent pour le C_{LG} et les deux variantes.

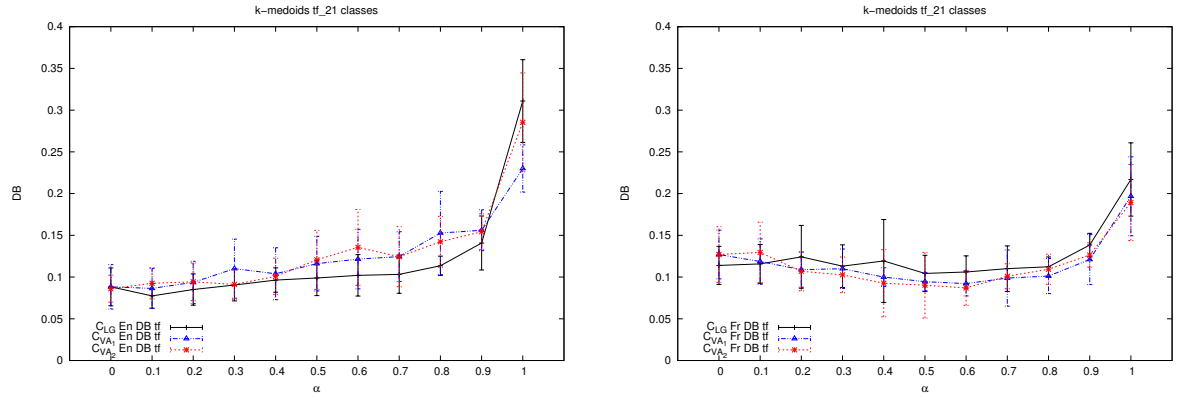

FIGURE 4.19 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *DB* avec la pondération *tf* sur *Wikipedia_B*

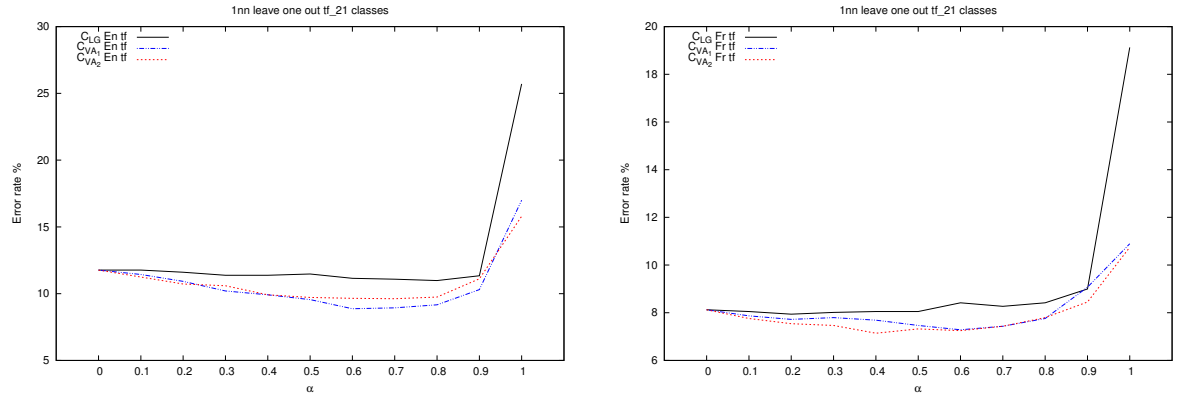

FIGURE 4.20 – Impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de "leave one out" de la classification $1 - PPV$ avec la pondération *tf* sur *Wikipedia_C*

4.5.3.3 Impact du modèle de fusion similarités/comparabilité sur le clustering hiérarchique ascendant avec la pondération *tf*

Nous étudions enfin l'impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering hiérarchique ascendant (HAC) [126] [2] pour les trois mesures de comparabilité. Nous avons utilisé les deux mesures *AC* et *NMI* pour évaluer la qualité du clustering.

La Figure 4.25 montre qu'il y a un peu d'amélioration sur la mesure *AC* pour les deux langues. Cependant, sur la Figure 4.26, l'effet sur la mesure *NMI* est inverse. Vu que les deux phénomènes sont observés sur *AC* et *NMI*, ce n'est plus nécessaire de tester sur la mesure

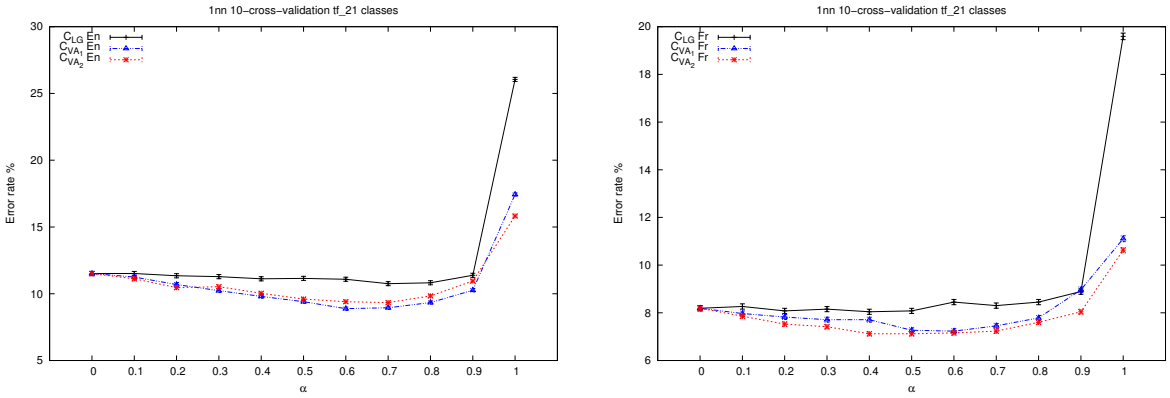

FIGURE 4.21 – Impact de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le taux d'erreur de "10 cross-validation" de la classification $1 - PPV$ avec la pondération *tf* sur *Wikipedia_C*

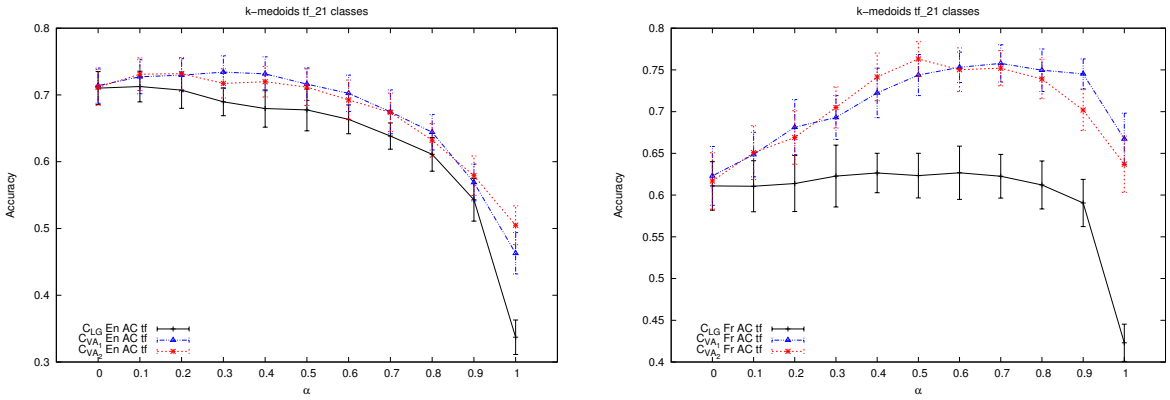

FIGURE 4.22 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *AC* avec la pondération *tf* sur *Wikipedia_C*

DB car nous pouvons conclure que le clustering HAC est moins adapté que le clustering k-médoides.

4.6 Analyse et éléments de conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche intitulée SCF-clustering et SCF-catégorisation pour le clustering et la classification des données bilingues thématiquement cohérentes. Cette approche est basée sur la notion de similarités *induites* par un graphe bipartite de comparabilité. La mise en œuvre de notre approche a été évaluée sur quatre corpus de test, le premier issu d'une collecte sur des flux RSS de type file d'agence de presse (*RSS7*) et les trois

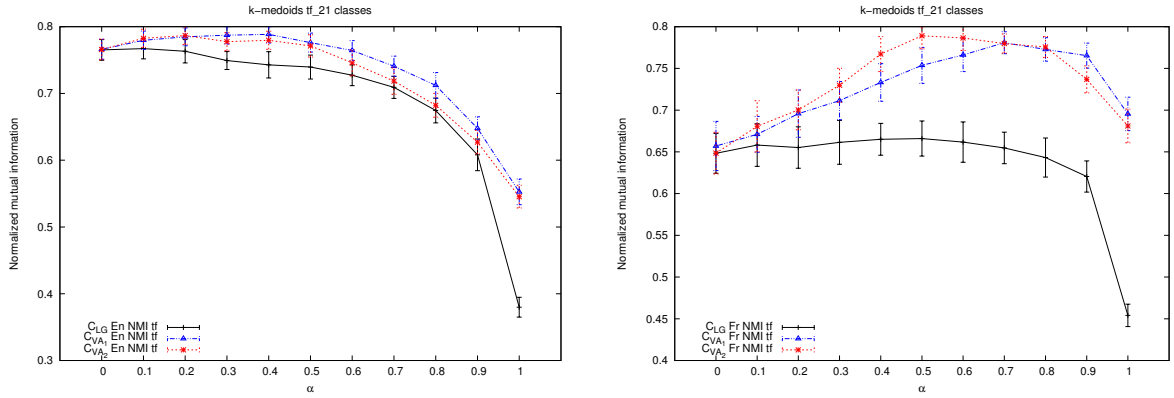

FIGURE 4.23 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *NMI* avec la pondération *tf* sur *Wikipedia_C*

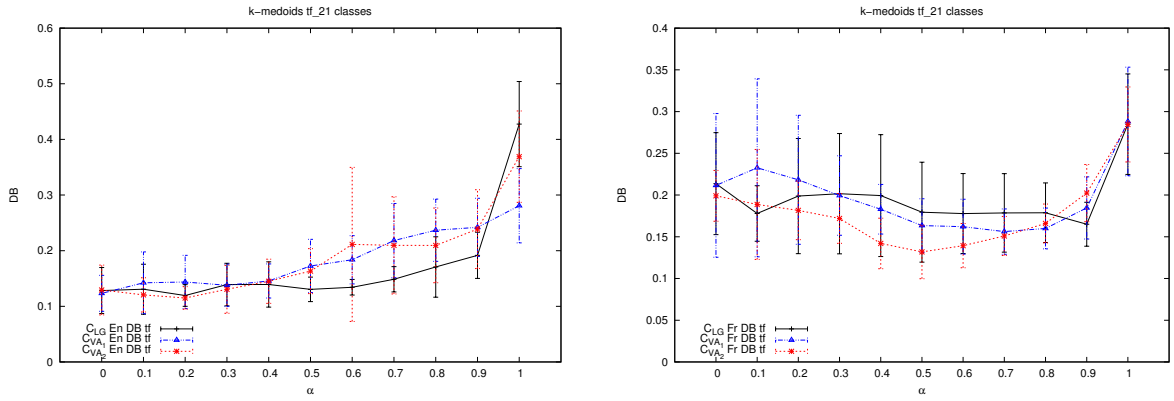

FIGURE 4.24 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering k-médoides en utilisant la mesure *DB* avec la pondération *tf* sur *Wikipedia_C*

autres, de taille plus importante, extraits à partir de quelques catégories Wikipédia et construits de manière à engendrer des cohésions thématiques variables : forte cohésion pour le corpus *Wikipedia_B*, faible cohésion pour le corpus *Wikipedia_A*, et cohésion thématique moyenne pour le corpus *Wikipedia_C*.

Nos expérimentations détaillées montre que notre modèle de mélange des similarités *natives* avec les similarités *induites* par une mesure de comparabilité a un impact significatif sur les qualités de classification et de clustering (surtout sur le clustering k-médoides) des données bilingues. Notre approche fonctionne particulièrement bien sur les tâches de classification et de clustering lorsque les mesures de comparabilité C_{VA_1} et C_{VA_2} sont exploitées, avec des résultats stables et robustes. Ce modèle a néanmoins un impact faiblement positif voire négatif lorsque la

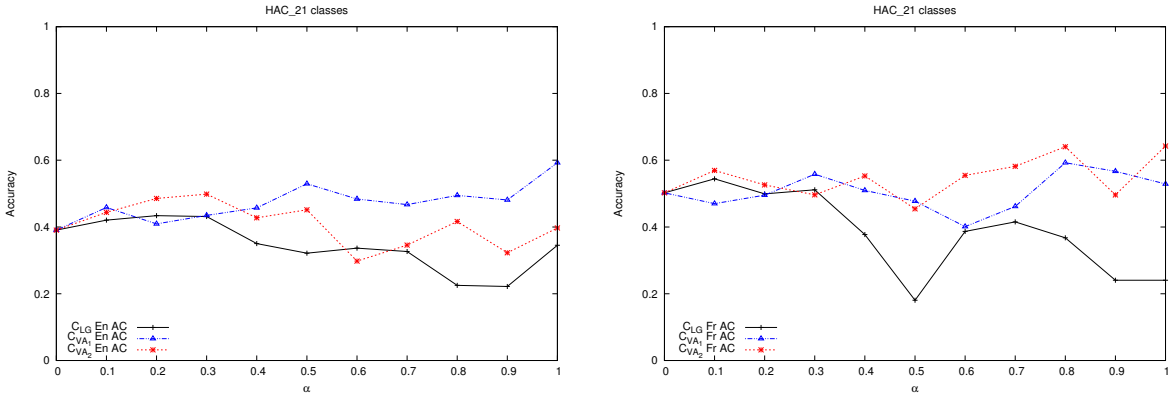

FIGURE 4.25 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering hiérarchique ascendant en utilisant la mesure *AC* avec la pondération *tf* sur *Wikipedia_C*

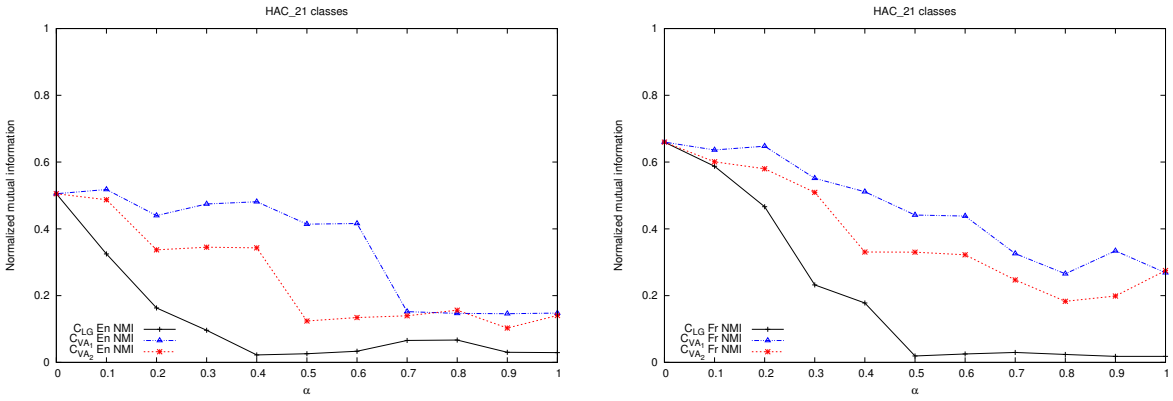

FIGURE 4.26 – Evaluation de la fusion des similarités *natives* avec les similarités *induites* par la comparabilité sur le clustering hiérarchique ascendant en utilisant la mesure *NMI* avec la pondération *tf* sur *Wikipedia_C*

mesure de comparabilité C_{LG} est exploitée. Ceci nous amène à conclure que la prise en compte des fréquences d'occurrences des entrées lexicales et des fréquences de leurs traductions dans la mesure de comparabilité a une importance cruciale pour la classification thématique ou le clustering thématique des documents bilingue anglais/français. Une explication qui peut être proposée est que ces fréquences, dans le contexte du modèle vectoriel à base de pondération de type *tf* (plutôt que *tf-idf*), s'intègre bien, c'est à dire de manière homogène, au sein du modèle de mélange des similarités *natives* avec les similarités *induites* proposé. Par ailleurs, d'après nos résultats, le choix de la valeur du paramètre de fusion α est lui aussi important. Une valeur α relativement moyenne (par exemple 0,5 ou 0,6), qui donne un poids égal ou légèrement plus important aux similarités *induites* comparativement aux similarités *natives*, est, en général, un

bon choix. Selon nos expérimentations, il apparaît que la mesure de comparabilité C_{VA_2} est la plus robuste des mesures de comparabilité testées. Enfin, nous avons également vérifié que le clustering k-médoides est plus adapté que HAC pour notre modèle.

Nous conseillons pour une tâche de clustering généralisée à base de k-médoids, l'utilisation de la mesure de comparabilité C_{VA_2}, avec le modèle vectoriel pondéré par *tf* et une valeur de α=0,5. Cette configuration est à la base de l'approche proposée dans un cadre de construction assistée de corpus comparables thématiques, décrite dans le chapitre qui suit.

Cinquième partie

Contribution à la construction assistée de corpus bilingues comparables thématiques

Dans cette dernière partie contributive, nous développons une approche semi-supervisée, en vue de développer une assistance à la construction de corpus comparables. Cette approche est basée sur l'identification et l'extraction de co-clusters comparables en "forte" cohérence thématique. Les co-clusters obtenus sont, du moins le supposons nous, autant de briques utilisables pour la construction de corpus comparables, exploitables notamment à des fins d'extraction de terminologies multilingues. Nous présentons tout d'abord les grandes étapes de l'approche proposée, puis nous étudions cette approche sur un cas d'usage construit à partir de la collecte de flux RSS accessibles sur le WEB. Nous traitons, au travers de quelques expériementations ciblées, des questions liées au choix des paramètres inhérents à cette approche et proposons une évaluation qualitative finale par le biais de l'alignement des clusters extraits par la méthode.

5

Quelques éléments pour la construction assistée de corpus comparables bilingues thématiques

Sommaire

5.1 Introduction

Nous présentons dans ce chapitre les bases d'une méthode exploitant le modèle de mélange de similarités natives et induites par comparabilité développé au chapitre précédent. Nous avons montré, au travers de nos expériences préliminaires, que ce modèle de mélange est relativement bien adapté à une tâche de co-clustering basée sur k-médoides lorsque le paramétrage suivant est considéré :

— un modèle vectoriel basé sur une pondération *tf* pour la représentation du contenu des documents,
— la mesure de comparabilité C_{VA_2} proposée dans le chapitre présentant notre première contribution,
— une valeur du paramètre de mélange α=0,5.

Nous explorons dans la suite de ce chapitre quelques pistes méthodologiques avec l'objectif d'assurer un *passage à l'échelle* des méthodes de co-clustering développées au chapitre précédent, afin d'être en capacité de proposer une assistance outillée à la construction de corpus comparables bilingues thématiques.

5.2 Construction semi-supervisée de corpus comparables par co-clustering de corpus bilingues

Nous proposons une approche incrémentale pour construire des corpus comparables thématiques à partir d'un corpus bilingue *brut* $\mathcal{C}_0$, collecté sur le web par exemple.

Nous détaillons ci-après les six étapes qui composent cette approche.

— **ETAPE-1 : Calcul et construction de la matrice de comparabilité pour les documents anglais et français du corpus *brut* initial $\mathcal{C}_0$.**
La mesure de comparabilité C_{VA_2} est utilisée pour calculer les comparabilités entre les paires de documents de langues différentes. La complexité pour évaluer la matrice de comparabilité est quadratique ($O(|\mathcal{C}_0|^2)$) en fonction de la taille du corpus $|\mathcal{C}_0|$.

— **ETAPE-2 : Filtrage du corpus initial $\mathcal{C}_0$ et production d'un corpus bilingue $\mathcal{C}_1$ plus dense au sens de la comparabilité.**
Cette étape, qui vise en premier lieu l'extraction d'un corpus plus dense du point de vue de la comparabilité, permet aussi de maintenir une complexité algorithmique acceptable. Le modèle de mélange proposé est caractérisé par une complexité $O(n^3)$, où n est le nombre de documents du corpus traité. Nous partons donc du principe que la taille du corpus $\mathcal{C}_1$ produit à l'issue de cette étape est sensiblement plus petite que la taille du corpus initial $\mathcal{C}_0$ ($|\mathcal{C}_1| \ll |\mathcal{C}_0|$).
Cette étape consiste à trier puis à filtrer les documents à partir de la matrice de comparabilité calculée lors de l'étape précédente, en utilisant un seuil de comparabilité minimal, β, un seuil γ portant sur le degré minimal des nœuds du graphe bipartite de comparabilité obtenu après élagage des liens associés à une comparabilité inférieure au seuil β, et le paramètre $\sigma = |C_1|/2$ définissant la taille du corpus filtré souhaité.
Nous présentons deux alternatives pour procéder au tri et au filtrage des documents.

1. **Tri séquentiel** sur les lignes et les colonnes de la matrice de comparabilité :
   a) Nous calculons pour chaque ligne de la matrice de comparabilité, le nombre *nc* de valeurs de comparabilité plus grandes que le seuil β et nous ne conservons que les lignes pour lesquelles $nc > \gamma$.
   b) Nous ordonnons les lignes par ordre décroissant (complexité $O(N \times log(N))$) et nous ne retenons que les σ premières lignes.
   c) De manière similaire, nous effectuons les deux traitements a) et b) précédents sur les colonnes de la matrice de comparabilité avec les mêmes valeurs pour les paramètres β, γ et σ.
   d) Le corpus $\mathcal{C}_1$ est constitué des documents anglais et français qui correspondent aux lignes et colonnes retenues.
2. **Tri simultané** sur les lignes et les colonnes de la matrice de comparabilité (en

Figure 5.1) :

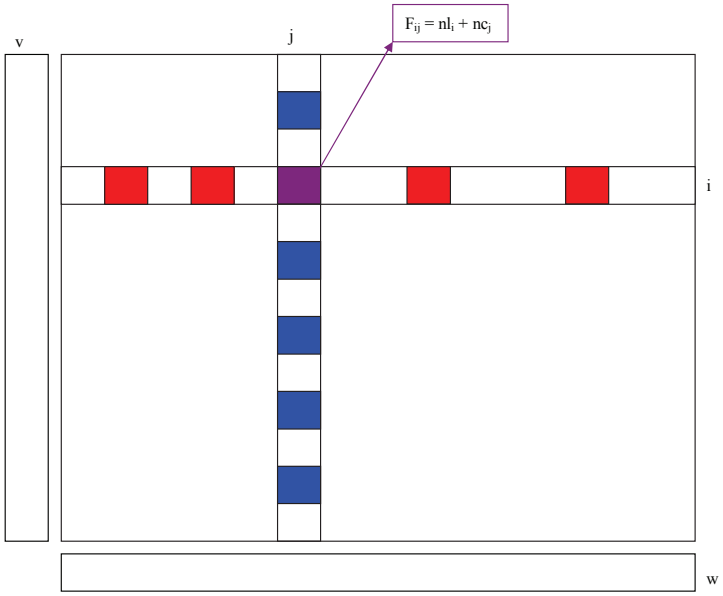

FIGURE 5.1 – Principe du **Tri simultané** basée sur le calcul de la matrice $F_{ij} = nl_i + nc_j$ et des vecteurs v et w

a) Nous calculons pour chaque ligne i de la matrice de comparabilité, le nombre nl_i de valeurs de comparabilité qui sont plus grandes que le seuil β et nous ne conservons que les lignes i pour lesquelles $nl_i > \gamma$.
b) De même, nous calculons pour chaque colonne j de la matrice de comparabilité, le nombre nc_j de valeurs de comparabilité plus grandes que le seuil β et nous ne conservons que les colonnes j pour lesquelles $nc_j > \gamma$.
c) Nous effectuons la somme du nombre de valeurs de chaque ligne i et celui de chaque colonne j et construisons la matrice $F_{ij} = nl_i + nc_j$ si la ligne i et la colonne j sont conservées, 0 sinon.
d) Nous calculons ensuite les vecteurs v et w définis de la manière suivante :
$v_i = Max_j F_{ij}$, $w_j = Max_i F_{ij}$
Les deux vecteurs sont triés par ordre décroissant (complexité $O(N \times log(N))$) et nous ne retenons que les σ premières valeurs pour chacun d'eux.
e) Le corpus $\mathcal{C}_1$ est constitué des documents français et anglais qui correspondent aux coordonnées des vecteurs v et w conservées.

Le **Tri simultané** est un peu plus complexe (complexité $O(3 \times N^2)$ au lieu de $O(2 \times N^2)$ pour le **Tri séquentiel**). Les trois paramètres β, γ et σ sont ajustables en fonction des données et des besoins.
A l'issue de cette étape, nous construisons la matrice de comparabilité et les matrices de

similarité *native* et *induite* pour le corpus $\mathcal{C}_1$ (chaque ligne conservée, respectivement chaque colonne, correspond à un document anglais, respectivement français).

— **ETAPE-3 : Détermination du nombre K_0 de clusters initiaux**.
En exploitant, pour chaque langue, les sorties produites par un clustering k-médoides en faisant varier le nombre de clusters k, nous effectuons le calcul des similarités intra et inter clusters moyennes δ_{intra} et δ_{inter} pour déterminer un nombre initial de clusters K_0. Si les clustering k-médoides sont effectués de manière indépendante pour chacune des deux langues, le modèle de mélange des similarités natives et induites par la comparabilité est néanmoins exploité, ce qui préserve donc la dimension de co-clustering.
Les valeurs δ_{intra} et δ_{inter} pour un clustering $\mathcal{C}$ contenant Nc clusters sont définis ci-dessous :

$$\delta_{intra}(\mathcal{C}) = \frac{1}{Nc}\sum_{i=1}^{Nc}\Big(\frac{1}{|C_i|}\sum_{d,d'\in C_i} S'_O(d,d')\Big) \tag{5.1}$$

$$\delta_{inter}(\mathcal{C}) = \frac{1}{Nc(Nc-1)}\sum_{i=1}^{Nc}\Big(\sum_{j}^{Nc} S'_O(m_i,m_j)\Big) \tag{5.2}$$

où $S'_O(d,d')$ et $S'_O(m_i,m_j)$ sont les similarités fusionnées, par modèle de mélange de similarités natives et induites, telles que définie dans l'équation 4.4. m_i est le médoide associé au cluster i, c'est-à-dire l'élément d'un cluster dont la distance moyenne à tous les autres éléments du cluster est minimale, ou encore l'élément le plus central dans ce cluster. A l'issue de cette étape, une fois que l'examen des courbes δ_{intra} et δ_{inter} lorsque K varie est effectué, K_0 est fixé manuellement.

— **ETAPE-4 : Filtrage des paires de clusters (Anglais-Français) fortement comparables**.
L'objectif ici est, à partir d'un ensemble de clustering k-médoides avec $k = K_0$, de déterminer les paires de clusters à conserver ainsi que les documents à conserver au sein de ces clusters. En pratique, nous fixons un deuxième seuil φ de comparabilité inter-cluster, puis nous calculons le degré moyen du graphe bipartite des clusters alignés obtenu après élagage des liens associés à une comparabilité inférieure au seuil φ. L'évolution du degré moyen de ce graphe lorsque le seuil de comparabilité φ varie est un critère de décision pour le choix des clusters à conserver. Dans l'idéal, nous ciblons des alignements de clusters relativement *purs*, i.e. en lien de comparabilité qui tend vers une relation 1-vers-1. Cela signifie que le degré moyen du graphe bipartite de comparabilité doit tendre vers 1, tout en limitant le nombre de clusters *orphelins*, c'est-à-dire non alignés.
Compte tenu de la dépendance de l'algorithme des k-medoides aux conditions initiales, nous traitons non pas un seul clustering, mais un ensemble de clustering concurrents.

L'analyse de l'impact du seuil de comparabilité φ sur : 1) le nombre de clusters conservés, 2) le nombre de documents conservés et 3) le degré du graphe bipartite des clusters alignés fournit quelques éléments d'information pour une sélection manuelle du seuil de comparabilité φ acceptable pour l'utilisateur.
Une fois que le seuil de comparabilité φ est fixé, nous alignons les clusters obtenus par exploitation du modèle de mélange des similarités natives avec les similarités induites par la comparabilité en exécutant plusieurs fois l'algorithme des k-médoides, afin de faire varier suffisamment les conditions initiales.
A l'issue de cette étape, nous obtenons un corpus $\mathcal{C}_2$ constitué de clusters bilingues alignés.

— **ETAPE-5 : Vérification manuelle des clusters alignés**
Cette étape à pour but de valider (ou d'invalider) à la main les paires de clusters alignés du corpus $\mathcal{C}_2$ obtenu lors de l'étape précédente. Les médoides jugés correctement alignés sont conservés (Puisqu'il y a des documents mal classés, nous ne conservons pas les documents qui leurs sont attachés. Nous réaffecterons les documents aux paires de médoides plus tard pour garantir une meilleure qualité thématique des clusters.). Les médoides doublons sont également éliminés.
Cette étape doit faire intervenir en général une expertise thématique et linguistique. Nous obtenons un corpus $\mathcal{C}_3$ validé à la fin de cette étape de vérification.

— **ETAPE-6 : Enrichissement du corpus**
L'objectif de la phase d'enrichissement est l'augmentation de la volumétrie du corpus, nécessaire si le corpus $\mathcal{C}_3$ est trop restreint pour l'application ciblée.
Nous exploitons ici deux variantes pour estimer le degré d'*adéquation* $S(d,m_i,m_j)$ entre un document d et une paire de médoides alignés (m_i,m_j)
a) la première variante consiste simplement à fusionner, en utilisant la valeur du paramètre α, la similarité native entre le document candidat et le médoide de même langue avec la comparabilité entre le document candidat et le médoide de langue différente. Si d et m_i sont de même langue, nous aurons :

$$S_{v1}(d,m_i,m_j) = \alpha S_{O_1}(d,m_i) + (1-\alpha)C(d,m_j)$$

b) la deuxième variante consiste à exploiter le modèle de mélange obtenu à partir du document d et de l'ensemble des paires médoides alignés considéré. Si C_d est la matrice de comparabilité calculée sur cette base, nous aurons alors :

$$S_{v2}(d,m_i,m_j) = \alpha S_{O_l,C_d}(d,m_i) + (1-\alpha)S_{O_l}(d,m_i)$$

Pour cette deuxième variante, la prise en compte du médoide m_j est effectuée par le

biais de la matrice $C_d.C_d^T$ exploitée pour calculer les similarités induites par la mesure de comparabilité dans l'espace linguistique O_l d'appartenance du document d.
Nous exploitons un seuil de rejet, τ, portant sur S_{v1} ou S_{v2}, pour décider si le document d viendra enrichir un cluster ou non. En pratique, chaque document du corpus initial $\mathcal{C}_0$ est testé et viendra enrichir le corpus si les valeurs S_{v1} ou S_{v2} sont supérieures au seuil τ fixé. Tout autre corpus complémentaire peut bien évidemment être exploité pour enrichir encore le corpus. Le seuil d'ajout de document τ peut être ajusté en fonction des exigences exprimées par les utilisateurs en matière de comparabilité. Si τ est faible, on aura plus de documents dans chaque paire de clusters mais les documents de ces paires de clusters seront moins comparables, par contre, si τ est grand, on aura moins de documents dans chaque paire de clusters mais ces documents y seront plus comparables.
A l'issue de cette étape d'enrichissement, le corpus bilingue comparable thématique final, $\mathcal{C}_F$ est produit.

Cette approche est semi-supervisée dans la mesure où elle nécessite un paramétrage adapté au cas d'usage et un contrôle manuel que nous avons cherché à réduire au maximum. Elle exploite 7 paramètres qu'il convient de positionner au mieux, en fonction des besoins et ressources disponibles. Nous listons ces paramètres de manière synthétique ci-dessous :

1. le paramètre α est utilisé dans notre modèle de mélange pour fusionner les similarités *natives* et *induites* par la comparabilité,
2. le paramètre β détermine la valeur de comparabilité minimale pour le filtrage du corpus initial $\mathcal{C}_0$,
3. le paramètre γ représente le degré minimal des nœuds (documents) dans le graphe bipartite de comparabilité après élagage conditionné par le seuil β,
4. le paramètre σ détermine le nombre de documents que nous conservons dans le corpus filtré et densifié du point de vue de la comparabilité, $\mathcal{C}_1$,
5. le paramètre K_0 spécifie le nombre initial de clusters extractibles du corpus $\mathcal{C}_1$,
6. le paramètre φ est un seuil de comparabilité utilisé pour l'extraction des paires de clusters fortement comparables qui constitueront les corpus $\mathcal{C}_2$ puis $\mathcal{C}_3$ à l'issue de la phase de vérification manuelle,
7. le paramètre τ est exploité en tant que seuil d'*adéquation* pour l'ajout de documents susceptibles d'enrichir le corpus $\mathcal{C}_3$ pour aboutir à la production du corpus bilingue final $\mathcal{C}_F$ constitué de clusters alignés, en principe fortement comparables et thématiques.

Cette approche est présentée de manière schématique dans la Figure 5.2. Les heuristiques d'instanciation (de sélection) des paramètres précédemment spécifiés sont précisées dans la section expérimentale qui suit.

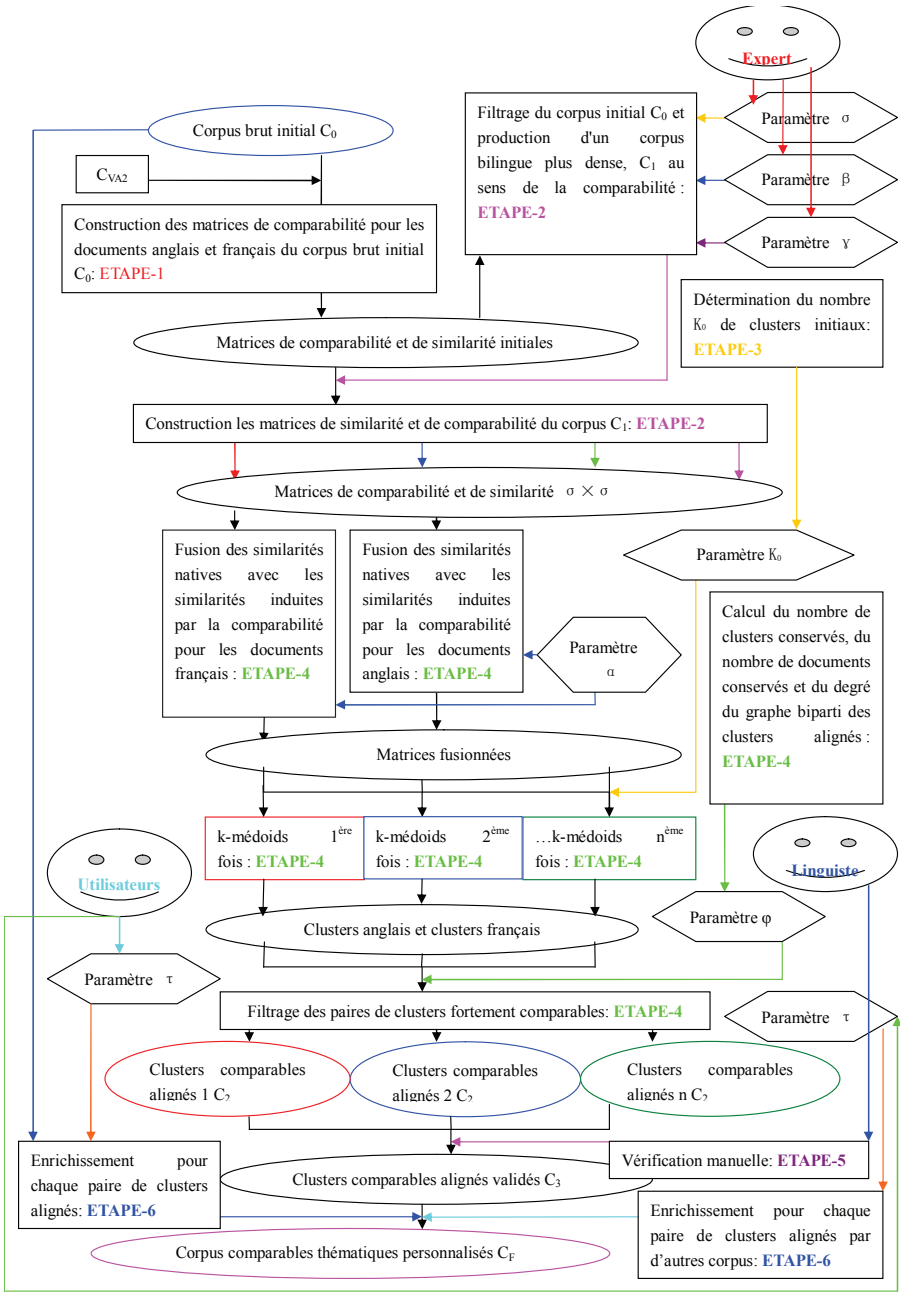

FIGURE 5.2 – Différentes étapes de notre approche pour la construction de corpus comparables thématiques

5.3 Corpus et dictionnaire exploités

Pour tester notre méthode semi-supervisée en situation réelle d'exploitation sur le WEB, nous avons exploité un crawler pour collecter des documents issus des 23 flux RSS sur une

période de six mois listés en Tableau 4.1. Le corpus *brut* initial ainsi constitué est composé de 39729 documents dont 18168 documents anglais et 21561 documents français. Ce corpus a été lemmatisé en utilisant le TreeTagger [115] [116] et la pondération *tf* a été utilisée dans le modèle vectoriel exploité pour représenter les contenus des documents pour chaque entrée du vocabulaire. Ce corpus pré-traité constitue le corpus $\mathcal{C}_0$ initial sur lequel nous appliquons notre approche semi-supervisée.

Le dictionnaire bilingue que nous avons utilisé pour évaluer la comparabilité entre documents et entre clusters de documents est à nouveau *dicElra*.

5.4 Expérimentations et résultats

Nous considérons en premier lieu les questions liées au choix des paramètres de l'approche proposée.

Le paramètre α utilisé pour la fusion des similarités *natives* et les similarités *induites* par la comparabilité à été étudié dans le chapitre précédent. Nous le fixons à la valeur 0,5 qui constitue en général un choix acceptable. En raison de la complexité de l'approche de fusion des similarités *natives* et des similarités *induites* par la comparabilité est $O(N^3)$, nous fixons le paramètre σ (le nombre de documents que nous gardons pour chacune des langues dans le corpus $\mathcal{C}_1$) à 1000. D'après nos expérimentations dans les deux chapitres précédents, nous avons constaté que si la valeur de comparabilité entre deux documents est supérieure à 0,1, les deux documents peuvent être relativement comparable. Nous affectons donc le paramètre β définissant la valeur de comparabilité minimale à 0,1 pour obtenir un compromis entre l'élémination des documents faiblement comparables et la conservation d'un nombre suffisant de documents. Nous fixons le paramètre γ (le nombre minimal de liens de comparabilité $> \beta$) à 10 pour garantir un degré minimal dans le graphe de comparabilité. Les paramètres β, γ, et σ sont définis dans l'ETAPE-2.

Le paramètre K_0 (le nombre initial de clusters (ETAPE-3)) et le paramètre φ (le seuil de rejet de comparabilité (ETAPE-4)) seront déterminés en fonction des expérimentations. Enfin, le paramètre τ (le seuil d'ajout des documents) est utilisé pour produire le corpus final. Il est ajustable par l'utilisateur en fonction des données traitées et des besoins (ETAPE-6).

Nous effectuons ensuite les expérimentations selon les six étapes et obtenons les résultats ci-dessous.

5.4.1 Expérimentations sur $\mathcal{C}_1$

5.4.1.1 Détermination du nombre initial de clusters K_0 en exploitant les similarités intra et inter clusters moyennes δ_{intra} et δ_{inter}

Nous déterminons ici un K_0 initial pour le clustering de $\mathcal{C}_1$ en analysant les variations des similarités intra et inter avec les valeurs des similarités intra et inter clusters moyennes δ_{intra} et δ_{inter} obtenues sur la base du clustering k-médoides, lorsque k varie (ETAPE-3).

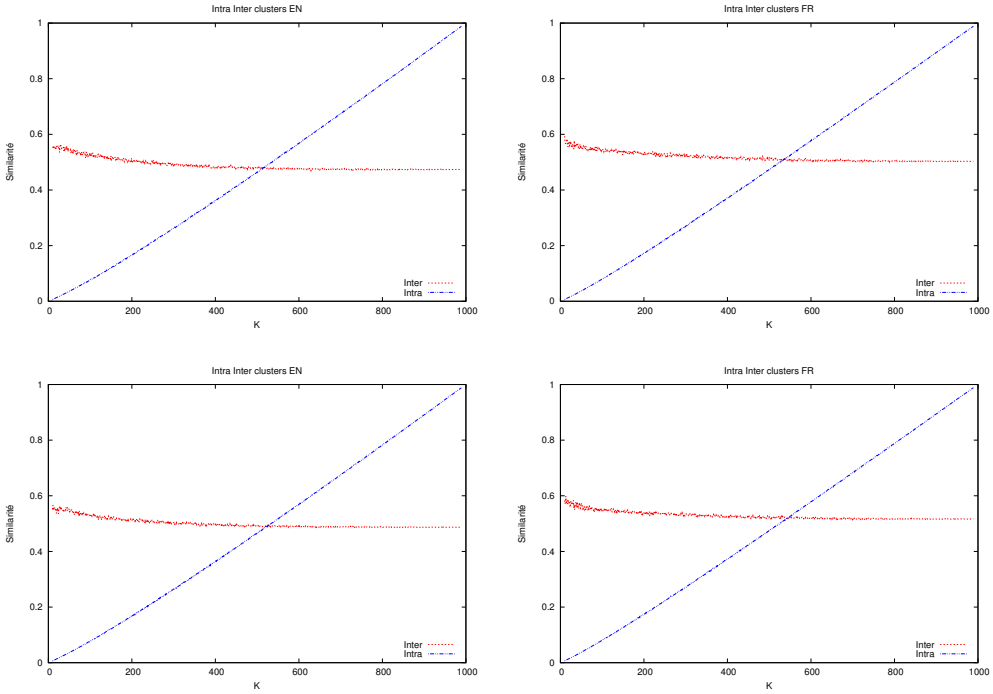

FIGURE 5.3 – Détermination du nombre initial de clusters K_0 pour $\mathcal{C}_1$ en exploitant les similarités intra et inter clusters moyennes δ_{intra} et δ_{inter} dans le clustering k-médoides, avec le **Tri séquentiel** en haut, et avec le **Tri simultané** en bas.

Dans la Figure 5.3, nous constatons que lorsque k augmente, les courbes δ_{intra} et δ_{inter} s'intersectent autour de $K_0 = 500$ avec le **Tri séquentiel** et autour de $K_0 = 550$ avec le **Tri simultané** pour les deux langues. Pour un bon clustering, il faut en général que la valeur de similarité intra clusters δ_{intra} soit grande et la valeur de similarité inter clusters δ_{inter} soit petite. Le point d'intersection ($K_0 = 500$ avec le **Tri séquentiel** et $K_0 = 550$ avec le **Tri simultané**) des deux courbes constitue d'après nous un bon compromis.

5.4.1.2 Détermination du seuil de comparabilité φ

Nous déterminons ici le seuil de comparabilité φ pour $\mathcal{C}_1$, en fonction du nombre de clusters conservés, du nombre de documents conservés et du degré du graphe bipartite des clusters alignés avec les seuils de comparabilité différents choisis (ETAPE-4).

Dans la Figure 5.4, nous essayons de déterminer une valeur φ de telle sorte que les clusters conservés demeurent en nombre suffisant, contiennent un nombre de documents également suffisant et que le graphe de comparabilité des médoides tende vers une relation 1-vers-1, i.e que le degré du graphe tende vers 1. Nous avons constaté et vérifié que lorsque φ est voisin de 0,45, avec le **Tri séquentiel** et également avec le **Tri simultané**, toutes les trois valeurs (le nombre de clusters conservés, le nombre de documents conservés et le degré du graphe

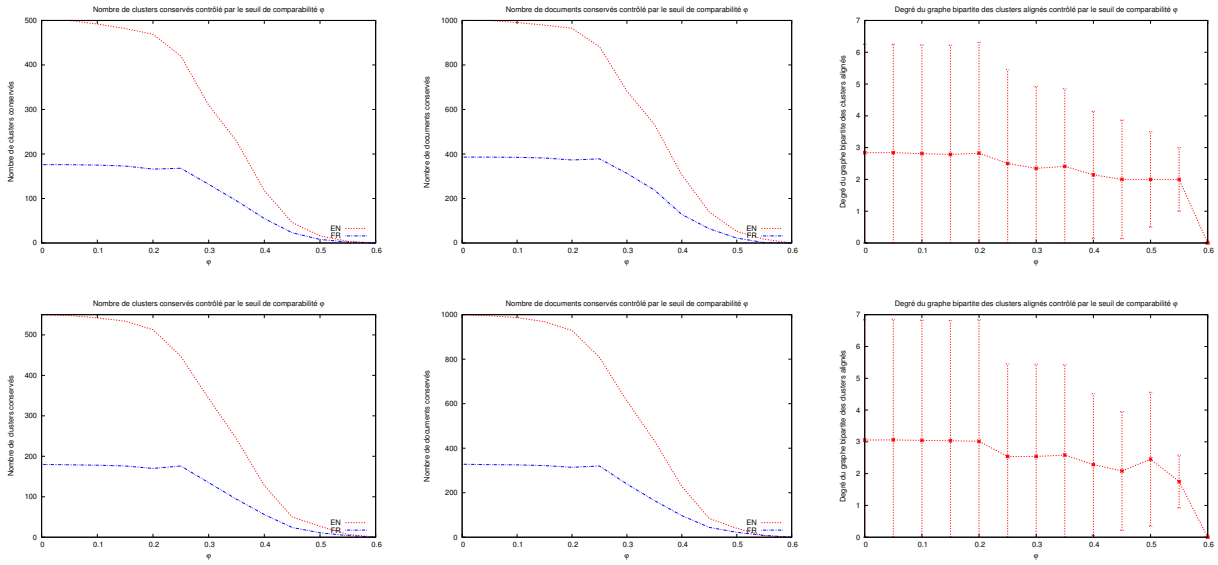

FIGURE 5.4 – Détermination du seuil de comparabilité φ en fonction du nombre de clusters conservés, du nombre de documents conservés et du degré du graphe bipartite des clusters alignés, avec le **Tri séquentiel** en haut, et avec le **Tri simultané** en bas.

bipartite des clusters alignés) sont stables. Nous fixons donc le seuil de comparabilité φ à 0,45. Ce seuil est utilisé pour aligner automatiquement les clusters de langues différentes.

5.4.1.3 Paires de clusters alignés

Nous présentons en Figure 5.5 deux exemples de paires de clusters alignés pour $\mathcal{C}_1$ (chaque cluster, étant représenté par son médoide) obtenus avec le **Tri séquentiel** et avec le **Tri simultané**. Chaque paire a obtenu la plus grande valeur de comparabilité après l'alignement des clusters comparables à l'issue de 4 exécutions des k-médoides avec le seuil de rejet de comparabilité $\varphi = 0,45$ (ETAPE-4). Les paires de clusters sont manuellement vérifiés (ETAPE-5).

Nous présentons en annexe les dix premiers clusters alignés triés par valeurs décroissantes de comparabilité avec le **Tri séquentiel** et avec le **Tri simultané**.

Dans la Figure 5.6, nous essayons de vérifier manuellement le nombre de clusters ajoutés et le nombre de clusters communs en fonction des itérations de k-médoides (ETAPE-5). Selon les résultats obtenus avec le **Tri séquentiel** et avec le **Tri simultané**, le nombre de clusters commun à tentance à augmenter et le nombre de clusters ajoutés à tentance à diminuer. Au niveau de la troisième ou quatrième itération, nous n'obtenons plus de nouveaux clusters à ajouter. Nous observons donc que les clusters à ajouter sont progressivement de moins en moins nombreux et au bout de quelques itérations (ici 3 ou 4), le nombre des clusters alignés devient stable.

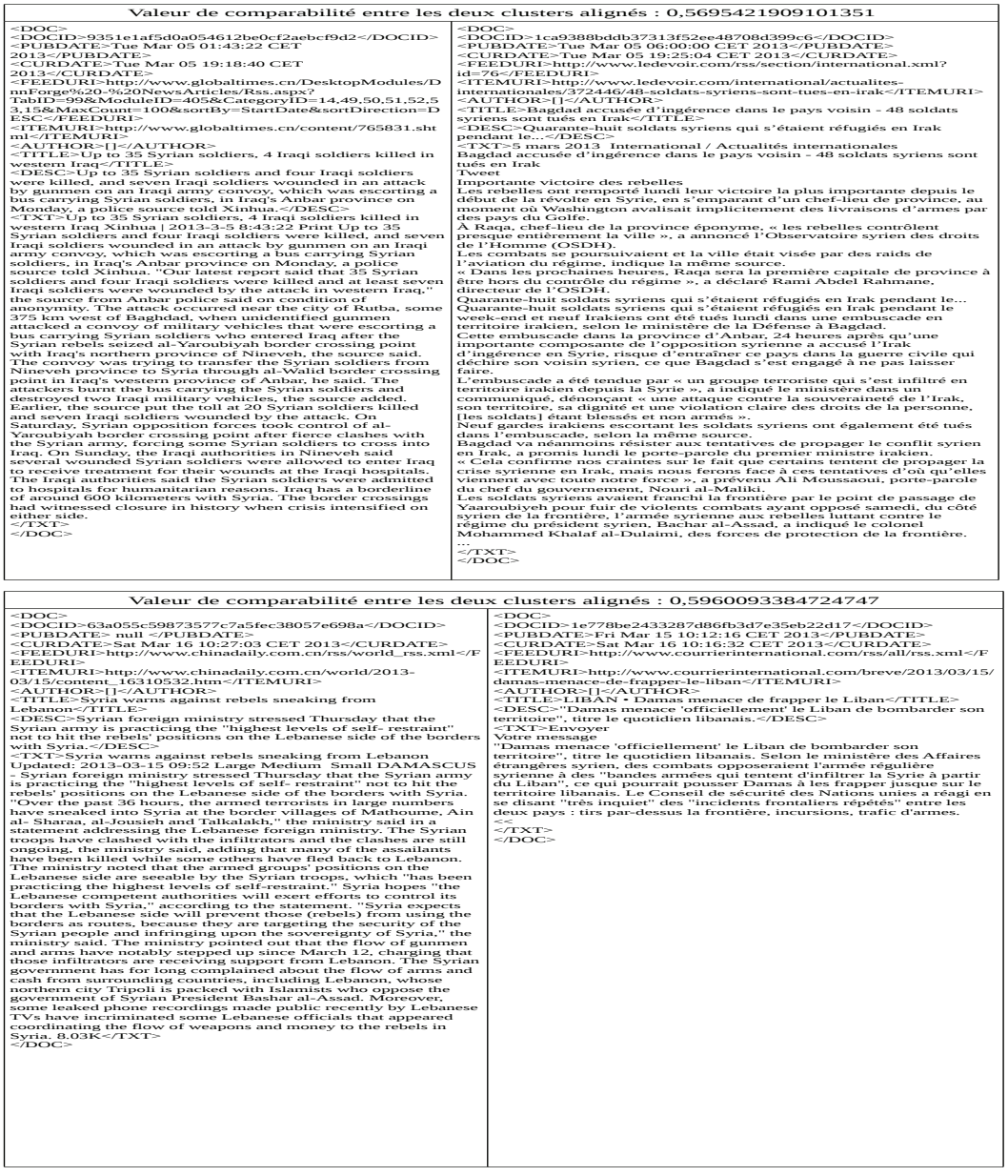

FIGURE 5.5 – Alignement des deux clusters (médoides) ayant la comparabilité la plus élevée, avec le **Tri séquentiel** en haut, et avec le **Tri simultané** en bas.

5.4.1.4 Nombre de documents conservés avec différents seuils d'ajout τ

Nous étudions ici le nombre de documents conservés avec différents seuils d'ajout τ à partir de $\mathcal{C}_1$, en fonction des deux mesures d'ajouts S_{v1} et S_{v2} (ETAPE-6).

Premièrement, nous illustrons le nombre de documents conservés avec différentes valeurs de S_{v1}, c'est-à-dire la somme de 50% de la valeur de similarité et 50% de la valeur de compa-

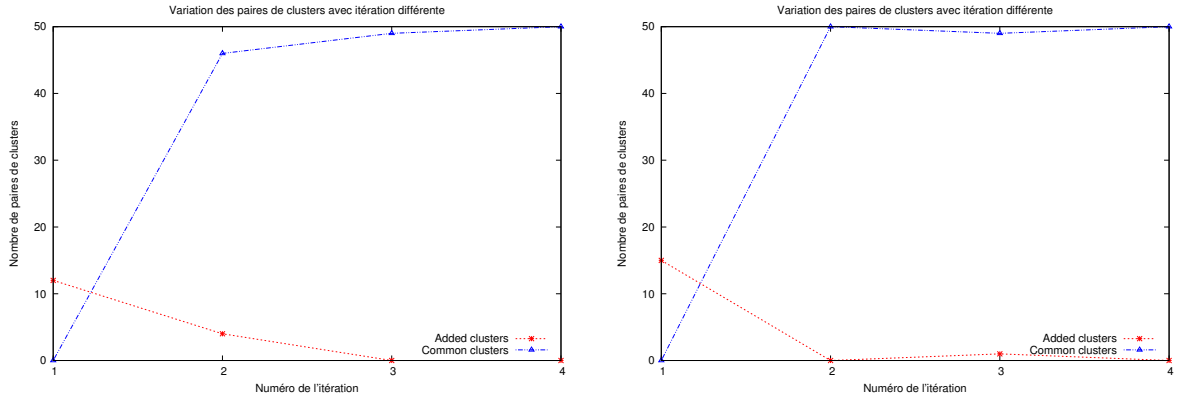

FIGURE 5.6 – Nombre de clusters ajoutés et nombre de clusters communs en fonction des itérations de k-médoides, avec le **Tri séquentiel** à gauche, et avec le **Tri simultané** à droite.

rabilité.

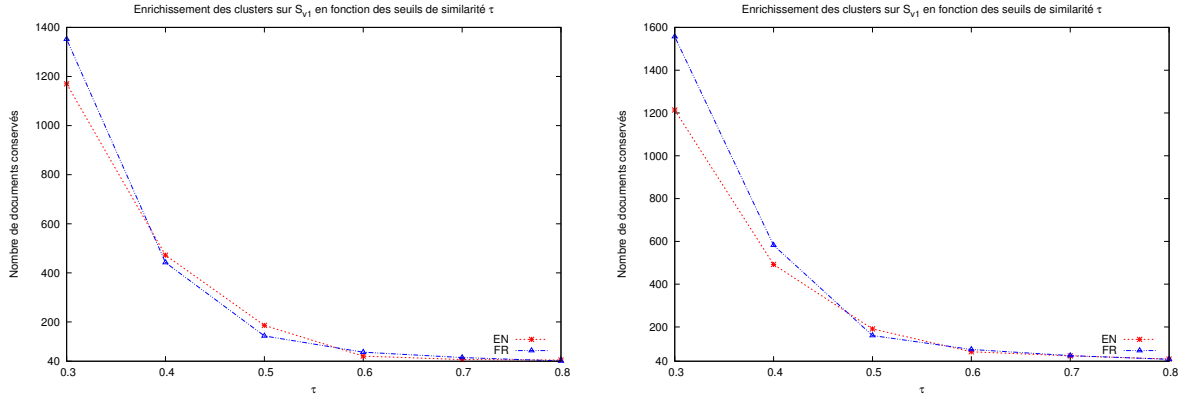

FIGURE 5.7 – Nombre de documents conservés avec différentes valeurs d'ajout en exploitant S_{v1}, avec le **Tri séquentiel** à gauche, et avec le **Tri simultané** à droite.

Deuxièmement, nous illustrons le nombre de documents conservés avec différentes valeurs de S_{v2}, c'est-à-dire la somme de 50% de la valeur de similarité native et 50% de la valeur de la similarité induite par la comparabilité.

Dans les Figures 5.7 et 5.8, nous présentons le nombres de documents ajoutés en fonction du seuil d'ajout τ (sur S_{v1} ou sur S_{v2}). Les valeurs de τ pour les deux mesures ne représentent pas le même niveau de comparabilité. Pour S_{v1}, la valeur de τ qui diminue considérablement le nombre de documents ajoutés est plus bas que pour S_{v2}. C'est-à-dire que pour S_{v1} et $\tau = 0,5$, on a presque le même niveau de comparabilité qu'en choisissant $\tau = 0,7$ et S_{v2}. Cependant, d'après nos expérimentations et notre intuition, S_{v2} est plus recommandée que S_{v1} pour "customiser" dans chaque paire de clusters alignés en fonction des besoins, les besoins différents sur

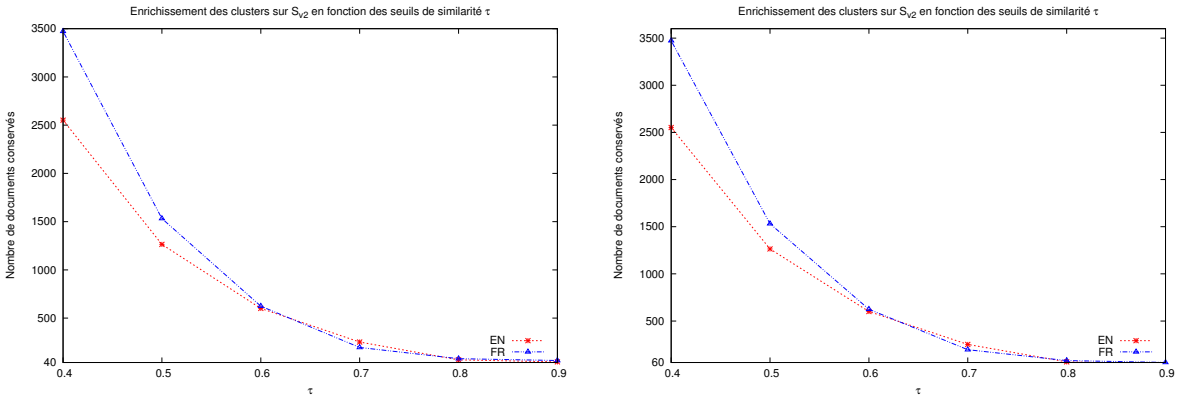

FIGURE 5.8 – Nombre de documents conservés avec différentes valeurs d'ajout en exploitant S_{v2}, avec le **Tri séquentiel** à gauche, et avec le **Tri simultané** à droite.

demande : les documents plus ou moins comparables au sein des clusters (plus grande est la comparabilité, plus petits sont les clusters). Par contre, les documents ajoutés à $\mathcal{C}_1$ avec le **Tri simultané** sont un peu plus nombreux qu'avec le **Tri séquentiel**.

5.4.1.5 Comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1} ou S_{v2}

Nous étudions ici sur $\mathcal{C}_1$, la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1} ou S_{v2} (ETAPE-6).

Premièrement, nous illustrons la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1}.

Deuxièmement, nous illustrons la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v2}.

Dans les Figures 5.9 et 5.10, nous montrons la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1} et sur S_{v2}. Dans la Figure 5.9, nous constatons qu'en général, la comparabilité moyenne de chaque paire de clusters augmente lorsque τ augmente, par contre, lorsque $\tau \geq 0,7$, la valeur de la comparabilité moyenne change très peu. Cela montre que sur S_{v1}, un τ pertinent peut être initialement défini autour de 0,5. Dans la Figure 5.10, la comparabilité moyenne de chaque paire de clusters augmente en général également lorsque τ augmente, mais lorsque $\tau \geq 0,8$, la valeur de la comparabilité moyenne devient très stable. Cela montre que sur S_{v2}, un τ pertinent peut être initialement défini autour de 0,7. En comparant les deux figures, nous constatons de nouveau que la fusion des similarités *natives* et les similarités *induites* par la comparabilité a un impact positif dans l'enrichissement car lorsque $\tau \geq 0,8$ (même $\tau = 0,9$), il y a plus des documents ajoutés. Et puis, intuitivement, encore une fois, S_{v2} est plus intéres-

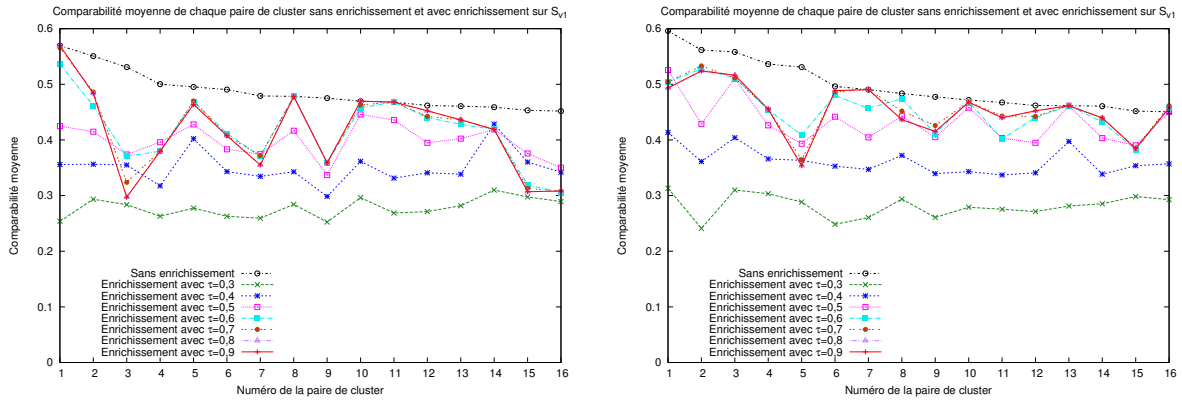

FIGURE 5.9 – Comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1}, avec le **Tri séquentiel** à gauche, et avec le **Tri simultané** à droite.

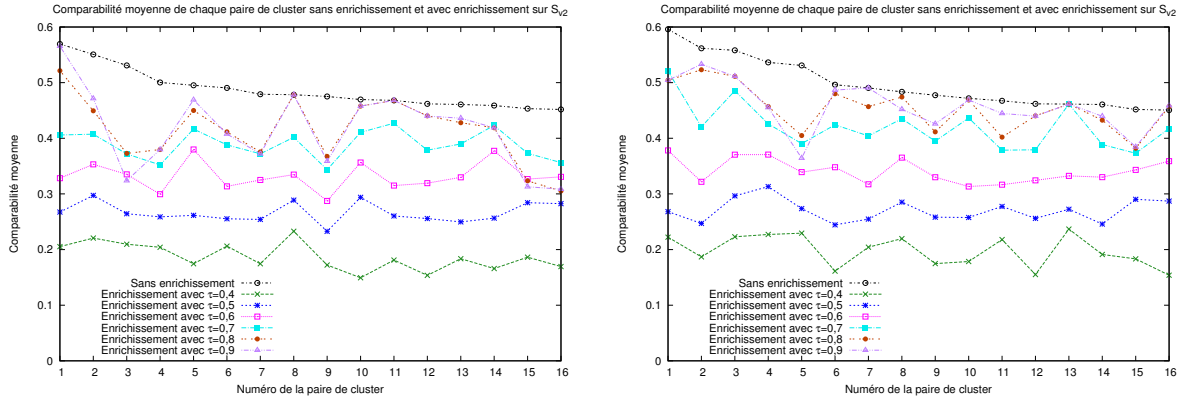

FIGURE 5.10 – Comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v2}, avec le **Tri séquentiel** à gauche, et avec le **Tri simultané** à droite.

sante que S_{v1}. Pour conclure sur ces deux Figures : 1) S_{v2} est plus adéquate que S_{v1}, et 2) un bon choix initial de τ (par exemple : 0,7 sur S_{v2}) peut rendre le corpus comparable plus dense en ajoutant de manière contrôlée des documents.

En comparant les figures à gauche et à droite, nous pouvons conclure que le **Tri simultané** est légèrement meilleur que le **Tri séquentiel**, car 1) il y a un peu plus de documents ajoutés pour une configuration donnée ; 2) lorsque τ devient très grand ($\tau \geq 0,8$) sur S_{v2}, les valeurs moyennes de la comparabilité sont un peu plus proches des valeurs de comparabilité des paires originales de médoides alignés (car les médoides alignés que nous avons validés manuellement sont ceux qui ont les valeurs de comparabilité les plus élevés). Par contre, en global, ces deux tris sont très comparables car le nombre de documents communs à l'issue des tris respectifs est

très élevé (934 documents communs sur 1000 en langue française et 911 documents communs sur 1000 en langue anglaise). Il en est de même pour le nombre de paires de clusters alignés après la validation manuelle, et les thèmes sont très voisins. Nous pouvons donc choisir l'un ou l'autre des tris.

5.4.2 Expérimentations complémentaires

Il y a aussi un point qui nous intéresse pour évaluer la solidité de notre approche : que se passe-t-il si nous prenons un *échantillon* autre que C_1 ? Puisque nous avons vérifié précédemment qu'avec le **Tri séquentiel** est très comparable avec le **Tri simultané**, et puisque nous considérons le **Tri du pire des cas** (pour le pire cas, il faut choisir inévitablement le **Tri séquentiel** car il ne prend pas en compte la relation entre les lignes et les colonnes), nous utilisons ici simplement le **Tri séquentiel**. Nous prenons donc un *échantillon* des 1000 **dernières** lignes et colonnes qui correspondent aux documents ayant les plus faibles degrés dans le graphe de comparabilité, afin de vérifier si notre approche peut encore extraire des médoides bien alignés.

5.4.2.1 Détermination du nombre initial de clusters K_0 en exploitant les similarités intra et inter clusters moyennes δ_{intra} et δ_{inter}

Nous déterminons ici un K_0 initial en analysant les variations des similarités intra et inter avec les valeurs des intra et inter clusters moyennes δ_{intra} et δ_{inter} obtenues sur la base de clustering k-médoides, lorsque k varie (ETAPE-3).

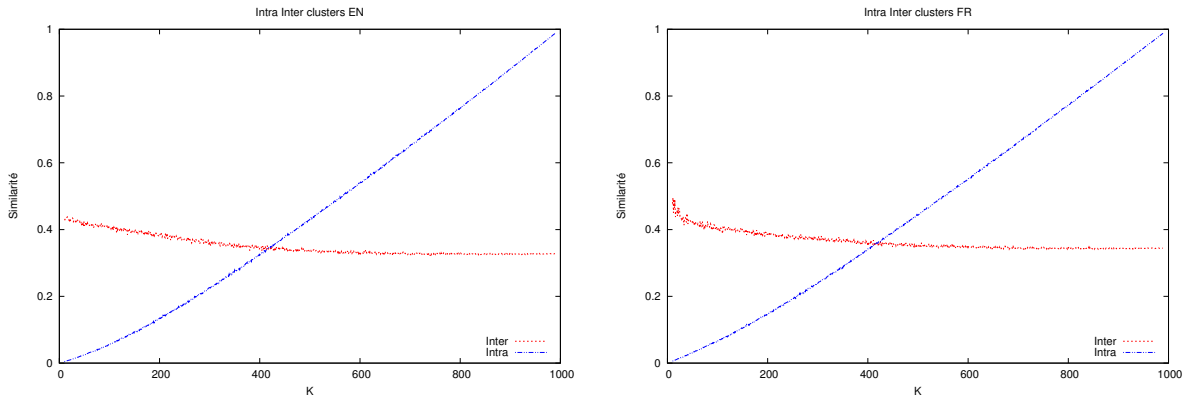

FIGURE 5.11 – Détermination du nombre initial de clusters K_0 en exploitant les similarités intra et inter clusters moyennes δ_{intra} et δ_{inter} dans le clustering k-médoides

Dans la Figure 5.11, pour les deux langues, nous observons une intersection pour le voisin de 400. Nous avons considéré que $K_0 = 400$ constitue ici un bon compromis.

5.4.2.2 Détermination du seuil de comparabilité φ

Nous déterminons le seuil de comparabilité φ en fonction du nombre de clusters conservés, du nombre de documents conservés et du degré du graphe bipartite des clusters alignés avec les seuils de comparabilité différents choisis (ETAPE-4).

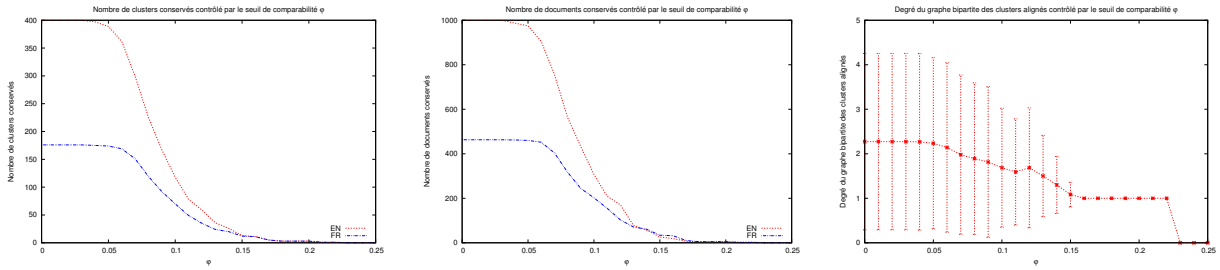

FIGURE 5.12 – Détermination du seuil de comparabilité φ en fonction du nombre de clusters conservés, du nombre de documents conservés et du degré du graphe bipartite des clusters alignés

Dans la Figure 5.12, nous essayons de déterminer une valeur φ de telle sorte que les clusters conservés demeurent en nombre suffisant, contiennent un nombre de documents également suffisant et que le graphe de comparabilité des médoides tende vers une relation 1-vers-1, i.e que le degré du graphe tende vers 1. Nous avons constaté et vérifié que lorsque φ est voisin de 0,157, toutes les trois valeurs (le nombre de clusters conservés, le nombre de documents conservés et le degré du graphe bipartite des clusters alignés) sont stables. Nous fixons donc le seuil de comparabilité φ à 0,157. Ce seuil est utilisé pour aligner automatiquement les clusters de langues différentes.

5.4.2.3 Paires de clusters alignés

Nous présentons en Figure 5.13 un exemple d'une paire de clusters alignés (chaque cluster, étant représenté par son médoide). Celle paire a obtenu la plus grande valeur de comparabilité après l'alignement des clusters comparables à l'issue de 6 exécutions des k-médoides avec le seuil de rejet de comparabilité φ = 0,157 (ETAPE-4). Les paires de clusters sont manuellement vérifiés (ETAPE-5).

Dans la Figure 5.13, nous observons que ces deux clusters ont été bien alignés car les deux médoides sont très comparables. Nous proposons en annexe les dix premiers clusters alignés triés par les valeurs de comparabilité décroissantes.

Dans la Figure 5.14, nous essayons de vérifier manuellement le nombre de clusters ajoutés et le nombre de clusters communs en fonction des itérations de k-médoides (ETAPE-5). Le nombre de clusters communs à tentance à augmenter et le nombre de clusters ajoutés à tentance à augmenter au début et à diminuer plus tard. Nous croyons donc que les clusters à ajouter

Valeur de comparabilité entre les deux clusters alignés : 0,2514	
<DOC> <DOCID>6b4e8896232d8455cd2630d1765f99e8</DOCID> <PUBDATE>Wed Feb 20 15:59:58 CET 2013</PUBDATE> <CURDATE>Wed Feb 20 20:41:27 CET 2013</CURDATE> <FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI> <ITEMURI>http://www.bbc.co.uk/news/world-asia-21519560#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Lost Antarctic penguin found in NZ</TITLE> <DESC>A royal penguin is being cared for at a Wellington Zoo after washing up on the coast of New Zealand, 2,000km (1,240 miles) from its Antarctic home.</DESC> <TXT>Lost Antarctic Royal penguin found in New Zealand Vets said the penguin could make its way home if it recovered Lost penguin Happy Feet 'missing' A royal penguin is being cared for at a New Zealand zoo after being found stranded on a beach 2,000km (1,240 miles) from home its Antarctic. The young male bird, which was dehydrated and starving, is thought to be only the fourth royal penguin to wash up there in more than a century. He is believed to have come from a breeding colony in the sub-Antarctic Macquarie Island. Vets said the bird, dubbed Happy Feet Jr, may have been drifting for a year. Lisa Argilla, a vet at Wellington Zoo, said the penguin had possibly struggled to find enough food or had had problems hunting and had come ashore as he needed to go through his seasonal moulting. He was found on Tora beach, on the coast to the south of Wellington, on Sunday. "It's very weak, doesn't want to stand. It's making very small progress every day but it's still in critical condition," Ms Argilla told the TVNZ channel. She told AFP his kidneys were not functioning properly, adding: "Hopefully we can reverse that, feed him up and bring him back to good health but it's touch and go at the moment." If he recovered, she said, he would be released to make his way home. "They're amazing at navigation so that shouldn't be a problem for him," she said. Last year, an emperor penguin, the original Happy Feet, made headlines when he appeared on New Zealand's shores. He had surgery to remove 3kg (6.6lb) of sand from his stomach, which he is thought to have eaten thinking it was snow, before being released with a tracking device. But he disappeared soon after and was believed to have been eaten. More on This Story</TXT> </DOC>	<DOC> <DOCID>e35912719c9d6731b04887102f207e12</DOCID> <PUBDATE>Fri Feb 22 01:56:03 CET 2013</PUBDATE> <CURDATE>Fri Feb 22 07:14:05 CET 2013</CURDATE> <FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI> <ITEMURI>http://www.romandie.com/news/n.asp?n=Triste_fin_pour_le_manchot_Happy_Feet_Junior4022022013056.asp</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Triste fin pour le manchot Happy Feet Junior</TITLE> <DESC>Un manchot surnommé Happy Feet junior qui avait été rejeté sur une côte néo-zélandaise à 2000 km de sa région d'origine est mort malgré tous les efforts mis en oeuvre pour le sauver, a annoncé vendredi le Zoo de Wellington. Selon sa...</DESC> <TXT>Tweet Triste fin pour le manchot Happy Feet Junior Un manchot surnommé Happy Feet junior qui avait été rejeté sur une côte néo-zélandaise à 2000 km de sa région d'origine est mort malgré tous les efforts mis en oeuvre pour le sauver, a annoncé vendredi le Zoo de Wellington. Selon sa vétérinaire en chef Lisa Argilla, il a succombé à la malnutrition et à un problème rénal. Une équipe de vétérinaires a passé cinq jours à soigner l'oiseau, un jeune manchot royal, qui a dérivé très loin de sa colonie d'origine, l'île subantarctique de Macquarie, après avoir passé environ 12 mois en mer. "A son arrivée, le manchot pesait près de trois kilos de moins que la normale, il n'avait absolument aucune réserve et de ce fait, nous supposons que cela a provoqué chez lui de multiples défaillances viscérales, après l'insuffisance rénale diagnostiquée à son arrivée", a-t-elle dit. Spécialité difficile "La médecine vétérinaire pour les animaux sauvages est une spécialité très difficile, et bien que nous ayions fait le mieux que nous pouvions, malheureusement le manchot n'a pas survécu", a-t-elle ajouté. La découverte de l'oiseau a rappelé l'histoire de Happy Feet, un manchot empereur qui s'était échoué sur une plage près de Wellington en juin 2011 et qui était devenu une attraction mondiale pendant les huit semaines où il avait été soigné au zoo. Le manchot avait finalement été relâché par un navire de recherche néo-zélandais dans l'océan Antarctique, après avoir reçu les visites de personnalités telles que l'acteur et réalisateur britannique Stephen Fry et les meilleurs voeux du Premier ministre néo-zélandais John Key. Cependant, le boîtier de localisation GPS qui avait été accroché à l'oiseau avait cessé de transmettre après quelques jours, faisant craindre que le manchot ait été dévoré par un requin. (ats / 22.02.2013 01h56)</TXT> </DOC>

FIGURE 5.13 – Alignement des deux clusters (médoides) ayant la comparabilité la plus élevée

potentiellement sont de moins en moins nombreux et après quelques itérations, le nombre des clusters alignés devient stable.

5.4.2.4 Nombre de documents conservés avec différents seuils d'ajout τ

Nous étudions ici le nombre de documents conservés avec différents seuils d'ajout τ pour les deux mesures d'ajouts S_{v1} et S_{v2} (ETAPE-6).

Premièrement, nous illustrons le nombre de documents conservés avec différentes valeurs de S_{v1}, c'est-à-dire la somme de 50% de la valeur de similarité et 50% de la valeur de comparabilité.

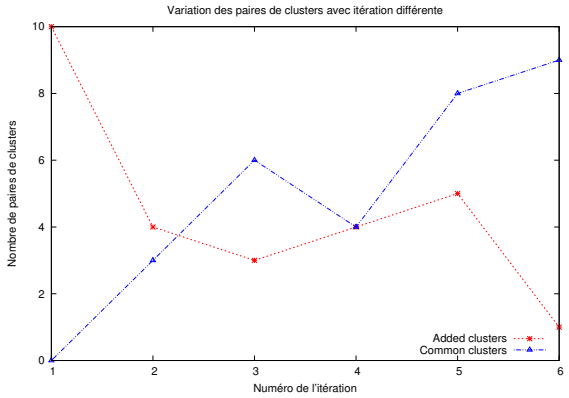

FIGURE 5.14 – Nombre de clusters ajoutés et nombre de clusters communs par rapport à chaque itération de k-médoides

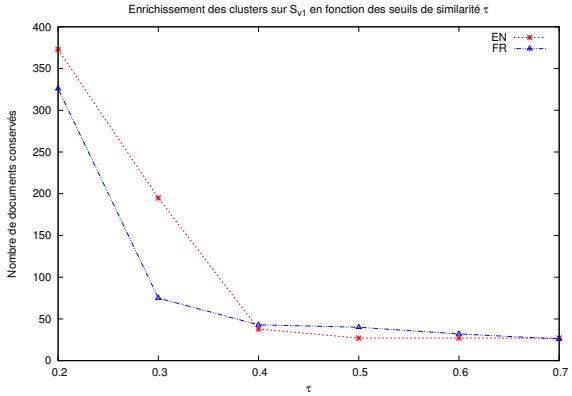

FIGURE 5.15 – Nombre de documents conservés avec différentes valeurs d'ajout sur S_{v1}

Deuxièmement, nous illustrons le nombre de documents conservés avec différentes valeurs de S_{v2}, c'est-à-dire la somme de 50% de la valeur de similarité native et 50% de la valeur de la similarité induite par la comparabilité.

Dans les Figures 5.15 et 5.16, nous montrons le nombre de documents ajoutés en fonction du seuil d'ajout τ (sur S_{v1} ou sur S_{v2}). Les valeurs de τ pour les deux mesures ne représentent pas le même niveau de comparabilité. Pour S_{v1}, la valeur de τ qui diminue considérablement le nombre de documents ajoutés est plus bas que pour S_{v2}. C'est-à-dire que pour S_{v1} et $\tau = 0,4$, on a presque le même niveau de comparabilité qu'en choisissant pour $\tau = 0,6$ et S_{v2}. Cependant, d'après nos expérimentations et conformément à l'intuition, S_{v2} est plus recommandée que S_{v1} pour "customiser" chaque paire de clusters alignés en fonction des besoins.

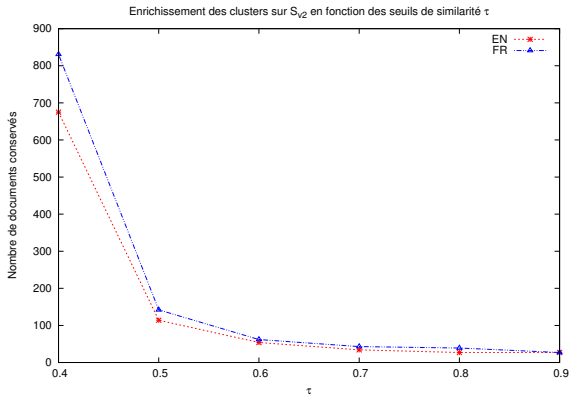

FIGURE 5.16 – Nombre de documents conservés avec différentes valeurs d'ajout sur S_{v2}

5.4.2.5 Comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1} ou S_{v2}

Nous étudions ici la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1} ou S_{v2} (ETAPE-6).

Premièrement, nous illustrons la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1}.

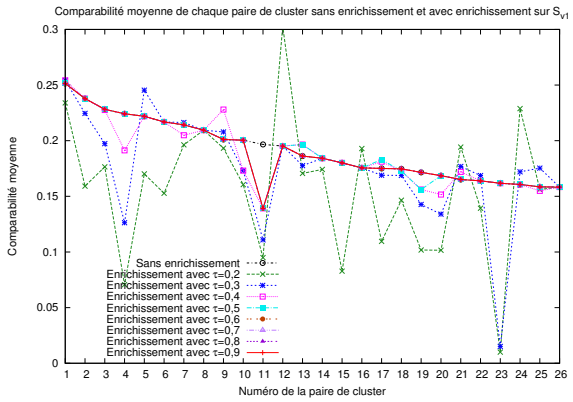

FIGURE 5.17 – Comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1}

Deuxièmement, nous illustrons la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v2}.

Dans les Figures 5.17 et 5.18, nous montrons la variation de la comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v1} et sur S_{v2}. Dans la Figure 5.17, lorsque $\tau \geq 0,3$, la comparabilité moyenne de

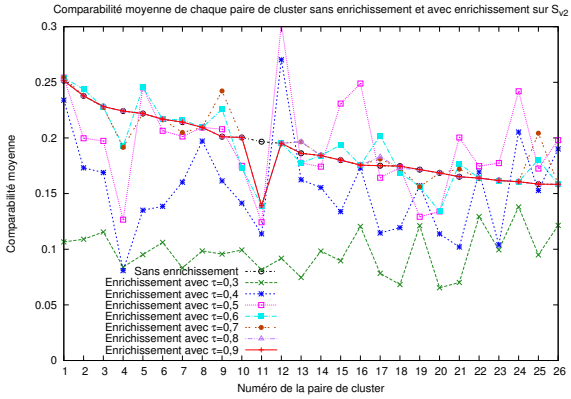

FIGURE 5.18 – Comparabilité moyenne de chaque paire de clusters sans enrichissement et avec enrichissement contrôlé par le seuil d'ajout τ sur S_{v2}

chaque paire de clusters augmente en général, mais lorsque $\tau \geq 0,6$, il n'a plus de documents ajoutés. Cela montre que sur S_{v1}, un τ pertinent peut être initialement défini autour de 0,3 ou 0,4. Par contre, dans la Figure 5.18, lorsque $\tau \geq 0,5$, la comparabilité moyenne de chaque paire de clusters augmente en général, mais lorsque $\tau \geq 0,8$, il y a encore des documents ajoutés, certes de moins en moins. Cela montre que sur S_{v2}, un τ pertinent peut être initialement défini autour de 0,5 ou 0,6. En comparant les deux figures, nous constatons de nouveau que la fusion des similarités *natives* et les similarités *induites* par la comparabilité a un impact positif dans l'enrichissement car lorsque $\tau \geq 0,8$ (même $\tau = 0,9$), il y a encore des documents ajoutés. Et puis, intuitivement, encore une fois, S_{v2} est plus adaptée que S_{v1}. Pour conclure sur ces deux Figures : 1) S_{v2} est plus adéquate que S_{v1}, et 2) un bon choix initial de τ (par exemple : 0,5 ou 0,6 sur S_{v2}) peut rendre le corpus comparable plus dense en ajoutant des documents.

Nous pouvons donc conclure ici que bien que nous prenions le pire des cas (les documents ayant les plus faibles degrés dans le graphe de comparabilité), nous pouvons toujours trouver des clusters (médoides) bien alignés. Alors, si nous avons besoin de plus de clusters, nous pouvons prendre d'autres échantillons pour élargir la base des corpus comparables.

5.5 Conclusion

Nous avons proposé une approche semi-supervisée pour la construction de corpus comparables en forte cohésion thématique. Cette approche vise à produire des clusters thématiques alignés plus ou moins comparables en utilisant le clustering k-médoides, la mesure de comparabilité C_{VA_2}, *tf* et une valeur pour le paramètre de mélange des similarités *natives* et *induites* α=0,5. Cette approche est basée sur 6 étapes et nécessite de fixer 7 paramètres importants comme le nombre initial de cluster K_0 pour k-médoides, le seuil de rejet de comparabilité φ lors de l'alignement les clusters comparables, le seuil d'ajout τ lors de l'enrichissement, etc.

Nous avons testé sur un cas réel l'utilisabilité de notre approche. Nous avons pu étudier certains effets de ces paramètres comme l'impact du nombre de clusters ajoutés et le nombre de clusters communs par rapport à chaque itération de k-médoides, et aussi l'impact du nombre de documents conservés avec différentes valeurs d'ajout τ pour les deux mesures d'ajout S_{v1} et S_{v2}. Nous avons obtenu 16 paires de clusters bien alignés après 4 itérations de k-médoides sur $\mathcal{C}_1$. Nous avons également obtenu 26 paires de clusters bien alignés après 6 itérations de k-médoides sur un *échantillon* construit sur la base du pire des cas. Nous avons vérifié que les deux alternatives de tri proposées sont très comparables. Nous avons également vérifié que S_{v2} est plus adéquate que S_{v1}, et qu'un bon choix initial de τ (par exemple : 0,6 ou 0,7 sur S_{v2}) peut rendre le corpus comparable plus dense en ajoutant des documents.

Un grand avantage de cette approche est que nous pouvons obtenir des clusters bien alignés pour construire des corpus comparables de "bonne qualité". Un paramètre d'enrichissement τ a été proposé pour personnaliser les corpus comparables en fonction des besoins des utilisateurs. Cette approche nous permet également de construire des corpus comparables de grand volume car nous pouvons les enrichir non seulement à partir du corpus bilingue "brut", mais aussi à partir de tout corpus thématique disponible.

Cependant, comme nous avons intégré une étape de vérification manuelle pour garantir la qualité de l'alignement des clusters, cette approche n'est pas tout-à-fait automatique. En matière de qualité et de quantité, nous avons fait un compromis. En plus, les paramètres ne sont pas facilement exploitables car ils peuvent varier un peu selon les différents corpus. Enfin, comme le clustering k-médoides dépend des conditions initiales, il faut que nous l'executions un certain nombre de fois pour extraire plus de clusters, ce qui complique également notre tâche.

Sixième partie

CONCLUSIONS

6

Conclusions et perspectives

Sommaire

6.1 Introduction

Dans cette thèse, nous avons principalement ciblé l'amélioration de la qualité des corpus comparables lors de leur construction. La problématique est justifiée par le manque de corpus parallèles et l'attrait des corpus comparables vis-à-vis des corpus parallèles. Les corpus parallèles sont coûteux à développer et leur transposition d'un domaine de spécialité à l'autre est souvent difficile. Par rapport aux corpus parallèles, les corpus comparables peuvent résoudre en partie ces problèmes car ils peuvent être engendrés à partir de ressources très larges et riches (comme le WEB par exemple), et sans trop de limitation sur des textes originaux et sur des langues. Les corpus comparables sont indispensables dans plusieurs domaines [158] : l'extraction des terminologies ou des lexiques bilingues [37], [94], [55], [79], [52], [152], [136], la fouille de textes multilingues, la traduction automatique ou assistée, l'apprentissage des langues [141], etc. Par ailleurs, dans [134], [87], [107], [78], les auteurs ont vérifié que la qualité d'alignement des corpus comparables est plus importante que leur taille. Pour ces raisons, la constitution des corpus comparables thématique de bonne qualité devient un enjeu essentiel. Cette thèse est donc induite par cette problématique et tente de résoudre le mieux possible les questions posées lors de la construction de ces corpus.

Afin de développer une assistance à la construction de corpus comparables de bonne qualité, nous avons suivi une démarche en trois phases :

1. Nous avons proposé des nouvelles mesures de comparabilité pour estimer le degré de comparabilité entre documents de langues différentes et re-développé un cadre général d'évaluation de ces mesures de comparabilité.
2. Nous avons développé une nouvelle approche de clustering et de classification adaptée aux corpus comparables bilingues pour améliorer l'extraction et l'alignement des clusters comparables.
3. Nous avons exploité ces deux premières contributions pour proposer une méthode d'assistance à la construction de corpus comparables de bonne qualité. Cette méthode semi-supervisée est paramétrable et ajustable en fonction des différents niveaux de comparabilité requis et de la volumétrie des corpus souhaitée.

Dans les sections suivantes, nous discutons brièvement ces trois contributions principales de cette thèse et les perspectives envisageables.

6.2 Sommaire des contributions

6.2.1 Mesures de comparabilité proposées

Dans le chapitre 3, nous avons présenté et évalué deux nouvelles variantes de mesure de comparabilité en intégrant la fréquence des mots et leurs nombres de traductions présentes dans le dictionnaire bilingue. Nous les avons comparées avec la mesure de comparabilité de référence [77] et constaté que nos deux variantes se comportent comparativement plutôt bien à un niveau de comparablité moyen (non parallèle), ce qui répond bien notamment lorsque l'on cherche à construire des corpus comparables à partir de ressources "bruitées" et hétérogènes comme le WEB. Par ailleurs, les deux variantes fonctionnent bien lorsque les documents ne sont pas très grands (par exemple, les documents de moins de 1000 phrases, ce qui est presque toujours le cas en réalité sur le WEB).

6.2.2 SCF-clustering, SCF-classification et alignement des clusters comparables

Nous avons présenté dans le chapitre 4, une nouvelle approche de clustering et de classification qui fusionne les similarités *natives* et les similarités *induites* par une mesure quantitative de comparabilité pour effectuer des tâches de clustering multilingue (en utilisant K-médoides et HAC), de classification multilingue (en utilisant k-PPV) et d'alignement des clusters comparables. Nous avons testé ce modèle de mélange sur deux corpus que nous avons collectés sur le WEB, dont un corpus de 7 classes issu de 23 Flux RSS et un autre corpus de 21 classes issu de Wikipédia. Nos expérimentations ont montré que cette nouvelle approche améliore significativement les performances de clustering et de classification, ainsi que la qualité d'alignement

des clusters comparables. Cela nous a permis de proposer une méthode semi-supervisée de construction des corpus comparables.

6.2.3 Généralisation pour la constitution des corpus comparables

Nous avons enfin présenté la généralisation des mesures proposées précédemment pour construire des corpus comparables avec une nouvelle approche dans le chapitre 5.

Cette approche est basée sur 6 étapes.

— **ETAPE-1 : Calcul et construction des matrices de comparabilité pour les documents anglais et français du corpus *brut* initial**, C_0.

— **ETAPE-2 : Filtrage du corpus initial C_0 et production d'un corpus bilingue plus dense, C_1 au sens de la comparabilité**.

— **ETAPE-3 : Détermination du nombre K_0 de clusters initiaux**.

— **ETAPE-4 : Filtrage des paires de clusters (Anglais-Français) fortement comparables**.

— **ETAPE-5 : Vérification manuelle des clusters alignés**

— **ETAPE-6 : Enrichissement du corpus**

L'ETAPE-5 intègre une procédure de vérification manuelle, cette approche est donc semi-supervisée. Bien que la procédure de vérification ait besoin d'une intervention manuelle (c'est un compromis que nous avons fait pour pouvoir obtenir des clusters comparables de bonne qualité), celle-ci reste gérable et même efficace car le nombre de clusters n'est pas très grand en général. En plus, grâce à cette procédure, la qualité d'alignement des clusters comparables est bien validée et on peut ainsi "customiser" les corpus à un certain niveau de comparabilité en fonction des besoins des utilisateurs.

6.3 Conclusions générales

Nous pouvons donc proposer quelques éléments de réponse aux questions posées en introduction pour étayer nos motivations comme suit :

— Le cadre d'évaluation proposé par Li et Gaussier, indépendante du contexte applicatif constitue-t-elle une opportunité pour hiérarchiser les mesures de comparabilité ?
Nous considérons que ce cadre est relativement bien adapté mais celui-ci doit être complété en fonction de la tâche considérée (par exemple, par des critères de qualité de

clustering, de classification ou d'alignement de documents).

— Jusqu'où peut-on aller en matière d'assistance automatisée pour la construction des corpus comparables thématiques ?
Nous avons exploité nos deux premières contributions pour proposer une méthodologie semi-supervisée d'assistance à la construction de corpus comparables de bonne qualité. Cette méthode semi-supervisée est paramétrable et ajustable en fonction des différents niveaux de comparabilité requis et de la volumétrie des corpus souhaitée. La méthodologie proposée passe bien à l'échelle (complexité $O(n^3)$ où n est le nombre de médoides alignés retenus). La vérification manuelle impliquée dans notre approche est gérable puisque l'on s'intéresse principalement à l'alignement des paires de médoides des clusters extraits.

— Peut-on optimiser à la fois sur la qualité d'alignement des documents (ou clusters) et également sur la volumétrie dans la construction des corpus comparables thématiques ?
Nous avons proposé une nouvelle approche de co-clustering et de co-classification, qui non seulement permet d'aligner des clusters bilingues comparables de qualité. Ces cluters peuvent également être enrichie en exploitant tout autre corpus thématique disponible. La méthodologie proposée permet un contrôle assez fin pour déterminer le meilleur compromis entre qualité d'alignement et volumétrie des corpus comparables thématiques.

6.4 Perspectives

Nous proposons deux types de perspectives : l'amélioration de notre approche et l'extension de notre approche.

Au niveau de l'amélioration de notre approche :

1. Il est possible de réduire la période temporelle pour filtrer les documents et soulager le calcul des matrices de comparabilité et de similarité (*natives* et *induites*) et ensuite intégrer des caractéristiques comme TNC (titre et contenu), LIU (unité indépendante linguistique) et MTD (distribution des termes monolingue) comme proposées dans [149].
2. Nous pouvons combiner les différentes méthodes de détermination du nombre de clusters K pour obtenir des clusters encore plus significatif.
3. Nous pouvons intégrer les mesures de désambiguïsation comme dans [129], [73] : soit pour créer une nouvelle mesure de comparabilité avec encore une meilleure qualité, soit pour mettre une procédure de prétraitement pour filtrer les mots et les documents.
4. Nous pouvons également modifier la dernière étape de l'approche de généralisation : il faut calculer les variations du nombre de documents en ajoutant les documents pour chaque paire de clusters en variant le seuil d'ajout (valeur de similarités combinées entre un document et les médoides). Dans nos expérimentations, pour effectuer l'ajout

des documents, nous avons calculé la valeur des similarités combinées (similarités *natives* et similarités *induites* par comparabilité) et si la valeur est supérieure au seuil de similarité, le document est ajouté dans la paire de clusters comparables. Nous avons opté pour cette méthode afin de diminuer le temps de calcul. Mais si le temps n'est pas un facteur limitant ou si la machine est assez puissante, nous pouvons proposer une autre solution. Au lieu de calculer simplement la similarité combinée entre un document et les médoides, nous les regroupons dans les matrices initiales : nous ajoutons ce document comme une nouvelle ligne de similarité à la fin de la matrice de similarité native en calculant toutes les valeurs de similarité entre ces documents et tous les autres documents dans cette matrice et nous ajoutons ce document comme une nouvelle ligne de comparabilité à la fin de la matrice de comparabilité en calculant les valeurs de comparabilité entre ce document et tous les autres documents dans cette matrice. Après cela, nous combinons la matrice de similarité native et la matrice transformée de la matrice de comparabilité et enfin nous affectons ce document au cluster avec lequel le document traité est le plus similaire.

5. Comme les mesures de comparabilité dépendent de la couverture du dictionnaire bilingue, nous pouvons utiliser les corpus comparables que nous avons construits (éventuellement les autres corpus comparables déjà existants) pour extraire les lexiques bilingues ou les terminologies thématiques. Grâce à cela, nous pouvons enrichir notre dictionnaire bilingue. Ensuite, nous pouvons utiliser le dictionnaire bilingue enrichi pour reconstruire des corpus comparables de meilleure qualité. C'est un cercle vertueux d'amélioration, comme montré en Figure 6.1.

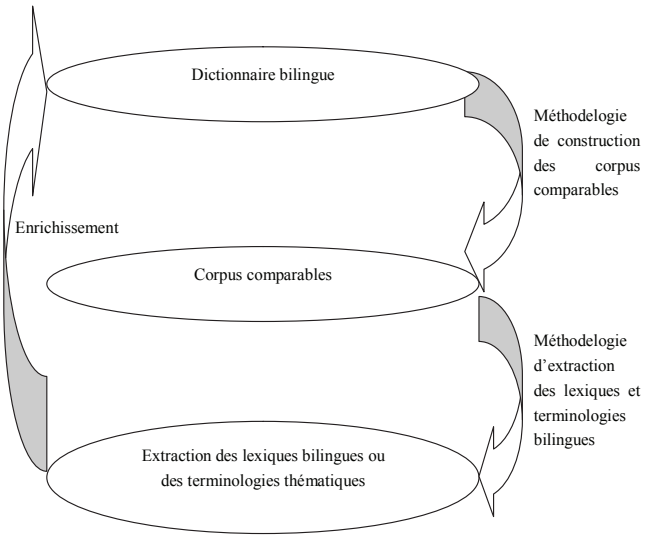

FIGURE 6.1 – Cercle vertueux d'amélioration itérative par raffinement du dictionnaire bilingue

Au niveau de l'extension de notre approche :

1. Nous pouvons étendre cette approche pour la construction des corpus comparables dans d'autres paires de langues comme anglais/chinois, français/chinois, etc. Ou bien nous pouvons même étendre cette approche pour la constitution des corpus comparables multilingues comme anglais/chinois/français, etc.
2. Par ailleurs, si dans certaines langues, les dictionnaires bilingues ne sont pas disponibles, soit nous pouvons créer ou utiliser les mesures de création automatique des entrées lexicales comme dans [132], soit nous pouvons exploiter par une langue pivot comme l'anglais qui est sans doute la plus riche : nous pouvons traduire tous les textes des deux langues différentes en anglais et grâce à cela, nous pouvons enfin construire les corpus comparables bilingues.

Bibliographie

[1] Tlfi, 1960.

[2] M. N. Murty A. K. Jain and P. J. Flynn. Data clustering : a review. *ACM Comput. Surv.*, 31(3) :264â323, 1999.

[3] Rakesh Agrawal, Christos Faloutsos, and Arun N. Swami. Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, FODO '93, pages 69–84, London, UK, UK, 1993. Springer-Verlag.

[4] Karin AIJMER and Bengt ALTENBERG. *Advances in Corpus Linguistics*. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Amsterdam/New York, NY, 2004, IX, 419 pp, 2004.

[5] Massih-Reza Amini and Cyril Goutte. A co-classification approach to learning from multilingual corpora. *Mach. Learn.*, 79(1-2) :105–121, May 2010.

[6] Kirk Baker. Singular value decomposition tutorial. *The Ohio State University*, 2005.

[7] Mona Baker et al. Corpus linguistics and translation studies : Implications and applications. *Text and technology : in honour of John Sinclair*, 233 :250, 1993.

[8] Simona Balbi and Michelangelo Misuraca. Rotated canonical correlation analysis for multilingual corpora. *JADT'06 : Actes Des 8es Journées Internationales D'analyse Statistique Des Données Textuelles, Besançon, 19-21 Avril 2006*, 1 :99, 2006.

[9] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. *SIGIR Forum*, 31(SI) :84–91, July 1997.

[10] Sergio Barrachina and Juan Miguel Vilar. Bilingual clustering using monolingual algorithms. In *8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1999)*, pages 77–87, 1999.

[11] Regina Barzilay and Noemie Elhadad. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 25–32, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[12] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[13] Kai bo Duan and S. Sathiya Keerthi. Which is the best multiclass svm method ? an empirical study. In *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, pages 278–285, 2005.

[14] Lynne Bowker and Jennifer Pearson. *Working with Specialized Language - A practical guide to using corpora*. London : Routledge, 2002.

[15] Martin Braschler and Peter Scäuble. Multilingual information retrieval based on document alignment techniques. In *Research and Advanced Technology for Digital Libraries*, volume 1513 of *Lecture Notes in Computer Science*, pages 183–197. Springer Berlin Heidelberg, 1998.

[16] Martin Braschler and Peter Schäuble. Experiments with the eurospider retrieval system for clef 2000. In *Proceedings of CLEF 2000*, pages 140–148. Springer-Verlag, 2001.

[17] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2) :79–85, June 1990.

[18] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4) :467–479, 1992.

[19] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1) :1–27, 1974.

[20] Fazli Can and Esen A Ozkarahan. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Transactions on Database Systems (TODS)*, 15(4) :483–517, 1990.

[21] František Čermák and Alexandr Rosen. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3) :411–427, 2012.

[22] Hsin-Hsi Chen and Chuan-Jie Lin. A multilingual news summarizer. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 159–165, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[23] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

[24] Jorge Civera, Elsa Cubel, and Enrique Vidal. Bilingual text classification. In *Pattern Recognition and Image Analysis*, pages 265–273. Springer, 2007.

[25] Jorge Civera and Alfons Juan-císcar. Bilingual text classification using the ibm 1 translation model.

[26] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1) :21–27, 1967.

[27] Ido Dagan, Yael Karov, and Dan Roth. Mistake-driven learning in text categorization. In *IN EMNLP-97, THE SECOND CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING*, pages 55–63, 1997.

[28] D. L. Davies and D. W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, PAMI-1(2) :224–227, 1979.

[29] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4) :364â366, 1977.

[30] H. Déjean and E. Gaussier. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Numéro spécial, corpus alignés :1–22, 2002.

[31] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[32] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM.

[33] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, 29(2-3) :103–130, November 1997.

[34] Dubreil E. *La dimention argumentative des collocations textuelles en corpus électronique spécialisé au domaine du TAL(N)*. PhD thesis, Université de Nantes, 2006.

[35] Kanti Mardia et al. Multivariate analysis. *Academic Press*, 1979.

[36] E. Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21 :768–780, 1965.

[37] Pascale Fung. A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Machine Translation and the Information Soup*, volume 1529 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg, 1998.

[38] Pascale Fung and Kathleen Mckeown. Finding terminology translations from non-parallel corpora, 1997.

[39] Pascale Fung and Lo Yuen Yee. 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, 1998. Association for Computational Linguistics.

[40] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 1993.

[41] Jahann Gamper. Encoding a parallel corpus for automatic terminology extraction. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 275–276, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[42] E. Gaussier, J. m. Renders, I. Matveeva, C. Goutte, and H. Déjean. A geometric view on bilingual lexicon extraction from comparable corpora. In *In Proceedings of ACL-04*, pages 527–534, 2004.

[43] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268) :pp. 732–764, 1954.

[44] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. ii : Further discussion and references. *Journal of the American Statistical Association*, 54(285) :pp. 123–163, 1959.

[45] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications iii : Approximate sampling theory. *Journal of the American Statistical Association*, 58(302) :pp. 310–364, 1963.

[46] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications, iv : Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67(338) :pp. 415–421, 1972.

[47] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure : an efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, pages 73–84, New York, NY, USA, 1998. ACM.

[48] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock : A robust clustering algorithm for categorical attributes. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 0 :512, 1999.

[49] John A Hartigan. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337) :123–129, 1972.

[50] Erik Hatcher and Otis Gospodnetic. Lucene in action (in action series). In *Manning Publications Co., Greenwich, CT, USA*, 2004.

[51] B. Havert. *Des corpus représentatifs : de quoi, pour quoi, comment ?* Presse Universitaires de Perpignan, 2000.

[52] Amir Hazem, Emmanuel Morin, and Sebastian Peña Saldarriaga. Bilingual lexicon extraction from comparable corpora as metasearch. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, BUCC '11, pages 35–43, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[53] Mehrdad Honarkhah and Jef Caers. Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences*, 42(5) :487–517, 2010.

[54] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Trans. Neur. Netw.*, 13(2) :415–425, March 2002.

[55] Azniah Ismail and Suresh Manandhar. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 481–489, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[56] Véronis J. *Alignement de corpus multilingues*. Ingénierie des langues, edition hermès edition, 2000.

[57] M Kamber J Han. *Data Mining : Concepts and Techniques*. version 2, 2006.

[58] Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. From bilingual dictionaries to interlingual document representations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers - Volume 2*, HLT '11, pages 147–152, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[59] Jagadeesh Jagarlamudi, Raghavendra Udupa, Hal Daumé, III, and Abhijit Bhole. Improving bilingual projections via sparse covariance matrices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 930–940, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[60] Thorsten Joachims. Text categorization with support vector machines : Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK, 1998. Springer-Verlag.

[61] G. Karypis, Eui-Hong Han, and V. Kumar. Chameleon : hierarchical clustering using dynamic modeling. *Computer*, 32(8) :68–75, 1999.

[62] L. Kaufman and P.J. Rousseeuw. *Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1âNorm and Related Methods*. North-Holland, 1987.

[63] L. Kaufman and P.J. Rousseeuw. *Finding Groups in DataâAn Introduction to Cluster Analysis*. JohnWiley& Sons, Inc, 1990.

[64] Martin Kay and Martin Röscheisen. Text-translation alignment. *Comput. Linguist.*, 19(1) :121–142, March 1993.

[65] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2) :pp. 81–93, 1938.

[66] Kevin Knight and Yaser Al-Onaizan. Translation with finite-state devices. In *Machine translation and the information soup*, pages 421–437. Springer, 1998.

[67] P. Koehn. Europarl : A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand, 2005.

[68] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[69] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9) :1464–1480, 1990.

[70] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 170–178, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[71] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 289–297, New York, NY, USA, 1996. ACM.

[72] Lianhau Lee, Aiti Aw, Min Zhang, and Haizhou Li. Em-based hybrid model for bilingual terminology extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 639–646, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[73] Els Lefever and Veronique Hoste. Semeval-2010 task 3 : Cross-lingual word sense disambiguation, 2010.

[74] Els Lefever, Lieve Macken, and Veronique Hoste. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 496–504, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[75] Lawrence J Leftin. Newsblaster russian-english clustering performance analysis. *Computer Science Technical Report Series*, 2003.

[76] David Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. pages 298–306, 1996.

[77] B. Li and E. Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING*, pages 644–652, 2010.

[78] Bo Li. *Measuring and Improving Comparable Corpus Quality*. PhD thesis, Université de Grenoble, 2012.

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

[80] Xiaoyi Ma and Mark Y. Liberman. BITS : A Method for Bilingual Text Search over the Web. 1999.

[81] Lieve Macken, Els Lefever, and Veronique Hoste. Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 529–536, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[82] Daniel Marcu and William Wong. A phrase-based, joint probability model for statistical machine translation. In *In Proceedings of EMNLP*, pages 133–139, 2002.

[83] Stella Markantonatou, Sokratis Sofianopoulos, Vassiliki Spilioti, George Tambouratzis, Marina Vassiliou, and Olga Yannoutsou. Using patterns for machine translation (mt. In *In Proceedings of the European Association for Machine Translation*, pages 239–246, 2006.

[84] Pierre-Francois Marteau and Gildas Ménier. Similarités induites par mesure de comparabilité : signification et utilité pour le clustering et lâalignement de textes comparables. In *TALN*, page 515â522, 2013.

[85] Benoit Mathieu, Romaric Besançon, and Christian Fluhr. Multilingual document clusters discovery. In *RIAO*, pages 116–125. Citeseer, 2004.

[86] A.M. McEnery. Multilingual Corpora – Current Practice and Future Trends. In *13th ASLIB Machine Translation Conference*, pages 75–86, London, 1997.

[87] Paul McNamee, James Mayfield, and Charles Nicholas. Translation corpus source and size in bilingual retrieval. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, NAACL-Short '09, pages 25–28, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[88] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209 :415–446, 1909.

[89] Boris Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Publishers, 1996.

[90] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[91] T.M. MITCHELL. *Machine Learning. McGraw Hill, New York, NY*. McGraw Hill, New York, NY, 1996.

[92] Soto Montalvo, Raquel Martínez, Arantza Casillas, and Víctor Fresno. Bilingual news clustering using named entities and fuzzy similarity. In *Text, Speech and Dialogue*, pages 107–114. Springer, 2007.

[93] Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. Bilingual terminology mining - using brain, not brawn comparable corpora. In *ACL*, 2007.

[94] Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. Brains, not brawn : The use of smart comparable corpora in bilingual terminology mining. *ACM Trans. Speech Lang. Process.*, 7(1) :1 :1–1 :23, October 2008.

[95] Dragos Stefan Munteanu. *Exploiting comparable corpora*. PhD thesis, Los Angeles, CA, USA, 2006. AAI3257825.

[96] Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pages 265–272, 2004.

[97] Dragos Stefan Munteanu and Daniel Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, July 2006. Association for Computational Linguistics.

[98] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the international NATO symposium on Artificial and human intelligence*, pages 173–180, New York, NY, USA, 1984. Elsevier North-Holland, Inc.

[99] Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 67–73, New York, NY, USA, 1997. ACM.

[100] D. Oard and Diekema A. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, Vol. 33 :223–256, 1998.

[101] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1) :19–51, March 2003.

[102] Pablo Gamallo Otero and Isaac González López. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 21–25, 2009.

[103] Pablo Gamallo Otero, Issac González López, SL Cilenis, and Santiago de Compostela. Measuring comparability of multilingual corpora extracted from wikipedia. *on Iberian Cross-Language Natural Language Processings Tasks (ICL 2011)*, page 8, 2011.

[104] Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1) :25–45, 1920.

[105] David Picó and Francisco Casacuberta. Some statistical-estimation methods for stochastic finite-state transducers. *Machine Learning*, 44(1-2) :121–141, 2001.

[106] Emmanuel Ep Prochasson. *Alignement multilingue en corpus comparables spécialisés*. PhD thesis, Université de Nantes, 2009.

[107] Zahra Rahimi and Azadeh Shakery. Topic based creation of a persian-english comparable corpus. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology*, AIRS'11, pages 458–469, Berlin, Heidelberg, 2011. Springer-Verlag.

[108] Philip Resnik. Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 527–534, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[109] Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3) :349–380, September 2003.

[110] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.

[111] Mathias Rossignol and Pascale Sébillot. Extraction statistique sur corpus de classes de mots-clés thématiques. *TAL. Traitement automatique des langues*, 44(3) :217–246, 2003.

[112] Peter J. Rousseeuw. Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0) :53 – 65, 1987.

[113] X. Saralegi, I. San Vicente, and A. Gurrutxaga. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *6th International Conference on Language Resources and Evaluations - Building and using Comparable Corpora workshop*, 2008.

[114] Xabier Saralegi and Inaki Alegria. Similitud entre documentos multilingues de caracter cientifico-tecnico en un entorno web. In *Proceedings of the SEPLN*, pages 71–78, 2007.

[115] Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[116] Helmut Schmid. TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, 2009.

[117] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[118] Sam Scott and Stan Matwin. Feature engineering for text classification. In *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pages 379–388. Morgan Kaufmann Publishers, 1999.

[119] Fabrizio Sebastiani. A tutorial on automated text categorisation. In *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pages 7–35, Buenos Aires, AR*, pages 7–35, 1999.

[120] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1) :1–47, March 2002.

[121] Serge Sharoff, Bogdan Babych, and Anthony Hartley. Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 739–746, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[122] Yirong Shen and Jing Jiang. Improving the performance of naive bayes for text classification, cs224n, 2003.

[123] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 58–65, New York, NY, USA, 1996. ACM.

[124] R. Sibson. Slink : An optimally efficient algorithm for the single link cluster method. *The Computer Journal*, 16(1) :30–34, 1973.

[125] J. Sinclair. Preliminary recommendations on corpus typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), 1996.

[126] R. Sokal and C. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38(22) :1409–1438, 1958.

[127] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 :11–21, 1972.

[128] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1) :72–101, 1904.

[129] Lucia Specia, Mark Stevenson, Maria Das Graçasvolpe Nunes, Gabriela Castelo, and Branco Ribeiro. Multilingual versus monolingual wsd. In *In Proceedings of the EACL Workshop "Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together", April 3-7*, pages 33–40, 2006.

[130] Ralf Steinberger, Johan Hagman, and Stefan Scheer. Using thesauri for automatic indexing and for the visualisation of multilingual document collections. In *Proceedings of the workshop on Ontologies and lexical knowledge bases*, pages 130–141, 2000.

[131] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. The jrc-acquis : A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, pages 2142–2147, 2006.

[132] Fangzhong Su and Bogdan Babych. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation*

(HyTra), EACL 2012, pages 10–19, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[133] John Swales. *Genre analysis : English in academic and research settings*. Cambridge University Press, 1990.

[134] Tuomas Talvensaari. Effects of aligned corpus quality and size in corpus-based clir. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 114–125, Berlin, Heidelberg, 2008. Springer-Verlag.

[135] Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, 25(1), February 2007.

[136] Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 24–36, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[137] Tao Tao. Mining comparable bilingual text corpora for cross-language information integration. In *In KDD*, pages 691–696, 2005.

[138] Wolfgang Teubert. Comparable or parallel corpora ? *International Journal of Lexicography*, 9(3) :238–264, 1996.

[139] RobertL. Thorndike. Who belongs in the family ? *Psychometrika*, 18(4) :267–276, 1953.

[140] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63(2) :411–423, 2001.

[141] Elena Tognini-Bonelli. *Patterns and Meanings. Using Corpora for English Language Research and Teaching*, volume Studies in Corpus Linguistics Vol. 2. Amsterdam : John Benjamins Publishing Company, 1998.

[142] Richard Xiao Tony McEnery. *Parallel and comparable corpora : what is happening ?* Multilingual Matters, 2007.

[143] Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. Mining named entity transliteration equivalents from comparable corpora. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1423–1424, New York, NY, USA, 2008. ACM.

[144] Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck. Two-mode clustering methods : a structured overview. *Statistical methods in medical research*, 13(5) :363–394, 2004.

[145] De Boeck P Van Mechelen I, Bock HH. Two-mode clustering methods :a structured overview. *Statistical Methods in Medical Research*, 13(5) :363–394, 2004.

[146] C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2) :106–199, 1977.

[147] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[148] Leon Versteegen. The simple bayesian classifier as a classification algorithm.

[149] Thuy Vu, Ai Ti Aw, and Min Zhang. Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 843–851, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[150] Chih-Ping Wei, Christopher C Yang, and Chia-Min Lin. A latent semantic indexing-based approach to multilingual document clustering. *Decision Support Systems*, 45(3) :606–620, 2008.

[151] Xin Liu Wei Xu and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR'03*, pages 267–273, 2003.

[152] Geoffrey Williams. A cultivated audience : Comparable corpora and cross language collocation. *Rassegna Italiana di Linguistica Applicata*, (1-2) :39–64, 2011.

[153] Dekai Wu and Pascale Fung. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proceedings of the Second international joint conference on Natural Language Processing*, IJCNLP'05, pages 257–268, Berlin, Heidelberg, 2005. Springer-Verlag.

[154] Rui Xu and D. Wunsch, II. Survey of clustering algorithms. *Trans. Neur. Netw.*, 16(3) :645–678, May 2005.

[155] Christopher C. Yang and Kar Wing Li. Automatic construction of english/chinese parallel corpora. *Journal of the American Society for Information Science and Technology*, 54 :730–742, 2003.

[156] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49, New York, NY, USA, 1999. ACM.

[157] Ye yi Wang and Alex Waibel. Fast decoding for statistical machine translation. In *In Proc. Int. Conf. Spoken Language Processing*, pages 2775–2778, 1998.

[158] Federico Zanettin. Bilingual comparable corpora and the training of translators. *Meta : Translators' Journal*, Volume 43 :616–630, 1998.

[159] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 25–32, New York, NY, USA, 2001. ACM.

[160] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch : an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, SIGMOD '96, pages 103–114, New York, NY, USA, 1996. ACM.

[161] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

A

Mots vides anglais et français

Mots vides anglais :

a’s able about above according accordingly across actually after afterwards again against ain’t all allow allows almost alone along already also although always am among amongst an and another any anybody anyhow anyone anything anyway anyways anywhere apart appear appreciate appropriate are aren’t around as aside ask asking associated at available away awfully be became because become becomes becoming been before beforehand behind being believe below beside besides best better between beyond both brief but by c’mon c’s came can can’t cannot cant cause causes certain certainly changes clearly co com come comes concerning consequently consider considering contain containing contains corresponding could couldn’t course currently definitely described despite did didn’t different do does doesn’t doing don’t done down downwards during each edu eg eight either else elsewhere enough entirely especially et etc even ever every everybody everyone everything everywhere ex exactly example except far few fifth first five followed following follows for former formerly forth four from further furthermore get gets getting given gives go goes going gone got gotten greetings had hadn’t happens hardly has hasn’t have haven’t having he he’s hello help hence her here here’s hereafter hereby herein hereupon hers herself hi him himself his hither hopefully how howbeit however i’d i’ll i’m i’ve ie if ignored immediate in inasmuch inc indeed indicate indicated indicates inner insofar instead into inward is isn’t it it’d it’ll it’s its itself just keep keeps kept know knows known last lately later latter latterly least less lest let let’s like liked likely little look looking looks ltd mainly many may maybe me mean meanwhile merely might more moreover most mostly much must my myself name namely nd near nearly necessary need needs neither never nevertheless new next nine no nobody non none noone nor normally not nothing novel now nowhere obviously of off often oh ok okay old on once one ones only onto or other others otherwise ought our ours ourselves out outside over overall own particular particularly per perhaps placed please plus possible presumably probably provides que quite qv rather rd re really reasonably regarding regardless regards relatively respectively right said same saw say saying says second secondly see seeing seem seemed seeming seems seen self selves sensible sent serious seriously seven several shall she should shouldn’t since six so some somebody somehow someone something sometime sometimes somewhat somewhere soon sorry specified

specify specifying still sub such sup sure t's take taken tell tends th than thank thanks thanx that that's thats the their theirs them themselves then thence there there's thereafter thereby therefore therein theres thereupon these they they'd they'll they're they've think third this thorough thoroughly those though three through throughout thru thus to together too took toward towards tried tries truly try trying twice two un under unfortunately unless unlikely until unto up upon us use used useful uses using usually value various very via viz vs want wants was wasn't way we we'd we'll we're we've welcome well went were weren't what what's whatever when whence whenever where where's whereafter whereas whereby wherein whereupon wherever whether which while whither who who's whoever whole whom whose why will willing wish with within without won't wonder would would wouldn't yes yet you you'd you'll you're you've your yours yourself yourselves zero

Mots vides français :

à allô aucuns auriez auxdits aviez ayons bof çà certaines chez comment da desquels deviez devras doit dues dût es êtes eurêka excepté fouchtra fûmes ho hurrah laquelle leur mazette mâtin ne nulle or outre pas plein pourraient pourvu pouviez puis pussent que quoi saperlipopette serait sien sommes ta telles touchant une veuillez voilà voudrez voulante voulue vôtre afin alors auquel aurions auxquelles avions aïe boum car certains chic concernant dans devaient devions devrez doive duquel eh et étiez eus eûmes furent fût holà hé le leurs me miséricorde ni nulles ôté palsambleu patatras plouf pourrais pouvaient pouvions puisque put quel quoique sapristi seras sienne son tandis tels tous unième veuillons vos voudriez voulantes voulues vôtres ah apr. aura aurons auxquels avoir bah bravissimo ce ces chiche contre de devais devoir devriez doivent durant elle étaient étions eusse eût fus fûtes hop il ledit lorsque merci moi nonobstant nuls ou pan pechère plus pourrait pouvais pouvoir puisse pécaïre quelle rataplan sauf serez siennes sont tant tes tout unièmes veulent votre voudrions voulants voulurent zut ai as aurai auront avaient avons basta bravo ceci cet chouette corbleu debout devait devons devrions doives durent elles étais être eussent eûtes fusse grâce hormis ils lequel lui merde moins nos ô où par pendant plusieurs pourras pouvait pouvons puissent pût quelles revoici se seriez siens sous taratata tien toute v'lan veut voudra voudrons voulez voulus aie attendu auraient autant avais ayant beaucoup ç'a cela cette chut coucou depuis devant devra devrons donc dus en était eu eusses évohé fussent ha hors jarnicoton les là mes mon notre oh ouais parbleu peu pouah pourrez pouvant psitt puisses qq. quelqu'un revoilà selon serions sinon soyez tayaut tienne toutes va veux voudrai voudront vouliez voulussent aient au aurais autre avait ayante bernique ç'aura celle ceux ciao couic des devante devrai devront dont dussent encontre étant eue eussiez évoé fusses hein hou je lesdites ma mien morbleu nôtre ohé ouf parce peuchère pour pourriez pouvante pst puissiez qqch. quelqu'une rien sera serons soi soyons taïaut tiennes tu vers via voudraient voulaient voulions voulut aies aucun aurait autres avant ayantes bien ç'aurait celles chacun clic crac desdites devantes devraient dia du dut endéans étante eues eussions fi fussiez hem hourra jusque lesdits made mienne motus nôtres olé ouille pardi peut pourquoi pourrions pouvantes pu puissions qqn quels sa serai seront soient stop te tiens tudieu veuille vivement

voudrais voulais vouloir voulût ait aucune auras aux avec ayants bigre ç'avait celui chacune clac cric desdits devants devrais diantre dudit dès entre étantes euh eut fichtre fussions hep hue la lesquelles mais miennes moyennant nous on oust pardieu peuvent pourra pourrons pouvants pue purent quand qui sacristi seraient ses sois suis tel toi turlututu veuillent vlan voudrait voulait voulons vous al. aucunes aurez auxdites avez ayez bis ça cependant chaque comme crénom desquelles devez devrait dois due dû envers étants eurent eux fors fut heu hum ladite lesquels malgré miens na nul ont ouste parmi peux pourrai pourront pouvez pues pus quant quiconque sans serais si soit sur telle ton un veuilles voici voudras voulant voulu vu à allô aucuns auriez auxdits aviez ayons bof çà certaines chez comment da desquels deviez devras doit dues dût es êtes eurêka excepté fouchtra fûmes ho hurrah laquelle leur mazette mâtin ne nulle or outre pas plein pourraient pourvu pouviez puis pussent que quoi saperlipopette serait sien sommes ta telles touchant une veuillez voilà voudrez voulante voulue vôtre afin alors auquel aurions auxquelles avions aïe boum car certains chic concernant dans devaient devions devrez doive duquel eh et étiez eus eûmes furent fût holà hé le leurs me miséricorde ni nulles ôté palsambleu patatras plouf pourrais pouvaient pouvions puisque put quel quoique sapristi seras sienne son tandis tels tous unième veuillons vos voudriez voulantes voulues vôtres ah apr. aura aurons auxquels avoir bah bravissimo ce ces chiche contre de devais devoir devriez doivent durant elle étaient étions eusse eût fus fûtes hop il ledit lorsque merci moi nonobstant nuls ou pan pechère plus pourrait pouvais pouvoir puisse pécaïre quelle rataplan sauf serez siennes sont tant tes tout unièmes veulent votre voudrions voulants voulurent zut ai as aurai auront avaient avons basta bravo ceci cet chouette corbleu debout devait devons devrions doives durent elles étais être eussent eûtes fusse grâce hormis ils lequel lui merde moins nos ô où par pendant plusieurs pourras pouvait pouvons puissent pût quelles revoici se seriez siens sous taratata tien toute v'lan veut voudra voudrons voulez voulus aie attendu auraient autant avais ayant beaucoup ç'a cela cette chut coucou depuis devant devra devrons donc dus en était eu eusses évohé fussent ha hors jarnicoton les là mes mon notre oh ouais parbleu peu pouah pourrez pouvant psitt puisses qq. quelqu'un revoilà selon serions sinon soyez tayaut tienne toutes va veux voudrai voudront vouliez voulussent aient au aurais autre avait ayante bernique ç'aura celle ceux ciao couic des devante devrai devront dont dussent encontre étant eue eussiez évoé fusses hein hou je lesdites ma mien morbleu nôtre ohé ouf parce peuchère pour pourriez pouvante pst puissiez qqch. quelqu'une rien sera serons soi soyons taïaut tiennes tu vers via voudraient voulaient voulions voulut aies aucun aurait autres avant ayantes bien ç'aurait celles chacun clic crac desdites devantes devraient dia du dut endéans étante eues eussions fi fussiez hem hourra jusque lesdits made mienne motus nôtres olé ouille pardi peut pourquoi pourrions pouvantes pu puissions qqn quels sa serai seront soient stop te tiens tudieu veuille vivement voudrais voulais vouloir voulût ait aucune auras aux avec ayants bigre ç'avait celui chacune clac cric desdits devants devrais diantre dudit dès entre étantes euh eut fichtre fussions hep hue la lesquelles mais miennes moyennant nous on oust pardieu peuvent pourra pourrons pouvants pue purent quand qui sacristi seraient ses sois suis tel toi turlututu veuillent vlan voudrait voulait voulons vous al. aucunes aurez auxdites avez ayez bis ça cependant chaque comme crénom desquelles devez

devrait dois due dû envers étants eurent eux fors fut heu hum ladite lesquels malgré miens na nul ont ouste parmi peux pourrai pourront pouvez pues pus quant quiconque sans serais si soit sur telle ton un veuilles voici voudras voulant voulu vu

B

Dix premières paires de clusters obtenues sur la base du **Tri séquentiel**

Nous présentons ici dix premières paires de clusters alignés, représentées par leurs médoides, obtenues sur la base du **Tri séquentiel** des plus fortes comparabilités pour le corpus Flux RSS.

B.1 Premières paire de clusters : ”Syrie-Iraq”

Valeur de comparabilité entre les deux clusters alignés : 0,5695421909101351

<DOC>
<DOCID>9351e1af5d0a054612be0cf2aebcf9d2</DOCID>
<PUBDATE>Tue Mar 05 01:43:22 CET 2013</PUBDATE>
<CURDATE>Tue Mar 05 19:18:40 CET 2013</CURDATE>
<FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI>
<ITEMURI>http://www.globaltimes.cn/content/765831.shtml</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Up to 35 Syrian soldiers, 4 Iraqi soldiers killed in western Iraq</TITLE>
<DESC>Up to 35 Syrian soldiers and four Iraqi soldiers were killed, and seven Iraqi soldiers wounded in an attack by gunmen on an Iraqi army convoy, which was escorting a bus carrying Syrian soldiers, in Iraq's Anbar province on Monday, a police source told Xinhua.</DESC>
<TXT>Up to 35 Syrian soldiers, 4 Iraqi soldiers killed in western Iraq Xinhua | 2013-3-5 8:43:22 Print Up to 35 Syrian soldiers and four Iraqi soldiers were killed, and seven Iraqi soldiers wounded in an attack by gunmen on an Iraqi army convoy, which was escorting a bus carrying Syrian soldiers, in Iraq's Anbar province on Monday, a police source told Xinhua. "Our latest report said that 35 Syrian soldiers and four Iraqi soldiers were killed and at least seven Iraqi soldiers were wounded by the attack in western Iraq," the source from Anbar police said on condition of anonymity. The attack occurred near the city of Rutba, some 375 km west of Baghdad, when unidentified gunmen attacked a convoy of military vehicles that were escorting a bus carrying Syrian soldiers who entered Iraq after the Syrian rebels seized al-Yaroubiyah border crossing point with Iraq's northern province of Nineveh, the source said. The convoy was trying to transfer the Syrian soldiers from Nineveh province to Syria through al-Walid border crossing point in Iraq's western province of Anbar, he said. The attackers burnt the bus carrying the Syrian soldiers and destroyed two Iraqi military vehicles, the source added. Earlier, the source put the toll at 20 Syrian soldiers killed and seven Iraqi soldiers wounded by the attack. On Saturday, Syrian opposition forces took control of al-Yaroubiyah border crossing point after fierce clashes with the Syrian army, forcing some Syrian soldiers to cross into Iraq. On Sunday, the Iraqi authorities in Nineveh said several wounded Syrian soldiers were allowed to enter Iraq to receive treatment for their wounds at the Iraqi hospitals. The Iraqi authorities said the Syrian soldiers were admitted to hospitals for humanitarian reasons. Iraq has a borderline of around 600 kilometers with Syria. The border crossings had witnessed closure in history when crisis intensified on either side.
</TXT>
</DOC>

<DOC>
<DOCID>1ca9388bddb37313f52ee48708d399c6</DOCID>
<PUBDATE>Tue Mar 05 06:00:00 CET 2013</PUBDATE>
<CURDATE>Tue Mar 05 19:25:04 CET 2013</CURDATE>
<FEEDURI>http://www.ledevoir.com/rss/section/international.xml?id=76</FEEDURI>
<ITEMURI>http://www.ledevoir.com/international/actualites-internationales/372446/48-soldats-syriens-sont-tues-en-irak</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Bagdad accusée d’ingérence dans le pays voisin - 48 soldats syriens sont tués en Irak</TITLE>
<DESC>Quarante-huit soldats syriens qui s’étaient réfugiés en Irak pendant le...</DESC>
<TXT>5 mars 2013 International / Actualités internationales
Bagdad accusée d’ingérence dans le pays voisin - 48 soldats syriens sont tués en Irak
Tweet
Importante victoire des rebelles
Les rebelles ont remporté lundi leur victoire la plus importante depuis le début de la révolte en Syrie, en s’emparant d’un chef-lieu de province, au moment où Washington avalisait implicitement des livraisons d’armes par des pays du Golfe.
À Raqa, chef-lieu de la province éponyme, « les rebelles contrôlent presque entièrement la ville », a annoncé l’Observatoire syrien des droits de l’Homme (OSDH).
Les combats se poursuivaient et la ville était visée par des raids de l’aviation du régime, indique la même source.
« Dans les prochaines heures, Raqa sera la première capitale de province à être hors du contrôle du régime », a déclaré Rami Abdel Rahmane, directeur de l’OSDH.
Quarante-huit soldats syriens qui s’étaient réfugiés en Irak pendant le...
Quarante-huit soldats syriens qui s’étaient réfugiés en Irak pendant le week-end et neuf Irakiens ont été tués lundi dans une embuscade en territoire irakien, selon le ministère de la Défense à Bagdad.
Cette embuscade dans la province d’Anbar, 24 heures après qu’une importante composante de l’opposition syrienne a accusé l’Irak d’ingérence en Syrie, risque d’entraîner ce pays dans la guerre civile qui déchire son voisin syrien, ce que Bagdad s’est engagé à ne pas laisser faire.
L’embuscade a été tendue par « un groupe terroriste qui s’est infiltré en territoire irakien depuis la Syrie », a indiqué le ministère dans un communiqué, dénonçant « une attaque contre la souveraineté de l’Irak, son territoire, sa dignité et une violation claire des droits de la personne, [les soldats] étant blessés et non armés ».
Neuf gardes irakiens escortant les soldats syriens ont également été tués dans l’embuscade, selon la même source.
Bagdad va néanmoins résister aux tentatives de propager le conflit syrien en Irak, a promis lundi le porte-parole du premier ministre irakien.
« Cela confirme nos craintes sur le fait que certains tentent de propager la crise syrienne en Irak, mais nous ferons face à ces tentatives d’où qu’elles viennent avec toute notre force », a prévenu Ali Moussaoui, porte-parole du chef du gouvernement, Nouri al-Maliki.
Les soldats syriens avaient franchi la frontière par le point de passage de Yaaroubiyeh pour fuir de violents combats ayant opposé samedi, du côté syrien de la frontière, l’armée syrienne aux rebelles luttant contre le régime du président syrien, Bachar al-Assad, a indiqué le colonel Mohammed Khalaf al-Dulaimi, des forces de protection de la frontière.
...
</TXT>
</DOC>

FIGURE B.1 – Alignement des deux médoides : ”Syrie-Irak”

B.2 Deuxième paire de clusters : ”Iran”

Valeur de comparabilité entre les deux clusters alignés : 0,5506803919337105

<DOC>
<DOCID>d7b4a9027836d84e06210fa3d8519f63</DOCID>
<PUBDATE>Sun Apr 07 03:17:51 CEST 2013</PUBDATE>
<CURDATE>Sun Apr 07 09:28:47 CEST 2013</CURDATE>
<FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI>
<ITEMURI>http://www.globaltimes.cn/content/772969.shtml</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Despite large differences, Iran, world powers still favor talks</TITLE>
<DESC>Despite major differences between Iran and the world powers over the Islamic republic's disputed nuclear program during the latest round of talks, analysts say gestures for the continuation of negotiations are still resonated from both sides, which is favored to be the best choice.</DESC>
<TXT>Despite large differences, Iran, world powers still favor talks Xinhua | 2013-4-7 9:17:51 Print Despite major differences between Iran and the world powers over the Islamic republic's disputed nuclear program during the latest round of talks, analysts say gestures for the continuation of negotiations are still resonated from both sides, which is favored to be the best choice.
The two-day nuclear talks between Iran and so-called P5+1 group (the five permanent members of the UN Security Council plus Germany) in Almaty of Kazakhstan concluded on Saturday with both sides saying that the gap of views over the key points could not be bridged. Iran and the P5+1 group remained "far apart" on key issues, Catherine Ashton, the European Union's head of foreign policy, said at the end of the meeting. Echoing her words, Iranian chief nuclear negotiator Saeed Jalili side as far as Iran's nuclear issue is concerned, " differences in the views of the parties exist." While Iranian officials stressed that any mechanism to settle Iran's nuclear issue in the talks should take into consideration the recognition of Iran's right to enrichment activities, the revised proposal by the world powers asked Iran to suspend its uranium enrichment and shut down its underground Fordow enrichment facilities in return for limited sanction relief.
"Either 5 percent or 20 percent uranium enrichment is part of Iranian nation's right," and as a means to building confidence, the West should give up its "hostile attitude" toward the Islamic republic, Jalili asserted, implying the torrent of sanctions that have been imposed against the country for years. Calling the sanctions as a tool to press Iran to refrain from its suspected nuclear activities, Ashton said that "the purpose of sanctions is to put pressure on Iran to see if this process works. " Still divided on the main topic of the meeting -- uranium enrichment -- the two sides disinclined to leave the talks closed- ended. Although no specific date and venue was designated for the follow-up meetings, the Iranian chief nuclear negotiator said that his country stresses the continuation of talks and Ashton will contact him later to discuss the means of how to proceed with the negotiations.
…
</TXT></DOC>

<DOC>
<DOCID>0be102555caffa2ef3d7804ac52aa9cb</DOCID>
<PUBDATE>Sat Apr 06 13:54:00 CEST 2013</PUBDATE>
<CURDATE>Sat Apr 06 23:09:38 CEST 2013</CURDATE>
<FEEDURI>http://www.lapresse.ca/rss/179.xml</FEEDURI>
<ITEMURI>http://rss.lapresse.ca/c/33663/f/607972/s/2a6705bf/l/0L0Slapresse0Bca0Cinternational0Cdossiers0Cnucleaire0Eiranien0C20A130A40C0A60C0A10E4638180A0Enucleaire0Eliran0Eet0Eles0Egrandes0Epuissances0Eencore0Etres0Eeloignes0Bphp/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Nucléaire: l'Iran et les grandes puissances encore «très éloignés»</TITLE>
<DESC>L'Iran et les grandes puissances ne sont pas parvenus à faire une percée dans les négociations sur le programme nucléaire iranien controversé, les positions des deux parties restant «très éloignées» après deux journées d'intenses pourparlers au Kazakhstan.</DESC>
<TXT>Agence France-Presse
ALMATY
L'Iran et les grandes puissances ne sont pas parvenus à faire une percée dans les négociations sur le programme nucléaire iranien controversé, les positions des deux parties restant «très éloignées» après deux journées d'intenses pourparlers au Kazakhstan.
«Il est devenu clair que les positions (des grandes puissances) et de l'Iran restent très éloignées sur le fond», a déclaré Catherine Ashton, la représentante de la diplomatie de l'Union européenne, qui dirige les négociations pour les grandes puissances.
Le chef des négociateurs iraniens, Saïd Jalili, a pour sa part reconnu «une certaine distance entre les positions des deux parties» à l'issue de deux longues journées de négociations à Almaty, la plus grande ville du Kazakhstan, en Asie centrale.
Les pays du groupe 5+1 (les cinq membres permanents du Conseil de sécurité de l'ONU - États-Unis, France, Grande-Bretagne, Russie et Chine - plus l'Allemagne) et l'Iran ne sont pas non plus parvenus à se mettre d'accord sur la date et le lieu de la prochaine rencontre, a ajouté Mme Ashton devant des journalistes.
En conséquence, «nous avons conclu que les représentants de toutes les parties allaient rentrer dans leur capitale pour évaluer où se situe le processus» des négociations, a souligné Mme Ashton.
«Un large fossé demeure entre les parties», a estimé dans un communiqué le chef de la diplomatie britannique, William Hague.
«L'actuelle position de l'Iran est très loin de ce qui est nécessaire pour réaliser une percée diplomatique», a-t-il ajouté.
Mme Ashton s'est engagée à «rester en contact» avec M. Jalili, afin de «voir comment continuer».
Ce dernier a insisté sur la reconnaissance internationale du droit de l'Iran à enrichir de l'uranium, le principal point sur lequel les grandes puissances exigent des concessions de Téhéran en promettant d'atténuer les sanctions visant Téhéran pour son programme nucléaire controversé.
«L'Iran veut faire des concessions très limitées concernant son programme nucléaire et s'attend en échange à des résultats significatifs», a résumé un haut responsable américain sous couvert de l'anonymat.
...
</TXT>
</DOC>

FIGURE B.2 – Alignement des deux médoides : ”Iran”

B.3 Troisième paire de clusters : ”Armes chimiques en Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,531008571627248

<DOC>
<DOCID>554b79fb9f7cde11a1c90552c62aefca</DOCID>
<PUBDATE>Tue Mar 19 17:33:00 CET 2013</PUBDATE>
<CURDATE>Tue Mar 19 22:07:30 CET 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEED URI>
<ITEMURI>http://www.bbc.co.uk/news/world-middle-east-21841217#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Syrians trade chemical attack claims</TITLE>
<DESC>Syrian rebels and the government accuse each other of firing chemical weapons, reportedly killing 25 people, but there is no confirmation of either claim.</DESC>
<TXT>Insurgency mapped Syrian rebels and the government have accused each other of firing chemical weapons, reportedly killing at least 25 people in the north of the country. A Syrian minister said it was a "dangerous escalation" and the "first act" of a new rebel authority. However, both a chemical weapons monitoring body and the US said there was no evidence they had been used. Both sides say the attack happened in the Khan al-Assal region north of the second city, Aleppo. The US says it is looking carefully at the allegations, while Russia has backed the Syrian government's claims. Analysis Jonathan Marcus BBC Diplomatic Correspondent At this stage it is impossible to verify if a chemical weapon has been used in northern Syria or indeed who may have fired one. The government and the rebels both accuse each other. Given Syria's extensive chemical weapons arsenal there have long been fears that either the government forces might use such weapons or that they might fall into the hands of Syrian opposition fighters. The exact status of Syria's arsenal which includes blister agents like mustard and persistent nerve agents like Sarin is unclear. There have been periodic "scares" when Western intelligence agencies claimed to see activity at weapons depots, but so far there has been no hard evidence that chemical warheads have been delivered to the units that might fire them. US President Barack Obama has made it clear that the use of such weapons would represent a "red line", which if crossed would lead to serious consequences. If confirmed, it would be the first time chemical weapons have been used in the two-year Syrian conflict. "Terrorists launched a missile containing chemical products into the region of Khan al-Assal in the province of Aleppo, killing 15 people, mainly civilians," Sana news agency said. The government routinely refers to rebels as "terrorists". State TV later said 25 people had died, while the pro-opposition Syrian Observatory for Human Rights put the figure at 26, including 16 soldiers. Senior rebel and spokesman for the Higher Military Council in Aleppo Qassim Saadeddine said the government had carried out a chemical attack. "We were hearing reports from early this morning about a regime attack on Khan al-Assal, and we believe they fired a Scud with chemical agents," he told Reuters news agency. "Then suddenly we learned that the regime was turning these reports against us. The rebels were not behind this attack."
</TXT>
</DOC>

<DOC>
<DOCID>04817110f793022e84e24573860bcfe1</DOCID>
<PUBDATE>Tue Mar 19 18:49:00 CET 2013</PUBDATE>
<CURDATE>Tue Mar 19 21:53:28 CET 2013</CURDATE>
<FEEDURI>http://www.lavenir.net/rss.aspx?foto=1&intro=1&section=info&info=1642237c-66b9-4e8a-a8c1-288d61fefe7e</FEEDURI>
<ITEMURI>http://www.lavenir.net/article/detail.aspx?articleid=DMF20130319_00284325</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Régime et rebelles syriens s’accusent d’user d’armes chimiques pour la première fois</TITLE>
<DESC>Le régime et les rebelles se sont mardi accusés mutuellement d’avoir utilisé des armes chimiques, pour la première fois en deux ans de conflit en Syrie, mais les États-Unis ont dit ne disposer d’aucune preuve sur un tel recours.</DESC>
<TXT>Guerre en Syrie
Le régime et les rebelles se sont mardi accusés mutuellement d’avoir utilisé des armes chimiques, pour la première fois en deux ans de conflit en Syrie, mais les États-Unis ont dit ne disposer d’aucune preuve sur un tel recours.
La Russie a repris à son compte les accusations du régime de Bachar al-Assad, son allié, en disant avoir «reçu des informations» selon lesquelles des rebelles ont utilisé des armes chimiques lors d’une attaque dans la province d’Alep (nord) qui a fait selon un dernier bilan officiel 31 morts.
Alors qu’il n’était pas possible de confirmer de source indépendante un tel recours, l’Observatoire syrien des droits de l’Homme (OSDH) a confirmé un tir de missile sol-sol contre l’armée dans la localité de Khan al-Assal mais a dit douter qu’il soit chargé de matières non conventionnelles.
Entre-temps à Istanbul, le «Premier ministre» intérimaire Ghassan Hitto, élu lundi par l’opposition, a dit qu’il ne dialoguerait pas avec le pouvoir et cité parmi ses priorités la chute du régime et l’envoi de l’aide aux populations des régions passées sous contrôle rebelle.
M. Hitto, de la mouvance islamiste, doit s’atteler à la formation de l’équipe attendue d’ici un mois et qui aura la charge de protéger les infrastructures et les ressources publiques et privées, gérer les postes frontières aux mains de la rébellion et coordonner l’aide humanitaire internationale.
Paris et Washington, qui soutiennent les rebelles, ont salué cette élection.
Une escalade dangereuse
«C’est une escalade dangereuse. Des terroristes ont tiré un missile contenant des produits chimiques à partir de Kfar Daël dans la région de Naïrab (est d’Alep) vers la région de Khan al-Assal (ouest de la métropole)», a déclaré le ministre syrien de l’Information Omrane al-Zohbi.
Il s’en est pris à «la Ligue arabe, la communauté internationale et les États qui arment, financent et hébergent les terroristes ainsi que le gouvernement (turc) d’Erdogan et le Qatar» après «ce crime perpétré par les terroristes qui ont utilisé une arme prohibée par la loi internationale».
Mais les rebelles de l’Armée syrienne libre (ASL) ont démenti cette allégation et accusé en retour le régime.
…
</TXT>
</DOC>

FIGURE B.3 – Alignement des deux médoides : ”Armes chimiques en Syrie”

B.4 Quatrième paire de clusters : ”Querre civile en Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,5002720842849248

<DOC>
<DOCID>95a66d7b1d8dd99c0963c90666c0406b</DOCID>
<PUBDATE>Thu Feb 28 02:10:44 CET 2013</PUBDATE>
<CURDATE>Fri Mar 01 18:27:35 CET 2013</CURDATE>
<FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI>
<ITEMURI>http://www.globaltimes.cn/content/764821.shtml</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Syria gov't shows more leniency toward dialogue than opposition</TITLE>
<DESC>Syria has made it clear that it is ready for dialogue, even with the armed parties, throwing thus the ball into the court of the opposition groups that have not so far shown any leniency in its stances and made the dialogue conditional on the departure of the current administration.</DESC>
<TXT>Syria gov't shows more leniency toward dialogue than opposition Xinhua | 2013-2-28 9:10:44 Print Syria has made it clear that it is ready for dialogue, even with the armed parties, throwing thus the ball into the court of the opposition groups that have not so far shown any leniency in its stances and made the dialogue conditional on the departure of the current administration. Syrian Foreign Minister Walid al-Moallem has recently announced that Syria was ready to open dialogue with the opposition, even the armed groups, a step which was seen by observers as aiming to placate the Syrian people who are eager to reach a political solution to the country's nearly two-year crisis. However, Salim Edris, the so-called chief of staff at the rebels Syrian Free Army, turned down al-Moallem's offer for dialogue, stipulating that President Bashar al-Assad should step down ahead of any dialogue and called for the cessation for all kinds of killing and the withdrawal of the Syrian army from cities. Edris' conditions mirrored the ones by the political opposition abroad that also said Assad's departure should be the result of any dialogue. Yet, the two parties of the conflict have apparently come under international pressure to embark on direct dialogue ahead of a Rome meeting of the Friends of Syria group slated for Thursday in the hope of narrowing the schism between both conflicting parties. Syrian observers, meanwhile, believe that the Syrian crisis could not be solved but through dialogue especially as military operations and attacks have reached the capital, prompting thousands of Syrians to flee the country and many others to lock themselves inside their houses. "I can see that the political solution is now more clear ... it 's very important factor in the Syrian crisis ... many of the countries in the world support it," Bassam Abu Abdullah, an expert in international relations, told Xinhua in an interview Wednesday. Abdullah, who is also a university professor in international law, said that al-Moallem's call for dialogue was "very important . .. and we considered it as some kind of an opening from the Syrian government toward dialogue," adding that "We are going toward the political solution."
...
</TXT>
</DOC>

<DOC>
<DOCID>8730f8c94ae11a12d7cb24a9faf86ce1</DOCID>
<PUBDATE>Mon Feb 25 16:14:00 CET 2013</PUBDATE>
<CURDATE>Tue Feb 26 16:06:47 CET 2013</CURDATE>
<FEEDURI>http://www.lapresse.ca/rss/179.xml</FEEDURI>
<ITEMURI>http://rss.lapresse.ca/c/33663/f/607972/s/28f046fd/l/0L0Slapresse0Bca0Cinternational0Cdossiers0Ccrise0Edans0Ele0Emonde0Earabe0Csyrie0C20A130A20C250C0A10E46251750Edamas0Epret0Ea0Edialoguer0Eavec0Eles0Erebelles0Bphp/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Damas prêt à dialoguer avec les rebelles</TITLE>
<DESC>Le régime du président syrien Bachar al-Assad s'est dit pour la première fois lundi prêt au dialogue avec les rebelles armés pour mettre fin au conflit, mais ces derniers ont rejeté toute négociation avant le départ du chef de l'Etat et le retrait de l'armée des villes.</DESC>
<TXT>> Damas prêt à dialoguer avec les rebelles
Damas prêt à dialoguer avec les rebelles
«Nous sommes prêts au dialogue avec tous ceux qui veulent le dialogue, y compris les groupes armés», a déclaré le ministre syrien des Affaires étrangères Walid al-Mouallem (ci-dessus) au début de ses entretiens avec son homologue russe, Sergueï Lavrov.
PHOTO YURI KADOBNOV, AFP
Consultez notre dossier complet sur les soulèvements populaires en Afrique. »
À lire aussi
Agence France-Presse
Moscou
Le régime du président syrien Bachar al-Assad s'est dit pour la première fois lundi prêt au dialogue avec les rebelles armés pour mettre fin au conflit, mais ces derniers ont rejeté toute négociation avant le départ du chef de l'Etat et le retrait de l'armée des villes.
Cependant, la Coalition de l'opposition syrienne a annoncé qu'elle participerait finalement jeudi à Rome à la réunion des Amis du peuple syrien qu'elle avait menacé de boycotter pour dénoncer «le silence international» sur les crimes commis par le régime.
«Nous sommes prêts au dialogue avec tous ceux qui veulent le dialogue, y compris les groupes armés», a déclaré le chef de la diplomatie syrienne, Walid al-Mouallem, au début d'entretiens à Moscou avec son homologue russe, Sergueï Lavrov.
«Nous restons favorables à un règlement pacifique du problème syrien. Une commission gouvernementale a été créée pour mener des pourparlers avec l'opposition dans le pays et même avec l'opposition à l'extérieur» de la Syrie, a-t-il ajouté.
Mais le chef d'état-major de l'armée rebelle en Syrie, Sélim Idriss, a affirmé que les insurgés refusaient toute négociation avec Damas avant le départ de M. Assad et le retrait de l'armée des villes.
«Walid Mouallem veut qu'on s'asseoie avec lui à la table de négociations (...). Je ne m'assiérai avec Mouallem ou quelqu'un d'autre de cette clique qu'après l'arrêt de toutes les tueries et le retrait de l'armée des villes» et le départ «du chef de la bande criminelle», a-t-il dit.
Le secrétaire d'Etat américain John Kerry, qui doit rencontrer M. Lavrov mardi à Berlin, s'est montré très sceptique quant à la proposition de M. Mouallem.
…
</TXT>
</DOC>

FIGURE B.4 – Alignement des deux médoides : ”Querre civile en Syrie”

B.5 Cinquième paire de clusters : ”Président chinois”

Valeur de comparabilité entre les deux clusters alignés : 0,49523592451707815

<DOC>
<DOCID>2bafd9e042c8463b842325c4235d917f</DOCID>
<PUBDATE>Sat Mar 23 07:30:29 CET 2013</PUBDATE>
<CURDATE>Sat Mar 23 10:04:21 CET 2013</CURDATE>
<FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI>
<ITEMURI>http://www.globaltimes.cn/content/770052.shtml</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Chinese, Russian presidents vow to enhance comprehensive strategic cooperation</TITLE>
<DESC>Visiting Chinese President Xi Jinping and his Russian counterpart, Vladimir Putin, held talks here Friday and vowed to enhance their countries' comprehensive strategic cooperation.</DESC>
<TXT>Chinese, Russian presidents vow to enhance comprehensive strategic cooperation Xinhua | 2013-3-23 14:30:29 Print Visiting Chinese President Xi Jinping and his Russian counterpart, Vladimir Putin, held talks here Friday and vowed to enhance their countries' comprehensive strategic cooperation. During their meeting in the Kremlin, Putin extended his warm welcome to Xi, and said the fact that Xi selected Russia as the first foreign country to visit after assuming presidency testifies to the great importance both sides attach to the development of their relations as well as the special and strategic nature of the relationship. Describing the visit as one of historic significance, Putin said he is confident that the trip will bear rich fruit and give a strong boost to the development of the comprehensive strategic cooperative partnership between Russia and China. Xi stressed that China and Russia are each other's major and most important strategic cooperative partners, and both accord priority to deepening their comprehensive strategic cooperative partnership in their overall diplomatic agenda and foreign policy. In face of the profoundly complex international situation and the still grave global economic environment, the two sides should work together more closely to enhance their comprehensive strategic cooperation, said the Chinese president. China and Russia, he proposed, should deepen mutual political support, steadfastly backing each other's efforts to safeguard national sovereignty, security and development interests, each other's independent choice of development paths and each other's cause of national rejuvenation. The two sides should also expand practical cooperation, translate the advantage of their high-level political ties into tangible results and thus achieve common development, Xi added. Meanwhile, he continued, Beijing and Moscow should strengthen coordination and cooperation on global and regional issues so as to safeguard the two countries' common strategic security. The collaboration on the world stage should also aim at promoting world peace, stability and prosperity by defending the principles of the UN Charter and the basic norms of international relations, the post-World War II international order as well as international justice and fairness, he said.
...
</TXT>
</DOC>

<DOC>
<DOCID>b4ad45990c1781b1841309fa4c2c9808</DOCID>
<PUBDATE>Fri Mar 22 04:06:00 CET 2013</PUBDATE>
<CURDATE>Sat Mar 23 10:17:21 CET 2013</CURDATE>
<FEEDURI>http://www.lapresse.ca/rss/179.xml</FEEDURI>
<ITEMURI>http://rss.lapresse.ca/c/33663/f/607972/s/29da3b2f/l/0L0Slapresse0Bca0Cinternational0Casie0Eoceanie0C20A130A30C210C0A10E46335850Ele0Epresident0Echinois0Een0Erussie0Epour0Esa0Epremiere0Evisite0Ea0Eletranger0Bphp/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Le président chinois en Russie pour sa première visite à l'étranger</TITLE>
<DESC>Le nouveau président chinois Xi Jinping entame vendredi une visite en Russie, premier déplacement à l'étranger depuis son investiture, afin de relancer les liens économiques et le «partenariat stratégique» entre les deux pays.</DESC>
<TXT>> Le président chinois en Russie pour sa première visite à l'étranger
Le président chinois en Russie pour sa première visite à l'étranger
Le nouveau président chinois Xi Jinping entame vendredi une visite en Russie pour son premier voyage à l'étranger.
Photo ANDY WONG, Agence France-Presse
Agence France-Presse
Moscou
Le nouveau président chinois Xi Jinping entame vendredi une visite en Russie, premier déplacement à l'étranger depuis son investiture, afin de relancer les liens économiques et le «partenariat stratégique» entre les deux pays.
«Le fait que la Russie amicale ait été choisie comme la première destination pour ma visite d'État témoigne du caractère particulier de nos relations stratégiques», a déclaré M. Xi dans une interview à des médias officiels russes.
Investi la semaine dernière comme président de la République populaire après avoir pris les rênes du Parti communiste en novembre, Xi Jinping répond à l'invitation de Vladimir Poutine, qui s'était rendu en Chine en juin 2012, moins d'un mois après sa prise de fonctions.
«Nous voulons la continuité dans nos relations avec la Chine qui ont un caractère stratégique (...) Nous apprécions beaucoup que M. Xi vienne en Russie pour sa première visite à l'étranger», a souligné le vice-ministre russe des Affaires étrangères Sergueï Riabkov.
Durant les deux dernières décennies, les échanges économiques ont dominé la relation sino-russes, Moscou fournissant à Pékin des technologies militaires et spatiales ainsi que du pétrole, tout en important massivement des produits de consommation courante chinois.
«Au cours des 20 dernières années, les échanges commerciaux bilatéraux ont été multipliés par 14 et ont atteint l'année dernière la somme record de 88,2 milliards de dollars», a souligné le chef de l'État chinois.
L'une des priorités est de porter les échanges «à 100 milliards de dollars d'ici 2015» et de «développer le partenariat énergétique», a-t-il ajouté.
Pour Sergueï Sanakoïev, secrétaire de la Chambre sino-russe pour la promotion du commerce de produits industriels d'innovation, la rencontre permettra d'établir un plan de coopération pour la décennie à venir.
…
</TXT>
</DOC>

FIGURE B.5 – Alignement des deux médoides : ”Président chinois”

B.6 Sixième paire de clusters : ”Israel et Turquie”

Valeur de comparabilité entre les deux clusters alignés : 0,4905583428832619

<DOC>
<DOCID>4f64ad06603f76414fbd0323e69e1dc9</DOCID>
<PUBDATE>Tue Mar 26 22:41:36 CET 2013</PUBDATE>
<CURDATE>Wed Mar 27 07:48:37 CET 2013</CURDATE>
<FEEDURI>http://www.france24.com/en/monde/rss</FEEDURI>
<ITEMURI>http://www.france24.com/en/20130326-israel-pay-turkey-millions-gaza-flotilla-deaths-haaretz?ns_campaign=editorial&ns_source=RSS_public&ns_mchannel=RSS&ns_fee=0&ns_linkname=20130326_israel_pay_turkey_millions_gaza_flotilla</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>HAARETZ: Israel to pay Turkey tens of millions over flotilla deaths</TITLE>
<DESC>Israel will transfer tens of millions of dollars to a humanitarian fund set up by the Turkish government to compensate for the deaths of nine Turkish activists aboard a flotilla bound for Gaza in 2010.</DESC>
<TXT>Israel to pay Turkey tens of millions over flotilla deaths Haaretz Israel will transfer tens of millions of dollars to a humanitarian fund set up by the Turkish government to compensate for the deaths of nine Turkish activists aboard a flotilla bound for Gaza in 2010. By HAARETZ (text) In the wake of Prime Minister Benjamin Netanyahus apology Friday to Turkish Prime Minister Recep Tayyip Erdogan over the deaths of nine Turkish activists aboard the 2010 Gaza flotilla, the two countries have set the wheels in motion to pay compensation over the deaths, with Israel set to pay out as much as tens of millions of dollars, according to sources in Turkey. High-level diplomatic contact between the two countries began on Monday when Turkish Foreign Minister Ahmet Davutoglu spoke with Justice Minister Tzipi Livni over the establishment of a joint committee that will formulate the terms of Israels agreement to pay compensation. The vice prime minister of Turkey, Bulent Arinc, told journalists on Monday that both sides agreed to establish a joint high-level committee over the coming days to discuss the details of the compensation transfer. Beyond the technical and legal questions over the compensation payments, the waiver of the legal claims and the extent of the blockade on Gaza, the Palestinian issue rather than the Syrian one will continue to be the focus of future relations between the two countries. The Turkish foreign minister made it clear during Tuesdays Arab League summit in Doha that Turkey will continue to stand with the Palestinian people and will act in order to end Israeli occupation. In Turkey, they estimate that the three-year long rift caused by the Palestinian question now gives Turkey leverage, and that the nature of the relationship between it and Israel will be largely dependent upon Israels behavior towards the Palestinians. As for Syria, Israel and Turkey see its future differently. While Israel is concerned by the possibility that Assads rule may fall and be replaced by an extremist Islamic regime, or that the state may be dismantled by armed forces that will control different sections, Turkey estimates that the Syrian opposition, which will also include Islamic movements, will be able to lead Syria and will not be a threat to the region.
…
</TXT>
</DOC>

<DOC>
<DOCID>065ca2c7abb83130f90c294aa62cf5c3</DOCID>
<PUBDATE>Sat Mar 23 08:44:40 CET 2013</PUBDATE>
<CURDATE>Sat Mar 23 10:12:36 CET 2013</CURDATE>
<FEEDURI>http://rss.lemonde.fr/c/205/f/3052/index.rss</FEEDURI>
<ITEMURI>http://www.lemonde.fr/proche-orient/article/2013/03/23/obama-met-un-terme-a-la-brouille-entre-la-turquie-et-israel_1853126_3218.html</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Obama met un terme à la brouille entre la Turquie et Israël</TITLE>
<DESC>Barack Obama a obtenu de Benyamin Nétanyahou qu'il présente ses excuses à son homologue turc pour l'abordage de la "Flottille de la liberté" en mai 2010.</DESC>
<TXT>Obama met un terme à la brouille entre la Turquie et Israël Le Monde |
• Mis à jour le
23.03.2013 à 09h46
C'est le succès diplomatique de dernière minute arraché par le président américain Barack Obama lors de sa visite en Israël , qui s'est achevée vendredi. Le premier ministre israélien Benyamin Nétanyahou a présenté ses excuses au chef du gouvernement turc Recep Tayyip Erdogan pour la mort de neuf Turcs à bord d'une flottille pour Gaza en 2010.
Dans un communiqué diffusé quelques minutes avant la fin de sa première visite officielle en Israël, Barack Obama indique que les deux hommes se sont entretenus par téléphone. "Les Etats-Unis sont très attachés à leur partenariat étroit avec la Turquie comme avec Israël et nous accordons une grande importance à la restauration de relations positives entre eux afin de consolider la paix et la sécurité dans la région", est-il écrit. L'appel d'une trentaine de minutes a été passé dans le véhicule qui conduisait Benyamin Nétanyahou et le président américain jusqu'à l'avion du second, sur l'aéroport de Tel Aviv, d'où il devait partir pour la Jordanie , a-t-on appris de sources américaines.
"Tous deux sont convenus de normaliser les relations entre les deux pays, y compris le retour des ambassadeurs", selon un communiqué officiel israélien. "Le premier ministre Nétanyahou a présenté ses excuses au peuple turc pour toute erreur ayant pu conduire à la perte de vies et accepté l'indemnisation" des victimes, assurant que "les résultats tragiques de la flottille du "Mavi Marmara" n'étaient pas intentionnels", selon le texte. M. Erdogan a accepté ces excuses "au nom du peuple turc" et les deux dirigeants "sont convenus de la conclusion d'un accord pour une indemnisation" des familles des victimes, selon un communiqué de ses services .
Déjà tendues depuis l'opération israélienne meurtrière "Plomb durci" dans la bande de Gaza (décembre 2008-janvier 2009), les relations entre la Turquie et Israël, alliés stratégiques dans les années 1990, se sont brutalement dégradées le 31 mai 2010 lors de l'assaut israélien contre une flottille tentant de briser le blocus israélien du territoire palestinien gouverné par le Hamas. Neuf passagers du navire turc Mavi Marmara avaient été tués, provoquant une crise diplomatique entre les deux pays. Ankara a abaissé le niveau de sa représentation diplomatique en Israël, dont il a expulsé l'ambassadeur, et suspendu la coopération militaire.
...
</TXT>
</DOC>

FIGURE B.6 – Alignement des deux médoides : ”Israel et Turquie”

B.7 Septième paire de clusters : ”Afghanistan”

Valeur de comparabilité entre les deux clusters alignés : 0,47923437154785387

<DOC>
<DOCID>b32ab408ce86943d951683d6feb2ef8e</DOCID>
<PUBDATE>Sun Apr 07 16:54:28 CEST 2013</PUBDATE>
<CURDATE>Sun Apr 07 19:30:09 CEST 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>
<ITEMURI>http://www.bbc.co.uk/news/world-asia-22058455#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Afghan children 'killed by Nato'</TITLE>
<DESC>Up to 12 civilians - 10 children and two women - are reported to have been killed in a Nato air strike in Kunar province in eastern Afghanistan.</DESC>
<TXT>Q&A: Foreign forces Up to 12 civilians - 10 children and two women - are reported to have been killed in a Nato air strike in eastern Afghanistan. A further six women are believed to have been injured in the incident in Shigal district, Kunar province. Villagers and officials told the BBC that the casualties were inside their homes when they died. Nato confirmed that "fire support" was used in Shigal after a US civilian adviser died in a militant attack. It did not have any reports of civilian deaths, but photographs apparently sent from the scene to international news agencies appeared to show the bodies of several dead young children, surrounded by Afghan villagers. A local official said eight Taliban insurgents had also died in the air strike on Saturday, which is reported to have caused the roofs of several houses in three villages to collapse. Analysis Bilal Sarwary BBC News, Kabul The narrow and mountainous valley of Shultan lies 30km (20 miles) away from the provincial capital of Asadabad, right on the border with Pakistan's Bajaur tribal agency. The area is covered with dense forest, offering the perfect cover for insurgents. Afghan intelligence officials in Kunar say Afghan and foreign fighters have a big presence in the area and often launch attacks against Afghan and international forces in Afghanistan, and against Pakistani military positions across the border. The area has been the site of intense fighting between Taliban and US/Afghan forces for the last 10 years, with US-led forces often carrying out operations in the valley targeting Afghan Taliban and foreign fighters with al-Qaeda backing. The Afghan government has struggled to exert its control in this strategic district despite several major American offensives, as the militants keep re-grouping. Afghan counter-terrorism officials in the province say foreign fighters have been training local fighters in the area for quite some time, and their presence has become a major threat for the security of Kunar province. He said the strikes were called in to support a major operation by US and Afghan government forces targeting senior Taliban commanders and a local weapons cache. Tribal elder Haji Malika Jan told the BBC: "The fighting started yesterday morning [Saturday] and continued for at least seven hours. There were heavy exchanges between both sides. "The area is very close to the Pakistani border and there are hundreds of local and foreign fighters, mostly Pakistanis, in the area." In a statement, the Nato-led International Security Assistant Force (Isaf) said: "We are aware of an incident yesterday in Kunar province in which insurgents engaged an Afghan and coalition force.
…
</TXT>
</DOC>

<DOC>
<DOCID>fcfd05837ed025622575d05dce3b32db</DOCID>
<PUBDATE>Sun Apr 07 21:12:26 CEST 2013</PUBDATE>
<CURDATE>Mon Apr 08 08:16:24 CEST 2013</CURDATE>
<FEEDURI>http://liberation.fr.feedsportal.com/c/32268/fe.ed/rss.liberation.fr/rss/10/</FEEDURI>
<ITEMURI>http://liberation.fr.feedsportal.com/c/32268/f/606159/s/2a6dcb7c/l/0L0Sliberation0Bfr0Cmonde0C20A130C0A40C0A70Cafghanistan0Eun0Ebombardement0Ede0El0Eotan0Etue0Edix0Eenfants0I894270A0Dxtor0Frss0E450A/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Afghanistan: Karzaï condamne un bombardement de l’Otan ayant tué 11 enfants</TITLE>
<DESC>Articles en rapport Violences au Caire entre Coptes et musulmans La Corée du Nord prépare un double essai missile/bombe nucléaire, indique Séoul Afghanistan: Karzaï condamne un bombardement de l’Otan ayant tué 11 enfants Monténégro: le président sortant et son adversaire revendiquent la victoire Afghanistan: 9 passagers d’un bus tués par une bombe artisanale</DESC>
<TXT>Par AFP
Libération
Le président Hamid Karzaï a «fermement condamné» dimanche un bombardement de l’Otan samedi dans l’est de l’Afghanistan qui a tué onze enfants afghans, malgré de multiples injonctions de sa part à cesser les attaques aériennes sur des zones d’habitations.
«Tout en condamnant l’utilisation de civils comme boucliers, le président a dénoncé toute opération qui cause la mort de civils», peut-on lire dans un communiqué de la présidence afghane, ajoutant qu’une «délégation» se rendrait sur les lieux pour enquêter.
L’Isaf, la force de l’Otan en Afghanistan, qui indiquait jusqu’alors que «jusqu’à dix femmes et enfants avaient été blessés mais non pas tués», selon l’un de ses porte-parole dimanche après-midi, a déclaré quelques heures plus tard qu’elle «prenait acte des informations sur la mort de dix enfants», selon un autre de ses communicants.
«Nous rassemblons les faits pour comprendre ce qui s’est produit. Nous prenons chaque perte civile très au sérieux», a poursuivi cet autre porte-parole.
Un premier bilan, confirmé par trois responsables de la province du Kunar, l’un des bastions talibans de l’Est du pays où l’incident s’est produit, faisait état de 10 enfants morts, auxquels s’ajoutait la mort d’une femme, selon l’une de ces sources.
Le bombardement s’est produit alors qu’un combat intense opposait des troupes afghanes et américaines à des insurgés talibans dans le district de Shigal, selon plusieurs sources afghanes et l’Isaf.
«Avant le bombardement, un Américain a été tué et quatre membres des forces de sécurité afghanes ont été blessés dans une attaque des insurgés», a commenté Wasifullah Wasifi, le porte-parole du gouvernement provincial du Kunar.
La mort d’un civil américain dans l’Est afghan a été annoncée samedi par les forces armées américaines par communiqué, sans plus de précisions. Le porte-parole de l’Isaf a confirmé à l’AFP qu’il s’agissait bien du même incident.
«On nous tirait dessus depuis plusieurs maisons de la zone. Un Américain a été tué et plusieurs de nos hommes blessés. La force de la coalition a répondu par un bombardement», a expliqué une source sécuritaire afghane présente pendant l’opération.
…
</TXT>
</DOC>

FIGURE B.7 – Alignement des deux médoides : ”Afghanistan”

B.8 Huitième paire de clusters : ”Chypre”

Valeur de comparabilité entre les deux clusters alignés : 0,4783106871667471	
<DOC> <DOCID>db6ad50538b90b63f7d2c01f800036a2</DOCID> <PUBDATE>Fri Mar 22 09:03:21 CET 2013</PUBDATE> <CURDATE>Fri Mar 22 09:07:31 CET 2013</CURDATE> <FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI> <ITEMURI>http://www.bbc.co.uk/news/world-europe-21890015#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Cyprus MPs due to vote on new plan</TITLE> <DESC>Cypriot MPs are to vote on new measures to raise the funds the country needs to secure vital bailout before Monday's deadline.</DESC> <TXT>Cyprus MPs due to vote on new plan to secure bailout The country is in a race against time to secure a bailout How damaging is it? MPs in Cyprus are due to begin voting on a series of bills that aim to raise the funds the country needs to secure an international bailout. The country is in a race against time after the European Central Bank gave Cyprus until Monday to find the money. If it does not, liquidity to the country's banks could be cut off and they could collapse. Talks on new Russian financial aid for Cyprus have failed, Russia's finance minister has confirmed. Anton Siluanov, speaking after talks with his Cypriot counterpart Michael Sarris, said Russian investors were not interested in Cyprus' offshore gas reserves. Mr Sarris, who has now left Moscow, had reportedly been seeking some 5bn euros (4.3bn; $6.5bn) in return for bonds in energy and other assets, according to a report by Bloomberg news agency . The country is surviving on a lifeline from the European Central Bank End Quote Cyprus: Mounting EU pressure Cyprus needs to find 5.8bn euros to qualify for a 10bn-euro bailout loan from the EU and International Monetary Fund (IMF). Parliamentarians flatly rejected a plan to tax bank deposits earlier this week. European Commission chief Jose Manuel Barroso, who is also in Moscow this week, said he was "very concerned" at recent developments in Cyprus but added: "We have in the past solved bigger problems." "I hope that this time a solution can also be found." Critical time MPs could be seen arriving at parliament in the capital, Nicosia, passing protesters. Eurozone finance ministers have said they are "ready to discuss with the Cypriot authorities a draft new proposal", which they expect "the Cyprus authorities to present as rapidly as possible". Political leaders discussed the options with President Nicos Anastasiades on Thursday, and the package was then discussed by the cabinet. But MPs said they needed more time to study the nine bills that make up the draft legislation. At the scene Mark Lowen BBC News, Nicosia The fear is catching. Outside cash machines, queues grew on Thursday with savers worried about their money - particularly in the two most troubled banks. Rumours that they might be closed altogether only sparked more concern. Elsewhere, businesses are demanding payment in cash, turning away credit cards for fear they won't get their money. It's led to a drop in business - with customers staying away. This is the price Cyprus is paying for its current crisis. And the race is on to resolve it by next Monday when the European Central Bank says it will turn off its emergency funds. … </TXT> </DOC>	<DOC> <DOCID>44d3817e0369e233f92e4f63ad69914d</DOCID> <PUBDATE>Fri Mar 22 16:40:51 CET 2013</PUBDATE> <CURDATE>Sat Mar 23 10:12:47 CET 2013</CURDATE> <FEEDURI>http://www.courrierinternational.com/rss/all/rss.xml</FEEDURI> <ITEMURI>http://www.courrierinternational.com/article/2013/03/22/le-kremlin-joue-la-montre</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>CHYPRE-RUSSIE • Le Kremlin joue la montre</TITLE> <DESC>C'est un véritable bras de fer qui oppose désormais Moscou et Bruxelles et il n'y aura pas d'issue simple à la crise chypriote. La Russie a beaucoup à perdre et peu à gagner, et va faire monter les enchères.</DESC> <TXT>VU DU PORTUGAL - Chypre : la politique des fous CHYPRE-RUSSIE • Le Kremlin joue la montre C'est un véritable bras de fer qui oppose désormais Moscou et Bruxelles et il n'y aura pas d'issue simple à la crise chypriote. La Russie a beaucoup à perdre et peu à gagner, et va faire monter les enchères. Votre message Dessin de Bojesen Dans la partie délicate qui l'oppose à l'Union européenne, le Kremlin vient de riposter. L'UE avait bien failli persuader Chypre d'instaurer une taxe confiscatoire qui aurait porté préjudice aux sociétés publiques russes. Il n'est désormais plus possible d'envisager de sortie de crise simple ou sans trop de pertes. Et le manque d'empressement de la Russie à proposer une nouvelle aide à Chypre est un moyen de pression commode sur l'Europe. Les fonctionnaires européens semblent en avoir pris conscience. Hier, la BCE a déclaré qu'elle allait cesser dès lundi [25 mars] de fournir l'argent permettant de couvrir les dépenses courantes aux banques de l'île. Ensuite, la crise devrait franchir un nouveau seuil. Chypre espère une aide de Moscou en échange de certains "avantages" pour l'économie russe. Michalis Sarris, le ministre chypriote des Finances, a souligné qu'il n'était pas venu à Moscou [mercredi 20 mars] les mains vides. Il est venu exposer les possibilités qui s'ouvriraient à la Russie si elle aidait son île. La liste comprend une participation au système bancaire et à l'exploitation du gaz. Ce que Chypre demande, ce n'est absolument pas un crédit, mais la conclusion d'une transaction censée être mutuellement profitable : "Malheureusement, un crédit ne serait d'aucun secours. Nous devons trouver des moyens de coopérer dans plusieurs secteurs afin d'obtenir des investissements, et non des crédits", a déclaré le ministre. Mais la Russie semble maintenant vouloir prendre son temps. Andreï Kostine, le directeur de la banque VTB [Vnechtorgbank], principale victime de la crise chypriote, a annoncé le 21 mars que son établissement n'était pas du tout intéressé par l'achat d'actifs bancaires de l'île : "Sur place, il y a deux banques dans une situation critique qui ont besoin d'être assainies. Il serait absurde de prétendre que nous aurions un intérêt là-dedans. Notre seul intérêt, c'est de retrouver au plus vite la faculté d'effectuer les paiements et de gérer les comptes de nos clients". Et il ajoute que sa banque va devoir "arrêter son activité et quitter purement et simplement le marché chypriote" en cas de "décisions violant le droit, dictées par la politique". Le paradis fiscal chypriote … </TXT> </DOC>

FIGURE B.8 – Alignement des deux médoides : ”Chypre”

B.9 Neuvième paire de clusters : ”Election Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,4751787785591162

<DOC>
<DOCID>5d6d9fe72eb51be8aa78b3cddba516d4</DOCID>
<PUBDATE>Sun Mar 03 02:25:23 CET 2013</PUBDATE>
<CURDATE>Mon Mar 04 09:50:14 CET 2013</CURDATE>
<FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI>
<ITEMURI>http://www.globaltimes.cn/content/765328.shtml</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Iran says Assad to run for president in 2014 Syrian elections</TITLE>
<DESC>Iranian Foreign Minister Ali-Akbar Salehi said here Saturday that Bashar al-Assad is currently the " legal" president of Syria and will take part in the presidential elections in 2014.</DESC>
<TXT>Iran says Assad to run for president in 2014 Syrian elections Xinhua | 2013-3-3 9:25:23 Print Iranian Foreign Minister Ali-Akbar Salehi said here Saturday that Bashar al-Assad is currently the " legal" president of Syria and will take part in the presidential elections in 2014. Salehi made the remarks at a joint press conference with his visiting Syrian counterpart Walid al-Moallem in Tehran. Syria, like any other country, has a president by elections, said Salehi, adding that by the next elections in Syria, Assad is the legal president of the Syrians and this is the official stance of the Islamic republic. In the next presidential elections of Syria, anybody who is elected by the people will be the president of the Syrians for the next term, said the Iranian foreign minister. Also, Salehi said that no country, except Syria, would be allowed to decide for the Syrians. No country outside has the right to make decisions for Syria, Salehi said, adding that nobody can inflict its will on the Syrians using force and arms. Salehi said there is no military solution to the Syrian issue, and the first and foremost demand of the Islamic republic is to end the violence in the Arab state. He said foreign intervention in the Middle-Eastern Arab state is not acceptable, adding that Iran supports the efforts by the joint special representative of the UN and Arab League for Syria, Lakhtar Brahimi, to bring peace to the war-torn state. Iran and Syria has a long history of relations and Iranians will never forget Syria's support for Iran during the eight-year Iran-Iraq War (1980-1988), said Salehi. Iran is the major regional ally of the Syrian government in its conflict with armed opposition groups. For his part, al-Moallem said Saturday that the talks between the Syrian government and the opposition groups have started and in order for the talks to bear "successful" results, the violence in the country should be stopped. Joint efforts should be done to stop violence, he said, emphasizing that those who back "terrorists" with arms and finance should give up their supports. "We should make efforts to end bloodshed in Syria," and those who avoid dialogue must be held accountable, he maintained. The current dialogue in Syria displays that the crisis in the country would be resolved "only through political means," said al- Moallem. He further stressed that the Syrian Army is duty-bound to defend the country against "terrorists."
…
</TXT>
</DOC>

<DOC>
<DOCID>d96575cd8655dd4effb5296ddd9c1091</DOCID>
<PUBDATE>Sat Mar 02 13:57:05 CET 2013</PUBDATE>
<CURDATE>Sat Mar 02 15:10:22 CET 2013</CURDATE>
<FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI>
<ITEMURI>http://www.romandie.com/news/n.asp?n=_Assad_president_legitime_de_Syrie_jusqu_a_l_election_de_2014_19020320131357.asp</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Assad président légitime de Syrie jusqu'à l'élection de 2014</TITLE>
<DESC>TEHERAN - Bachar al-Assad restera le président légitime de Syrie jusqu'à la prochaine élection prévue en 2014, a affirmé samedi à Téhéran le ministre iranien des Affaires étrangères Ali Akbar Salehi dont le pays est l'un des...</DESC>
<TXT>Tweet
Assad président légitime de Syrie jusqu'à l'élection de 2014
TEHERAN - Bachar al-Assad restera le président légitime de Syrie jusqu'à la prochaine élection prévue en 2014, a affirmé samedi à Téhéran le ministre iranien des Affaires étrangères Ali Akbar Salehi dont le pays est l'un des principaux alliés du pouvoir à Damas.
M. Salehi, lors d'un point de presse commun avec son homologue syrien Walid Mouallem, a aussi apporté son soutien à l'appel au dialogue avec l'opposition armée lancé par le régime, tout en réaffirmant que ce dernier n'avait pour l'instant pas d'autre choix que de continuer la lutte contre la rébellion.
Le ministre iranien a redit la position officielle de l'Iran qui est que M. Assad demeurera le président légitime jusqu'à la prochaine élection présidentielle en Syrie prévue en 2014.
En vue d'un règlement du conflit en Syrie qui a fait plus de 70.000 morts depuis près de deux ans selon l'ONU, les Occidentaux, plusieurs pays arabes, la Turquie ainsi que l'opposition syrienne appellent à un départ de M. Assad du pouvoir.
Ce dernier est resté intraitable en affirmant que ses troupes continueraient le combat jusqu'à en venir à bout des rebelles, qu'il assimile à des terroristes.
La crise syrienne n'a pas de solution militaire et la seule solution est un dialogue entre le pouvoir et l'opposition, a ajouté M. Salehi après une rencontre avec M. Mouallem, arrivé le matin à Téhéran, six jours après une visite à Moscou, l'autre allié du régime syrien.
Dans cet esprit, l'appel à un dialogue avec l'opposition armée lancé lundi pour la première fois par M. Mouallem lors de sa visite à Moscou est un pas positif, a-t-il estimé.
Toutefois, a dit le ministre iranien, personne ne peut demander au pouvoir syrien d'abandonner les armes, car il n'a pas d'autre choix que de combattre les mercenaires pour rétablir le calme.
M. Mouallem a dénoncé de son côté l'annonce jeudi par Washington de l'octroi de 60 millions de dollars à l'opposition syrienne et d'une aide directe non létale à la rébellion. Nous ne comprenons pas cette initiative alors que cette opposition tue des gens, a déclaré le ministre syrien.
Il a également appelé à faire pression sur la Turquie et le Qatar, accusés par la Syrie d'aider la rébellion.
Téhéran prône un dialogue national entre le pouvoir et ses opposants mais dénonce le soutien aux groupes d'opposition armée.
(©AFP / 02 mars 2013 13h56)

</TXT>
</DOC>

FIGURE B.9 – Alignement des deux médoides : ”Election Syrie”

B.10 Dixième paire de clusters : ”Liban”

Valeur de comparabilité entre les deux clusters alignés : 0,4697332316685679

<DOC>
<DOCID>4f97982b41c8f3cc0e729058312bff5d</DOCID>
<PUBDATE>Fri Mar 22 23:24:31 CET 2013</PUBDATE>
<CURDATE>Sat Mar 23 10:23:30 CET 2013</CURDATE>
<FEEDURI>http://www.abc.net.au/news/feed/52278/rss.xml</FEEDURI>
<ITEMURI>http://www.abc.net.au/news/2013-03-23/lebanon27s-pm-quits-over-election-impasse/4590140</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Lebanon's PM quits over election impasse</TITLE>
<DESC>The prime minister of Lebanon has resigned after his cabinet failed to agree on forthcoming elections, in a decision that automatically brings down the government.</DESC>
<TXT>Lebanon's PM quits over election impasse Posted March 23, 2013 09:24:31 Map: Lebanon The prime minister of Lebanon has resigned after his cabinet failed to agree on forthcoming elections, in a decision that automatically brings down the government. In a speech aired live on television, Najib Mikati said hoped his departure would be "an impetus for leaders to shoulder their responsibilities". "I announce the resignation of the government, hoping that this will open the way for the major political blocs to take responsibility and come together to bring Lebanon out of the unknown," he said. After months of wrangling, the parliament has been unable to agree on a law to govern elections slated for June. The cabinet was divided on two issues, including the formation of a commission to oversee the ballot. The government has held off on agreeing on the membership of the commission over fears it would ensure that elections are held on the basis of a decades-old electoral law.' Mr Mikati is said to favour the existing law. It gives his Sunni community and the Druze disproportionate strength in parliament, but is opposed by Christians who say it fails to give them representative weight. There was also disagreement over Mr Mikati's request to extend the tenure of the country's police chief. 'Salvation government' Mr Mikati hopes a new unity government can now be formed to save the country from going over the brink. "A national salvation government in which all Lebanese political forces are represented, in order to save the nation and deal with regional developments with a collective spirit of responsibility," he said. Mr Mikati said he was willing to resign last year, after a car bombing that killed the police intelligence chief, but president Michel Sleiman rejected it and he stayed in office. He became prime minister in 2011 after five months of negotiations, positioning himself as a political moderate able to deal with all political parties. He headed a government dominated by the so-called March 8 coalition, made up of Hezbollah and its allies, and drew fire from Sunnis who accused him of betraying his community and siding with the Syrian- and Iranian-backed group. The resignation throws Lebanon into new uncertainty, and comes as the violence in Syria increasingly affects the country. The conflict has exacerbated existing tensions in Lebanon's multi-confessional population and violence between opponents and supporters of Syrian president Bashar al-Assad has already spilled over the border. Damascus has warned Beirut against allowing fighters and weapons to enter the country. ABC/AFP</TXT>
</DOC>

<DOC>
<DOCID>c6e255a1042001bde97aef76e9bfb1c8</DOCID>
<PUBDATE>Fri Mar 22 19:55:00 CET 2013</PUBDATE>
<CURDATE>Sat Mar 23 10:16:23 CET 2013</CURDATE>
<FEEDURI>http://www.lapresse.ca/rss/179.xml</FEEDURI>
<ITEMURI>http://rss.lapresse.ca/c/33663/f/607972/s/29e2633a/l/0L0Slapresse0Bca0Cinternational0Cmoyen0Eorient0C20A130A30C220C0A10E46338660Ele0Epremier0Eministre0Elibanais0Edemissionne0Bphp/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Le premier ministre libanais démissionne</TITLE>
<DESC>Le premier ministre libanais Najib Mikati a créé la surprise en annonçant vendredi la démission de son cabinet, en place depuis juin 2011, et en préconisant la mise sur pied d'un «gouvernement de salut national».</DESC>
<TXT>> Le premier ministre libanais démissionne
Le premier ministre libanais démissionne
Le premier ministre libanais Najib Mikati.
PHOTO JAMAL SAIDI, REUTERS
Agence France-Presse
Beyrouth
Le premier ministre libanais Najib Mikati a créé la surprise en annonçant vendredi la démission de son cabinet, en place depuis juin 2011, et en préconisant la mise sur pied d'un «gouvernement de salut national».
Cette démission ouvre une période d'incertitude dans ce pays au moment où son voisin, la Syrie, ancienne puissance tutélaire, est ravagée par une guerre civile dont les répercussions se font sentir au Liban.
«J'annonce la démission du gouvernement en espérant que cela fera prendre conscience aux principaux blocs politiques au Liban de la nécessité d'assumer leurs responsabilités et de faire preuve de cohésion pour éviter l'inconnu au Liban», a-t-il dit aux journalistes à l'issue d'un conseil des ministres marqué par de profondes divisions.
Ce sunnite de 57 ans, qui dirigeait un cabinet dominé par le mouvement chiite Hezbollah, a appelé à la «formation d'un gouvernement de salut national où toutes les forces politiques libanaises seront représentées afin de sauver la patrie et suivre les événements régionaux dans un grand esprit de responsabilité collective».
Paradoxalement, ce n'est pas le conflit syrien, dont le Liban avait décidé dès le début de se distancier, qui l'a poussé à jeter l'éponge, mais des problèmes intérieurs.
M. Mikati a confié à la presse avoir pris sa décision en raison des divergences au sein du cabinet sur l'organisation des élections législatives en juin prochain et sur la prolongation du mandat du chef des Forces de sécurité intérieure (FSI, police), le général Achraf Rifi, dont le mandat se termine à la fin du mois.
Les partis chrétiens voudraient changer la loi électorale datant de 1960 car ils l'estiment défavorable à leur communauté. Selon eux, les circonscriptions actuelles favorisent les musulmans qui, plus nombreux, peuvent choisir les députés chrétiens qui leur conviennent.
Le Liban compte un tiers de chrétiens, un tiers de sunnites et un tiers de chiites, mais le Parlement compte 128 députés répartis en moitié de chrétiens et moitié de musulmans.
…
</TXT>
</DOC>

FIGURE B.10 – Alignement des deux médoides : ”Liban”

C

Dix premières paires de clusters obtenues sur la base du **Tri simultané**

Nous présentons ici dix premières paires de clusters alignés, représentées par leurs médoides, obtenues sur la base du **Tri simultané** des plus fortes comparabilités pour le corpus Flux RSS.

C.1 Premières paire de clusters : ”Syrie et Liban”

Valeur de comparabilité entre les deux clusters alignés : 0,5960093384724747

<DOC>
<DOCID>63a055c59873577c7a5fec38057e698a</DOCID>
<PUBDATE> null </PUBDATE>
<CURDATE>Sat Mar 16 10:27:03 CET 2013</CURDATE>
<FEEDURI>http://www.chinadaily.com.cn/rss/world_rss.xml</FEEDURI>
<ITEMURI>http://www.chinadaily.com.cn/world/2013-03/15/content_16310532.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Syria warns against rebels sneaking from Lebanon</TITLE>
<DESC>Syrian foreign ministry stressed Thursday that the Syrian army is practicing the "highest levels of self- restraint" not to hit the rebels' positions on the Lebanese side of the borders with Syria.</DESC>
<TXT>Syria warns against rebels sneaking from Lebanon Updated: 2013-03-15 09:52 Large Medium Small DAMASCUS - Syrian foreign ministry stressed Thursday that the Syrian army is practicing the "highest levels of self- restraint" not to hit the rebels' positions on the Lebanese side of the borders with Syria. "Over the past 36 hours, the armed terrorists in large numbers have sneaked into Syria at the border villages of Mathoume, Ain al- Sharaa, al-Jousieh and Talkalakh," the ministry said in a statement addressing the Lebanese foreign ministry. The Syrian troops have clashed with the infiltrators and the clashes are still ongoing, the ministry said, adding that many of the assailants have been killed while some others have fled back to Lebanon. The ministry noted that the armed groups' positions on the Lebanese side are seeable by the Syrian troops, which "has been practicing the highest levels of self-restraint." Syria hopes "the Lebanese competent authorities will exert efforts to control its borders with Syria," according to the statement. "Syria expects that the Lebanese side will prevent those (rebels) from using the borders as routes, because they are targeting the security of the Syrian people and infringing upon the sovereignty of Syria," the ministry said. The ministry pointed out that the flow of gunmen and arms have notably stepped up since March 12, charging that those infiltrators are receiving support from Lebanon. The Syrian government has for long complained about the flow of arms and cash from surrounding countries, including Lebanon, whose northern city Tripoli is packed with Islamists who oppose the government of Syrian President Bashar al-Assad. Moreover, some leaked phone recordings made public recently by Lebanese TVs have incriminated some Lebanese officials that appeared coordinating the flow of weapons and money to the rebels in Syria. 8.03K</TXT>
</DOC>

<DOC>
<DOCID>1e778be2433287d86fb3d7e35eb22d17</DOCID>
<PUBDATE>Fri Mar 15 10:12:16 CET 2013</PUBDATE>
<CURDATE>Sat Mar 16 10:16:32 CET 2013</CURDATE>
<FEEDURI>http://www.courrierinternational.com/rss/all/rss.xml</FEEDURI>
<ITEMURI>http://www.courrierinternational.com/breve/2013/03/15/damas-menace-de-frapper-le-liban</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>LIBAN • Damas menace de frapper le Liban</TITLE>
<DESC>"Damas menace 'officiellement' le Liban de bombarder son territoire", titre le quotidien libanais.</DESC>
<TXT>Envoyer
Votre message
"Damas menace 'officiellement' le Liban de bombarder son territoire", titre le quotidien libanais. Selon le ministère des Affaires étrangères syrien, des combats opposeraient l'armée régulière syrienne à des "bandes armées qui tentent d'infiltrer la Syrie à partir du Liban", ce qui pourrait pousser Damas à les frapper jusque sur le territoire libanais. Le Conseil de sécurité des Nations unies a réagi en se disant "très inquiet" des "incidents frontaliers répétés" entre les deux pays : tirs par-dessus la frontière, incursions, trafic d'armes.
<<
</TXT>
</DOC>

FIGURE C.1 – Alignement des deux médoides : ”Syrie et Liban”

C.2 Deuxième paire de clusters : ”Syrie et Iraq”

Valeur de comparabilité entre les deux clusters alignés : 0,5617286025467252

<DOC>
<DOCID>be2e889997fb981cf8c544f99f52d798</DOCID>
<PUBDATE>Tue Mar 05 01:38:26 CET 2013</PUBDATE>
<CURDATE>Tue Mar 05 19:18:44 CET 2013</CURDATE>
<FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI>
<ITEMURI>http://www.globaltimes.cn/content/765822.shtml</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>At least 20 Syrian soldiers killed in western Iraq</TITLE>
<DESC>At least 20 Syrian soldiers were killed and some seven Iraqi soldiers wounded in an attack by gunmen on an Iraqi army convoy, which was escorting a bus carrying Syrian soldiers in Iraq's Anbar province on Monday.</DESC>
<TXT>At least 20 Syrian soldiers killed in western Iraq Xinhua | 2013-3-5 8:38:26 Print At least 20 Syrian soldiers were killed and some seven Iraqi soldiers wounded in an attack by gunmen on an Iraqi army convoy, which was escorting a bus carrying Syrian soldiers in Iraq's Anbar province on Monday. The attack occurred near the city of Rutba, some 375 km west of Baghdad, when unidentified gunmen attacked a convoy of military vehicles escorting a bus carrying Syrian soldiers who entered Iraq after the Syrian rebels seized al-Yaroubiyah border crossing point with Iraq's northern province of Nineveh, the source from Anbar police said on condition of anonymity. The convoy was trying to transfer the Syrian soldiers from Nineveh province to Syria through al-Walid border crossing point in Iraq's western province of Anbar, the source said. The attackers burnt the bus carrying the Syrian soldiers and destroyed two Iraqi military vehicles, the source added. On Saturday, the Syrian opposition forces took control of al- Yaroubiyah border crossing point after fierce clashes with the Syrian army, forcing some Syrian soldiers to cross into Iraq. On Sunday, the Iraqi authorities in Nineveh said several wounded Syrian soldiers were allowed to enter Iraq to receive treatment for their wounds at the Iraqi hospitals. The Iraqi authorities said the soldiers were admitted to hospitals for humanitarian reasons. Iraq has a borderline of around 600 kilometers with Syria. The border crossings had witnessed closure in history when crisis intensified on either side.</TXT>
</DOC>

<DOC>
<DOCID>1ca9388bddb37313f52ee48708d399c6</DOCID>
<PUBDATE>Tue Mar 05 06:00:00 CET 2013</PUBDATE>
<CURDATE>Tue Mar 05 19:25:04 CET 2013</CURDATE>
<FEEDURI>http://www.ledevoir.com/rss/section/international.xml?id=76</FEEDURI>
<ITEMURI>http://www.ledevoir.com/international/actualites-internationales/372446/48-soldats-syriens-sont-tues-en-irak</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Bagdad accusée d’ingérence dans le pays voisin - 48 soldats syriens sont tués en Irak</TITLE>
<DESC>Quarante-huit soldats syriens qui s’étaient réfugiés en Irak pendant le...</DESC>
<TXT>5 mars 2013 International / Actualités internationales
Bagdad accusée d’ingérence dans le pays voisin - 48 soldats syriens sont tués en Irak
Tweet
Importante victoire des rebelles
Les rebelles ont remporté lundi leur victoire la plus importante depuis le début de la révolte en Syrie, en s’emparant d’un chef-lieu de province, au moment où Washington avalisait implicitement des livraisons d’armes par des pays du Golfe.
À Raqa, chef-lieu de la province éponyme, « les rebelles contrôlent presque entièrement la ville », a annoncé l’Observatoire syrien des droits de l’Homme (OSDH).
Les combats se poursuivaient et la ville était visée par des raids de l’aviation du régime, indique la même source.
« Dans les prochaines heures, Raqa sera la première capitale de province à être hors du contrôle du régime », a déclaré Rami Abdel Rahmane, directeur de l’OSDH.
Quarante-huit soldats syriens qui s’étaient réfugiés en Irak pendant le...
Quarante-huit soldats syriens qui s’étaient réfugiés en Irak pendant le week-end et neuf Irakiens ont été tués lundi dans une embuscade en territoire irakien, selon le ministère de la Défense à Bagdad.
Cette embuscade dans la province d’Anbar, 24 heures après qu’une importante composante de l’opposition syrienne a accusé l’Irak d’ingérence en Syrie, risque d’entraîner ce pays dans la guerre civile qui déchire son voisin syrien, ce que Bagdad s’est engagé à ne pas laisser faire.
L’embuscade a été tendue par « un groupe terroriste qui s’est infiltré en territoire irakien depuis la Syrie », a indiqué le ministère dans un communiqué, dénonçant « une attaque contre la souveraineté de l’Irak, son territoire, sa dignité et une violation claire des droits de la personne, [les soldats] étant blessés et non armés ».
Neuf gardes irakiens escortant les soldats syriens ont également été tués dans l’embuscade, selon la même source.
Bagdad va néanmoins résister aux tentatives de propager le conflit syrien en Irak, a promis lundi le porte-parole du premier ministre irakien.
« Cela confirme nos craintes sur le fait que certains tentent de propager la crise syrienne en Irak, mais nous ferons face à ces tentatives d’où qu’elles viennent avec toute notre force », a prévenu Ali Moussaoui, porte-parole du chef du gouvernement, Nouri al-Maliki.
…
</TXT>
</DOC>

FIGURE C.2 – Alignement des deux médoides : ”Syrie et Irak”

C.3 Troisième paire de clusters : ”Querre civile en Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,5581265658090165

<DOC>
<DOCID>ba614d6f31f0cd5cd0b7fc0da19fd611</DOCID>
<PUBDATE>Thu Mar 28 14:26:07 CET 2013</PUBDATE>
<CURDATE>Fri Mar 29 08:45:30 CET 2013</CURDATE>
<FEEDURI>http://rss.cnn.com/rss/edition_world.rss</FEEDURI>
<ITEMURI>http://rss.cnn.com/~r/rss/edition_world/~3/EZUXDyQoIjs/index.html</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Deaths reported at Damascus college</TITLE>
<DESC>As many as 12 people were killed Thursday when mortars struck a college in Syria's capital, the government said.</DESC>
<TXT>Mortar shells hit Damascus college; deaths reported By Joe Sterling and Hamdi Alkhshali, CNN March 29, 2013 -- Updated 0100 GMT (0900 HKT) Syrians carry the body of a Syrian army soldier during a funeral ceremony in Idlib province on Tuesday, March 19. Tensions in Syria flared in March 2011, escalating into a civil war that still rages today. This gallery contains the most compelling images taken since the start of the conflict. Syrian rebels take position in Aleppo, the largest city in the country, on March 11. Syrian men search for their relatives amongst the bodies of civilians executed and dumped in the Quweiq River on March 11. A Free Syrian Army fighter looks back as smoke rises during fighting between rebel fighters and forces loyal to Syria's President Bashar al-Assad on the outskirts of Aleppo on Saturday, March 2. Residents read Shaam News newspapers published by the Free Syrian Army in Aleppo on March 2. A member of the Free Syrian Army reacts to the death of a comrade who was killed in fighting, at Bustan al Qasr cemetery in Aleppo on Friday, March 1. A rebel fighter throws a home-made grenade at Syrian government forces in Aleppo on February 16. A member of the Free Syrian Army stands with his weapon as he looks at a rainbow in Aleppo on February 16. A Syrian woman looks through a bus window in Aleppo on February 14. Free Syrian Army fighters walk through a dust-filled stairwell in Damascus on February 7. A Syrian rebel gestures at comrades from inside a broken armored personnel carrier in Al-Yaqubia on February 6. A rebel fighter throws a hand grenade inside a Syrian Army base in Damascus on February 3. People stand in the dust of a building destroyed in an airstrike in Aleppo, Syria on February 3. Free Syrian Army fighters run as they enter a Syrian Army base during heavy fighting in the Arabeen neighborhood of Damascus on February 3. An unexploded mortar shell fired by the Syrian Army sits lodged in the ground in Damascus on January 25. Fighters from Fateh al Sham unit of the Free Syrian Army fire on Syrian Army soldiers at a check point in Damascus on January 20. A Free Syrian Army fighter walks between buildings damaged during Syrian Air Force strikes in Damascus on January 19. A Syrian rebel fighter tries to locate a government jet fighter in Aleppo on January 18. Syrian rebels launch a missile near the Abu Baker brigade in Albab on January 16. A Syrian boy walks near rubbish next to tents at a refugee camp near the northern city of Azaz on the Syria-Turkey border, on January 8. Syrians look for survivors amid the rubble of a building targeted by a missile in Aleppo on January 7.
…
</TXT>
</DOC>

<DOC>
<DOCID>d59cadd8de57e2631d3573ea2a85dc06</DOCID>
<PUBDATE>Mon Feb 25 20:14:00 CET 2013</PUBDATE>
<CURDATE>Tue Feb 26 16:06:33 CET 2013</CURDATE>
<FEEDURI>http://www.lapresse.ca/rss/179.xml</FEEDURI>
<ITEMURI>http://rss.lapresse.ca/c/33663/f/607972/s/28f1f682/l/0L0Slapresse0Bca0Cinternational0Cdossiers0Ccrise0Edans0Ele0Emonde0Earabe0Csyrie0C20A130A20C250C0A10E46252780Eles0Eviolences0Efont0E920Emorts0Een0Esyrie0Bphp/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Les violences font 92 morts en Syrie</TITLE>
<DESC>Au moins 30 soldats syriens et 23 rebelles ont été tués en 24 heures de combats dans l'ouest de la province d'Alep (nord), où des insurgés ont abattu un hélicoptère de l'armée, a rapporté l'Observatoire syrien des droits de l'homme (OSDH).</DESC>
<TXT>> Les violences font 92 morts en Syrie
Les violences font 92 morts en Syrie
À travers le pays, les violences ont fait lundi au moins 92 morts, selon un bilan provisoire de l'OSDH basée au Royaume-Uni et s'appuyant sur un vaste réseau de militants et de sources médicales civiles et militaires.
PHOTO GORAN TOMASEVIC, ARCHIVES REUTERS
Consultez notre dossier complet sur les soulèvements populaires en Afrique. »
À lire aussi
Agence France-Presse
Beyrouth, Liban
Au moins 30 soldats syriens et 23 rebelles ont été tués en 24 heures de combats dans l'ouest de la province d'Alep (nord), où des insurgés ont abattu un hélicoptère de l'armée, a rapporté l'Observatoire syrien des droits de l'homme (OSDH).
À travers le pays, les violences ont fait lundi au moins 92 morts, selon un bilan provisoire de cette organisation basée au Royaume-Uni et s'appuyant sur un vaste réseau de militants et de sources médicales civiles et militaires.
La plupart des combattants tués dans la province d'Alep sont tombés autour de l'académie de police de Khan al-Assal, théâtre d'affrontements acharnés depuis plusieurs jours, selon la même source.
«Les insurgés ont pris le contrôle d'un immeuble où s'étaient retranchés des soldats du régime après des combats féroces», et ont pris en otage «des dizaines de membres armés des comités populaires pro-régime», a rapporté l'OSDH.
Le quotidien al-Watan, proche du régime, a pour sa part affirmé lundi que «les membres de l'académie de police (avaient) repoussé pour la deuxième journée consécutive les attaques intensives des hommes armés», qui avaient subi des «pertes considérables».
Au nord d'Alep, un hélicoptère s'est écrasé et a explosé au sol après avoir été touché près de l'aéroport militaire de Mingh, assiégée depuis des mois par des rebelles qui l'attaquent avec des roquettes artisanales dans le cadre d'une campagne pour prendre le contrôle des aéroports, selon l'OSDH.
Un habitant de la zone joint par l'AFP a précisé que l'hélicoptère avait été abattu alors qu'il essayait d'atterrir sur la base.
Si les insurgés ont chassé les forces du régime de larges territoires dans les provinces d'Alep, Idlib (nord-ouest), Raqqa (nord) et Hassaké (nord-est), ils restent sous la menace de bombardements de l'aviation ou de tirs de missiles.
…
</TXT>
</DOC>

FIGURE C.3 – Alignement des deux médoides : ”Querre civile en Syrie”

C.4 Quatrième paire de clusters : ”Iran”

Valeur de comparabilité entre les deux clusters alignés : 0,5364881301529452

<DOC>
<DOCID>de9ce0c2557b5f39e5f0f83695fa1d52</DOCID>
<PUBDATE>Tue Feb 26 10:51:50 CET 2013</PUBDATE>
<CURDATE>Tue Feb 26 16:13:17 CET 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>
<ITEMURI>http://www.bbc.co.uk/news/world-middle-east-21572075#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Iranian nuclear talks get under way</TITLE>
<DESC>Talks between Tehran and world powers over Iran's nuclear programme - the first since since July 2012 - open in the Kazakh city of Almaty.</DESC>
<TXT>A new round of talks between world powers and Iran over its controversial nuclear programme has opened in the Kazakh city of Almaty. The discussions are the first since talks in July 2012 ended without a breakthrough. Negotiators from Iran are meeting counterparts from the US, UK, France, China, Russia and Germany - the P5+1. International powers suspect Iran of seeking to develop nuclear weapons - a charge Iran strongly denies. Iran insists its purposes are purely civilian, asserting it needs enriched uranium to make medical isotopes. Since 2010, Iran has been enriching uranium to a level of 20%, an important technological step towards being able to produce more highly enriched weapons-grade material. Analysis James Reynolds BBC News, Almaty By curious practice, each new round of nuclear talks with Iran is held in a new place. So, the well-travelled collection of nuclear negotiators now finds itself amid the wide avenues and frozen pavements of Kazakhstan's largest city. In this round of talks - the first since June 2012 - Western diplomats say that they will present Iran with an updated proposal which they describe as "serious and substantial." In previous rounds, the six major powers at the talks made three specific demands of Iran: the halting of all enrichment work; the shutting down of an underground facility near the holy city of Qom; and the export of the country's supply of medium-level enriched uranium. One Western official says that some form of sanctions relief may also be offered at this round. Previously, Iran has insisted that sanctions need to be lifted before it can consider any concessions of its own. Iran has repeatedly rejected Western calls to stop enriching uranium, insisting it is an inalienable right. Western negotiators at the meeting are expected to offer Iran incentives to compromise. "The offer addresses the international concern on the exclusively peaceful nature of the Iranian nuclear program, but it is also responsive to Iranian ideas," said EU spokesman Michael Mann. "We've put some proposals forward which will hopefully allow Iran to show some flexibility." The proposals might involve easing some of the sanctions which have been imposed on Iran, in return for shutting its Fordo uranium enrichment plant, reports say. Several rounds of sanctions have squeezed Iran's economy, with oil revenue slashed, a currency that has nosedived in value, and growing unemployment. Iran's Press TV said Iran would also offer "a new comprehensive package of proposals", without giving details.
…
</TXT>
</DOC>

<DOC>
<DOCID>dd9e24037bb5f85823a30cf5c1f9a667</DOCID>
<PUBDATE>Thu Feb 21 23:26:39 CET 2013</PUBDATE>
<CURDATE>Fri Feb 22 07:12:31 CET 2013</CURDATE>
<FEEDURI>http://liberation.fr.feedsportal.com/c/32268/fe.ed/rss.liberation.fr/rss/10/</FEEDURI>
<ITEMURI>http://liberation.fr.feedsportal.com/c/32268/f/606159/s/28d0df9e/l/0L0Sliberation0Bfr0Cmonde0C20A130C0A20C210Cnucleaire0El0Eiran0Ea0Einstalle0Edes0Enouvelles0Ecentrifugeuses0I8835420Dxtor0Frss0E450A/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Nucléaire: l'Iran installerait de nouvelles centrifugeuses plus modernes</TITLE>
<DESC>Articles en rapport Affaire Pistorius: décision attendue sur sa demande de libération sous caution Venezuela: l'insuffisance respiratoire de Chavez persiste, évolution défavorable Michelle Obama et son volatile jaune Syrie: près de 60 morts dans l'attentat le plus sanglant à Damas Français enlevés au Cameroun: les recherches se poursuivent au Nigeria</DESC>
<TXT>Par AFP
Libération
L'AIEA a annoncé jeudi que l'Iran avait commencé à installer des centrifugeuses plus modernes sur son site d'enrichissement d'uranium de Natanz, un développement accueilli avec inquiétude par les Occidentaux et par Israël à cinq jours de nouvelles négociations sur le nucléaire iranien.
"Le 6 février 2013, l'agence a observé que l'Iran avait commencé l'installation de centrifugeuses IR-2m" à Natanz, indique un rapport de l'Agence internationale de l'énergie atomique consulté par l'AFP. "C'est la première fois que des centrifugeuses plus avancées que les IR-1 ont été installées" sur le site de Natanz, dans le centre de l'Iran, souligne l'AIEA dans ce document.
Le 13 février, le chef de l'Organisation iranienne de l'énergie nucléaire (OIEA), Fereydoun Abbassi Davani, avait annoncé que l'installation de ces nouveaux équipements avait commencé, ce que l'AIEA n'avait pas confirmé avant ce jeudi.
L'enrichissement est au centre du conflit entre l'Iran et les grandes puissances, qui soupçonnent ce pays de vouloir enrichir de l'uranium jusqu'à 90%, niveau nécessaire pour fabriquer la bombe atomique, ce que les autorités iraniennes démentent régulièrement.
Téhéran affirme enrichir uniquement à des fins civiles - jusqu'à 5% pour produire de l'électricité et 20% pour alimenter son laboratoire de recherche médicale - et revendique son droit à enrichir de l'uranium en tant que signataire du Traité de non-prolifération nucléaire (TNP).
L'Iran est sous le coup de sanctions internationales pour ces activités, qui seront de nouveau au coeur de nouvelles négociations prévues le 26 février après huit mois d'interruption à Almaty (Kazakhstan) avec le Groupe 5+1 (Etats-Unis, Chine, Russie, France, Grande-Bretagne et Allemagne).
Washington a déclaré jeudi que "l'installation de nouvelles centrifugeuses modernes", si elle était confirmée, serait une nouvelle "provocation" de la part de l'Iran.
Ce serait "une nouvelle escalade et une poursuite de la violation des obligations de l'Iran conformément aux résolutions du Conseil de sécurité de l'ONU et de l'AIEA", a ajouté la porte-parole du département d'Etat, Victoria Nuland.
…
</TXT>
</DOC>

FIGURE C.4 – Alignement des deux médoides : ”Iran”

C.5 Cinquième paire de clusters : ”Armes chimiques en Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,531008571627248

<DOC>
<DOCID>554b79fb9f7cde11a1c90552c62aefca</DOCID>
<PUBDATE>Tue Mar 19 17:33:00 CET 2013</PUBDATE>
<CURDATE>Tue Mar 19 22:07:30 CET 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>
<ITEMURI>http://www.bbc.co.uk/news/world-middle-east-21841217#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Syrians trade chemical attack claims</TITLE>
<DESC>Syrian rebels and the government accuse each other of firing chemical weapons, reportedly killing 25 people, but there is no confirmation of either claim.</DESC>
<TXT>Insurgency mapped Syrian rebels and the government have accused each other of firing chemical weapons, reportedly killing at least 25 people in the north of the country. A Syrian minister said it was a "dangerous escalation" and the "first act" of a new rebel authority. However, both a chemical weapons monitoring body and the US said there was no evidence they had been used. Both sides say the attack happened in the Khan al-Assal region north of the second city, Aleppo. The US says it is looking carefully at the allegations, while Russia has backed the Syrian government's claims. Analysis Jonathan Marcus BBC Diplomatic Correspondent At this stage it is impossible to verify if a chemical weapon has been used in northern Syria or indeed who may have fired one. The government and the rebels both accuse each other. Given Syria's extensive chemical weapons arsenal there have long been fears that either the government forces might use such weapons or that they might fall into the hands of Syrian opposition fighters. The exact status of Syria's arsenal which includes blister agents like mustard and persistent nerve agents like Sarin is unclear. There have been periodic "scares" when Western intelligence agencies claimed to see activity at weapons depots, but so far there has been no hard evidence that chemical warheads have been delivered to the units that might fire them. US President Barack Obama has made it clear that the use of such weapons would represent a "red line", which if crossed would lead to serious consequences. If confirmed, it would be the first time chemical weapons have been used in the two-year Syrian conflict. "Terrorists launched a missile containing chemical products into the region of Khan al-Assal in the province of Aleppo, killing 15 people, mainly civilians," Sana news agency said. The government routinely refers to rebels as "terrorists". State TV later said 25 people had died, while the pro-opposition Syrian Observatory for Human Rights put the figure at 26, including 16 soldiers. Senior rebel and spokesman for the Higher Military Council in Aleppo Qassim Saadeddine said the government had carried out a chemical attack. "We were hearing reports from early this morning about a regime attack on Khan al-Assal, and we believe they fired a Scud with chemical agents," he told Reuters news agency. "Then suddenly we learned that the regime was turning these reports against us. The rebels were not behind this attack." The Aleppo Media Centre, which is affiliated to the rebels, said there had been cases of "suffocation and poison" among civilians in Khan al-Assal after a surface-to-surface missile was fired at the area.
…
</TXT>
</DOC>

<DOC>
<DOCID>04817110f793022e84e24573860bcfe1</DOCID>
<PUBDATE>Tue Mar 19 18:49:00 CET 2013</PUBDATE>
<CURDATE>Tue Mar 19 21:53:28 CET 2013</CURDATE>
<FEEDURI>http://www.lavenir.net/rss.aspx?foto=1&intro=1&section=info&info=1642237c-66b9-4e8a-a8c1-288d61fefe7e</FEEDURI>
<ITEMURI>http://www.lavenir.net/article/detail.aspx?articleid=DMF20130319_00284325</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Régime et rebelles syriens s’accusent d’user d’armes chimiques pour la première fois</TITLE>
<DESC>Le régime et les rebelles se sont mardi accusés mutuellement d’avoir utilisé des armes chimiques, pour la première fois en deux ans de conflit en Syrie, mais les États-Unis ont dit ne disposer d’aucune preuve sur un tel recours.</DESC>
<TXT>Guerre en Syrie
Le régime et les rebelles se sont mardi accusés mutuellement d’avoir utilisé des armes chimiques, pour la première fois en deux ans de conflit en Syrie, mais les États-Unis ont dit ne disposer d’aucune preuve sur un tel recours.
La Russie a repris à son compte les accusations du régime de Bachar al-Assad, son allié, en disant avoir «reçu des informations» selon lesquelles des rebelles ont utilisé des armes chimiques lors d’une attaque dans la province d’Alep (nord) qui a fait selon un dernier bilan officiel 31 morts.
Alors qu’il n’était pas possible de confirmer de source indépendante un tel recours, l’Observatoire syrien des droits de l’Homme (OSDH) a confirmé un tir de missile sol-sol contre l’armée dans la localité de Khan al-Assal mais a dit douter qu’il soit chargé de matières non conventionnelles.
Entre-temps à Istanbul, le «Premier ministre» intérimaire Ghassan Hitto, élu lundi par l’opposition, a dit qu’il ne dialoguerait pas avec le pouvoir et cité parmi ses priorités la chute du régime et l’envoi de l’aide aux populations des régions passées sous contrôle rebelle.
M. Hitto, de la mouvance islamiste, doit s’atteler à la formation de l’équipe attendue d’ici un mois et qui aura la charge de protéger les infrastructures et les ressources publiques et privées, gérer les postes frontières aux mains de la rébellion et coordonner l’aide humanitaire internationale.
Paris et Washington, qui soutiennent les rebelles, ont salué cette élection.
Une escalade dangereuse
«C’est une escalade dangereuse. Des terroristes ont tiré un missile contenant des produits chimiques à partir de Kfar Daël dans la région de Naïrab (est d’Alep) vers la région de Khan al-Assal (ouest de la métropole)», a déclaré le ministre syrien de l’Information Omrane al-Zohbi.
Il s’en est pris à «la Ligue arabe, la communauté internationale et les États qui arment, financent et hébergent les terroristes ainsi que le gouvernement (turc) d’Erdogan et le Qatar» après «ce crime perpétré par les terroristes qui ont utilisé une arme prohibée par la loi internationale».
Mais les rebelles de l’Armée syrienne libre (ASL) ont démenti cette allégation et accusé en retour le régime.
…
</TXT>
</DOC>

FIGURE C.5 – Alignement des deux médoides : ”Armes chimiques en Syrie”

C.6 Sixième paire de clusters : ”Président chinois”

Valeur de comparabilité entre les deux clusters alignés : 0,4963489422143512

<DOC>
<DOCID>57e8f5ad8e1a38dac78dbee83e2b2cb8</DOCID>
<PUBDATE>Fri Mar 22 07:42:23 CET 2013</PUBDATE>
<CURDATE>Fri Mar 22 09:07:32 CET 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>
<ITEMURI>http://www.bbc.co.uk/news/world-asia-china-21873944#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Xi Jinping embarks on Russia visit</TITLE>
<DESC>Xi Jinping heads to Russia on the first stop of his maiden overseas tour as president of China, with a visit to Africa to follow.</DESC>
<TXT>Xi Jinping heads to Russia in first foreign tour Chinese trade with Russia and Africa is on the rise China pledges credit for Africa Xi Jinping is heading to Russia on the first stop of his maiden overseas tour as president of China. Mr Xi is set to meet Russian President Vladimir Putin, with the two likely to discuss energy and investment deals. Speaking ahead of the visit, Mr Xi said the two countries were "most important strategic partners" who spoke a "common language". He will also visit Tanzania, South Africa and the Republic of Congo on his tour, which continues until 30 March. In South Africa, he will attend the fifth Brics summit from 26-27 March. Brics stands for Brazil, Russia, India, China and South Africa - five key emerging economies. The choice of Moscow as Mr Xi's first destination is seen as symbolic, and a move from China to counter the US pivot to Asia, correspondents say. Russia is one of the world's biggest energy producers, and China is the world's top energy consumer. Bilateral trade is booming, reaching a record $88bn (58bn) last year. Beijing and Moscow have held similar positions over a number of thorny diplomatic issues, from Iran to Syria to North Korea, and some analysts suggest the bond is likely to strengthen. 'Just world order' At a press conference, Mr Xi called Russia China's "friendly neighbour", and said that the fact that he was visiting so soon after assuming presidency was "a testimony to the great importance China places on its relations with Russia." "China-Russia relations have entered a new phase in which the two countries provide major development opportunities to each other," he said. In an interview with Russian press, Mr Putin said that Russia-China co-operation would produce "a more just world order". Russia and China both demonstrated a "balanced and pragmatic approach" to international crises, he said. In an article in 2012, the Russian president had called for further economic co-operation with China to "catch the 'Chinese wind' in [its] economic sails". China is also Africa's largest trading partner, surpassing the United States and its traditional European partners. "China-Africa co-operation is comprehensive," Mr Xi said. "It has contributed to Africa's international standing." Xi Jinping was confirmed as China's president last week, concluding a lengthy transition process that saw him assume the Communist Party leadership in November 2012. More on This Story</TXT>
</DOC>

<DOC>
<DOCID>431c3a7dc3fc555a062cfaa0cca8f457</DOCID>
<PUBDATE>Sat Mar 23 15:10:06 CET 2013</PUBDATE>
<CURDATE>Sun Mar 24 07:49:25 CET 2013</CURDATE>
<FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI>
<ITEMURI>http://www.romandie.com/news/n.asp?n=_Une_visite_a_Moscou_au_dela_des_attentes_pour_le_nouveau_president_chinois82230320131510.asp</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Une visite à Moscou au-delà des attentes, pour le nouveau président chinois</TITLE>
<DESC>MOSCOU - Le président chinois Xi Jinping a estimé samedi que sa visite en Russie, qui s'est soldée par une trentaine d'accords, a dépassé toutes les attentes, et a loué le rôle joué, selon lui, par Moscou et Pékin pour assurer...</DESC>
<TXT>Tweet
Une visite à Moscou au-delà des attentes, pour le nouveau président chinois
MOSCOU - Le président chinois Xi Jinping a estimé samedi que sa visite en Russie, qui s'est soldée par une trentaine d'accords, a dépassé toutes les attentes, et a loué le rôle joué, selon lui, par Moscou et Pékin pour assurer l'équilibre international.
Ma visite a atteint son but. Les résultats ont largement dépassé toutes mes attentes , a déclaré le président chinois, lors d'une rencontre avec le Premier ministre russe Dmitri Medvedev.
Je suis très content, a souligné M. Xi, cité dans un communiqué du service de presse du gouvernement russe, alors qu'une trentaine d'accords notamment dans le pétrole et le gaz ont été conclus la veille à l'issue de sa rencontre avec le président russe Vladimir Poutine.
Le président chinois, qui est arrivé à Moscou vendredi avec son épouse Peng Liyuan pour une visite de trois jours, a précisé avoir choisi la Russie pour son premier déplacement à l'étranger afin de montrer l'importance particulière des relations russo-chinoises.
Le temps passe, beaucoup de choses dans le monde changent, mais les relations russo-chinoises ne font que se renforcer et se développer, a renchéri M. Xi qui qualifie MM Medvedev et Poutine de vieux amis.
Pour sa part, le Premier ministre russe s'est félicité de nouveaux résultats importants obtenus dans le cadre de cette visite survenue quelques jours après l'investiture de M. Xi comme président de la République populaire.
Tout cela apporte une contribution considérable dans le développement de nos relations, et nous en sommes ravis, a déclaré M. Medvedev.
Le président russe Vladimir Poutine et son homologue chinois ont déjà affiché vendredi des relations au beau fixe, le chef d'État russe ayant salué cette visite historique avec des résultats positifs.
Vendredi, la Russie et la Chine se sont mises d'accord notamment sur une augmentation des ventes de pétrole par le groupe pétrolier russe Rosneft au chinois CNPC, ainsi que sur une coopération sur huit blocs d'exploration dans l'Arctique russe.
De son côté, le géant gazier russe Gazprom a signé un accord avec CNPC en vue de livrer à la Chine à compter de 2018 un volume de 38 milliards de mètres cubes de gaz par an.
...
</TXT>
</DOC>

FIGURE C.6 – Alignement des deux médoides : ”Président chinois”

C.7 Septième paire de clusters : ”Israel et Turquie”

Valeur de comparabilité entre les deux clusters alignés : 0,4905583428832619

<DOC>
<DOCID>4f64ad06603f76414fbd0323e69e1dc9</DOCID>
<PUBDATE>Tue Mar 26 22:41:36 CET 2013</PUBDATE>
<CURDATE>Wed Mar 27 07:48:37 CET 2013</CURDATE>
<FEEDURI>http://www.france24.com/en/monde/rss</FEEDURI>
<ITEMURI>http://www.france24.com/en/20130326-israel-pay-turkey-millions-gaza-flotilla-deaths-haaretz?ns_campaign=editorial&ns_source=RSS_public&ns_mchannel=RSS&ns_fee=0&ns_linkname=20130326_israel_pay_turkey_millions_gaza_flotilla</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>HAARETZ: Israel to pay Turkey tens of millions over flotilla deaths</TITLE>
<DESC>Israel will transfer tens of millions of dollars to a humanitarian fund set up by the Turkish government to compensate for the deaths of nine Turkish activists aboard a flotilla bound for Gaza in 2010.</DESC>
<TXT>Israel to pay Turkey tens of millions over flotilla deaths Haaretz Israel will transfer tens of millions of dollars to a humanitarian fund set up by the Turkish government to compensate for the deaths of nine Turkish activists aboard a flotilla bound for Gaza in 2010. By HAARETZ (text) In the wake of Prime Minister Benjamin Netanyahus apology Friday to Turkish Prime Minister Recep Tayyip Erdogan over the deaths of nine Turkish activists aboard the 2010 Gaza flotilla, the two countries have set the wheels in motion to pay compensation over the deaths, with Israel set to pay out as much as tens of millions of dollars, according to sources in Turkey. High-level diplomatic contact between the two countries began on Monday when Turkish Foreign Minister Ahmet Davutoglu spoke with Justice Minister Tzipi Livni over the establishment of a joint committee that will formulate the terms of Israels agreement to pay compensation. The vice prime minister of Turkey, Bulent Arinc, told journalists on Monday that both sides agreed to establish a joint high-level committee over the coming days to discuss the details of the compensation transfer. Beyond the technical and legal questions over the compensation payments, the waiver of the legal claims and the extent of the blockade on Gaza, the Palestinian issue rather than the Syrian one will continue to be the focus of future relations between the two countries. The Turkish foreign minister made it clear during Tuesdays Arab League summit in Doha that Turkey will continue to stand with the Palestinian people and will act in order to end Israeli occupation. In Turkey, they estimate that the three-year long rift caused by the Palestinian question now gives Turkey leverage, and that the nature of the relationship between it and Israel will be largely dependent upon Israels behavior towards the Palestinians. As for Syria, Israel and Turkey see its future differently. While Israel is concerned by the possibility that Assads rule may fall and be replaced by an extremist Islamic regime, or that the state may be dismantled by armed forces that will control different sections, Turkey estimates that the Syrian opposition, which will also include Islamic movements, will be able to lead Syria and will not be a threat to the region.
…
</TXT>
</DOC>

<DOC>
<DOCID>065ca2c7abb83130f90c294aa62cf5c3</DOCID>
<PUBDATE>Sat Mar 23 08:44:40 CET 2013</PUBDATE>
<CURDATE>Sat Mar 23 10:12:36 CET 2013</CURDATE>
<FEEDURI>http://rss.lemonde.fr/c/205/f/3052/index.rss</FEEDURI>
<ITEMURI>http://www.lemonde.fr/proche-orient/article/2013/03/23/obama-met-un-terme-a-la-brouille-entre-la-turquie-et-israel_1853126_3218.html</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Obama met un terme à la brouille entre la Turquie et Israël</TITLE>
<DESC>Barack Obama a obtenu de Benyamin Nétanyahou qu'il présente ses excuses à son homologue turc pour l'abordage de la "Flottille de la liberté" en mai 2010.</DESC>
<TXT>Obama met un terme à la brouille entre la Turquie et Israël Le Monde |
• Mis à jour le
23.03.2013 à 09h46
C'est le succès diplomatique de dernière minute arraché par le président américain Barack Obama lors de sa visite en Israël , qui s'est achevée vendredi. Le premier ministre israélien Benyamin Nétanyahou a présenté ses excuses au chef du gouvernement turc Recep Tayyip Erdogan pour la mort de neuf Turcs à bord d'une flottille pour Gaza en 2010.
Dans un communiqué diffusé quelques minutes avant la fin de sa première visite officielle en Israël, Barack Obama indique que les deux hommes se sont entretenus par téléphone. "Les Etats-Unis sont très attachés à leur partenariat étroit avec la Turquie comme avec Israël et nous accordons une grande importance à la restauration de relations positives entre eux afin de consolider la paix et la sécurité dans la région", est-il écrit. L'appel d'une trentaine de minutes a été passé dans le véhicule qui conduisait Benyamin Nétanyahou et le président américain jusqu'à l'avion du second, sur l'aéroport de Tel Aviv, d'où il devait partir pour la Jordanie , a-t-on appris de sources américaines.
"Tous deux sont convenus de normaliser les relations entre les deux pays, y compris le retour des ambassadeurs", selon un communiqué officiel israélien. "Le premier ministre Nétanyahou a présenté ses excuses au peuple turc pour toute erreur ayant pu conduire à la perte de vies et accepté l'indemnisation" des victimes, assurant que "les résultats tragiques de la flottille du "Mavi Marmara" n'étaient pas intentionnels", selon le texte. M. Erdogan a accepté ces excuses "au nom du peuple turc" et les deux dirigeants "sont convenus de la conclusion d'un accord pour une indemnisation" des familles des victimes, selon un communiqué de ses services .
Déjà tendues depuis l'opération israélienne meurtrière "Plomb durci" dans la bande de Gaza (décembre 2008-janvier 2009), les relations entre la Turquie et Israël, alliés stratégiques dans les années 1990, se sont brutalement dégradées le 31 mai 2010 lors de l'assaut israélien contre une flottille tentant de briser le blocus israélien du territoire palestinien gouverné par le Hamas. Neuf passagers du navire turc Mavi Marmara avaient été tués, provoquant une crise diplomatique entre les deux pays. Ankara a abaissé le niveau de sa représentation diplomatique en Israël, dont il a expulsé l'ambassadeur, et suspendu la coopération militaire.
…
</TXT>
</DOC>

FIGURE C.7 – Alignement des deux médoides : ”Israel et Turquie”

C.8 Huitième paire de clusters : ”Israel et Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,48343711155149277

<DOC>
<DOCID>e421b44e511f6434fc81e5c468ffac97</DOCID>
<PUBDATE>Sun Mar 24 15:41:44 CET 2013</PUBDATE>
<CURDATE>Mon Mar 25 07:08:45 CET 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>
<ITEMURI>http://www.bbc.co.uk/news/world-middle-east-21917351#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Israel 'returns fire' from Syria</TITLE>
<DESC>The Israeli government says its military has destroyed a machine-gun nest inside Syria after troops were shot at twice in the Golan Heights.</DESC>
<TXT>Israel opens fire after 'shooting from inside Syria' Israel seized the Golan Heights during the 1967 war Insurgency mapped The Israeli government has said its military destroyed a machine-gun nest inside Syria after troops were shot at twice in the Golan Heights. An Israeli military spokesman said it was his understanding that the shots had not been stray fire from fighting in the civil war between the Syrian government and rebels. Troops responded by firing a guided missile at the Syrian position. Israel has occupied the Golan Heights since the 1967 war. Defence Minister Moshe Yaalon said his country would not let "Syria's army or any other element" violate its sovereignty. No Israeli soldiers were hurt in the shooting, during which army vehicles were hit, the military said. More on This Story</TXT>
</DOC>

<DOC>
<DOCID>99e75e309efe73a2254500cc38142ba4</DOCID>
<PUBDATE>Tue Apr 02 22:46:05 CEST 2013</PUBDATE>
<CURDATE>Wed Apr 03 08:54:20 CEST 2013</CURDATE>
<FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI>
<ITEMURI>http://www.romandie.com/news/n.asp?n=_Echange_de_tirs_israelo_syriens_sur_les_hauteurs_du_Golan20020420132246.asp</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Echange de tirs israélo-syriens sur les hauteurs du Golan</TITLE>
<DESC>JERUSALEM - Un char israélien a ouvert le feu mardi soir en direction du territoire syrien après qu'un obus de mortier syrien, accompagné de tirs d'armes légères, fut tombé dans la partie du plateau du Golan occupée par Israël, a...</DESC>
<TXT>Tweet
Echange de tirs israélo-syriens sur les hauteurs du Golan
JERUSALEM - Un char israélien a ouvert le feu mardi soir en direction du territoire syrien après qu'un obus de mortier syrien, accompagné de tirs d'armes légères, fut tombé dans la partie du plateau du Golan occupée par Israël, a indiqué l'armée israélienne.
Lors d'un deuxième incident ce soir dans le plateau du Golan, une patrouille de Tsahal (l'armée israélienne, ndlr) a été la cible de tirs à l'arme légère à la frontière israélo-syrienne. Il n'y a eu ni blessé ni dégât, a indiqué un communiqué militaire.
En réponse, l'armée a ouvert le feu sur la source des tirs et l'a touchée avec précision, a précisé le communiqué.
Un obus de mortier venant de Syrie était auparavant tombé dans le sud du plateau du Golan, un secteur occupé par Israël, sans faire de victime ni de dégât, selon une porte-parole militaire.
L'armée israélienne, qui a qualifié ces incidents de graves, a transmis une plainte officielle à la FNUOD (Force de l'observation du désengagement sur le Golan), chargée de faire respecter le cessez-le-feu entre les deux voisins.
Depuis le début du conflit en Syrie il y a deux ans, la situation s'est tendue sur le Golan, dans le sud syrien, mais les incidents --obus syriens tombant côté israélien et tirs de semonce israéliens-- sont restés jusqu'à présent relativement limités.
Les responsables israéliens attribuent jusqu'à présent la chute récurrente d'obus syriens en territoire sous contrôle israélien à des erreurs de tirs, en raison de la proximité des combats entre les forces du régime et les rebelles.
Que ce soit notre territoire qui soit visé ou non, nous allons riposter pour faire taire la source des tirs, comme nous l'avons déjà fait la semaine dernière, a réaffirmé mardi à des journalistes le ministre israélien de la Défense, Moshé Yaalon, lors d'une visite sur le Golan.
Le 24 mars, M. Yaalon avait déjà promis de répondre immédiatement à tout tir syrien, ajoutant dans un communiqué qu'il tenait le régime de Damas responsable de toute violation de la souveraineté israélienne.
Le jour même, des soldats israéliens postés dans la partie du Golan occupée par Israël avaient ouvert le feu sur une position militaire syrienne après avoir essuyé des tirs du territoire syrien pour la deuxième fois en 24 heures, selon une porte-parole militaire israélienne.
Israël est officiellement en état de guerre avec la Syrie. Il occupe depuis 1967 quelque 1.200 km2 du plateau du Golan, qu'il a annexés, une décision que n'a jamais reconnue la communauté internationale, environ 510 km2 restant sous contrôle syrien.
</TXT>
</DOC>

FIGURE C.8 – Alignement des deux médoides : ”Israel et Syrie”

C.9 Neuvième paire de clusters : ”Afghanistan”

Valeur de comparabilité entre les deux clusters alignés : 0,4774912614103054

<DOC>
<DOCID>7891005eaee0e07a18427b1fb08b0737</DOCID>
<PUBDATE>Sun Apr 07 22:30:26 CEST 2013</PUBDATE>
<CURDATE>Mon Apr 08 08:24:55 CEST 2013</CURDATE>
<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>
<ITEMURI>http://www.bbc.co.uk/news/world-asia-22058455#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Afghan children 'killed by Nato'</TITLE>
<DESC>Eleven children are reported to have been among those killed in a Nato air strike in Kunar province in eastern Afghanistan.</DESC>
<TXT>Q&A: Foreign forces Eleven children have been killed in a Nato air strike in eastern Afghanistan, officials and witnesses say. At least one woman was reportedly killed and a further six are believed to have been injured in the incident in Shigal district, Kunar province. Nato confirmed that "fire support" was used in Shigal after a US civilian adviser died in a militant attack, but said it had no reports of deaths. Afghan President Hamid Karzai condemned the killings. A statement issued by his office said he had already issued a decree banning aerial attacks on civilian areas. Villagers and officials told the BBC that the casualties were inside their homes when they died. Photographs apparently sent from the scene to international news agencies appeared to show the bodies of several dead young children, surrounded by Afghan villagers. A local official said eight Taliban insurgents had also died in the air strike on Saturday, which is reported to have caused the roofs of several houses in three villages to collapse. Analysis Bilal Sarwary BBC News, Kabul The narrow and mountainous valley of Shultan lies 30km (20 miles) away from the provincial capital of Asadabad, right on the border with Pakistan's Bajaur tribal agency. The area is covered with dense forest, offering the perfect cover for insurgents. Afghan intelligence officials in Kunar say Afghan and foreign fighters have a big presence in the area and often launch attacks against Afghan and international forces in Afghanistan, and against Pakistani military positions across the border. The area has been the site of intense fighting between Taliban and US/Afghan forces for the last 10 years, with US-led forces often carrying out operations in the valley targeting Afghan Taliban and foreign fighters with al-Qaeda backing. The Afghan government has struggled to exert its control in this strategic district despite several major American offensives, as the militants keep re-grouping. Afghan counter-terrorism officials in the province say foreign fighters have been training local fighters in the area for quite some time, and their presence has become a major threat for the security of Kunar province. He said the strikes were called in to support a major operation by US and Afghan government forces targeting senior Taliban commanders and a local weapons cache. Tribal elder Haji Malika Jan told the BBC: "The fighting started yesterday morning [Saturday] and continued for at least seven hours. There were heavy exchanges between both sides. "The area is very close to the Pakistani border and there are hundreds of local and foreign fighters, mostly Pakistanis, in the area."
…
</TXT>
</DOC>

<DOC>
<DOCID>fcfd05837ed025622575d05dce3b32db</DOCID>
<PUBDATE>Sun Apr 07 21:12:26 CEST 2013</PUBDATE>
<CURDATE>Mon Apr 08 08:16:24 CEST 2013</CURDATE>
<FEEDURI>http://liberation.fr.feedsportal.com/c/32268/fe.ed/rss.liberation.fr/rss/10/</FEEDURI>
<ITEMURI>http://liberation.fr.feedsportal.com/c/32268/f/606159/s/2a6dcb7c/l/0L0Sliberation0Bfr0Cmonde0C20A130C0A40C0A70Cafghanistan0Eun0Ebombardement0Ede0El0Eotan0Etue0Edix0Eenfants0I894270A0Dxtor0Frss0E450A/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Afghanistan: Karzaï condamne un bombardement de l’Otan ayant tué 11 enfants</TITLE>
<DESC>Articles en rapport Violences au Caire entre Coptes et musulmans La Corée du Nord prépare un double essai missile/bombe nucléaire, indique Séoul Afghanistan: Karzaï condamne un bombardement de l’Otan ayant tué 11 enfants Monténégro: le président sortant et son adversaire revendiquent la victoire Afghanistan: 9 passagers d’un bus tués par une bombe artisanale</DESC>
<TXT>Par AFP
Libération
Le président Hamid Karzaï a «fermement condamné» dimanche un bombardement de l’Otan samedi dans l’est de l’Afghanistan qui a tué onze enfants afghans, malgré de multiples injonctions de sa part à cesser les attaques aériennes sur des zones d’habitations.
«Tout en condamnant l’utilisation de civils comme boucliers, le président a dénoncé toute opération qui cause la mort de civils», peut-on lire dans un communiqué de la présidence afghane, ajoutant qu’une «délégation» se rendrait sur les lieux pour enquêter.
L’Isaf, la force de l’Otan en Afghanistan, qui indiquait jusqu’alors que «jusqu’à dix femmes et enfants avaient été blessés mais non pas tués», selon l’un de ses porte-parole dimanche après-midi, a déclaré quelques heures plus tard qu’elle «prenait acte des informations sur la mort de dix enfants», selon un autre de ses communicants.
«Nous rassemblons les faits pour comprendre ce qui s’est produit. Nous prenons chaque perte civile très au sérieux», a poursuivi cet autre porte-parole.
Un premier bilan, confirmé par trois responsables de la province du Kunar, l’un des bastions talibans de l’Est du pays où l’incident s’est produit, faisait état de 10 enfants morts, auxquels s’ajoutait la mort d’une femme, selon l’une de ces sources.
Le bombardement s’est produit alors qu’un combat intense opposait des troupes afghanes et américaines à des insurgés talibans dans le district de Shigal, selon plusieurs sources afghanes et l’Isaf.
«Avant le bombardement, un Américain a été tué et quatre membres des forces de sécurité afghanes ont été blessés dans une attaque des insurgés», a commenté Wasifullah Wasifi, le porte-parole du gouvernement provincial du Kunar.
La mort d’un civil américain dans l’Est afghan a été annoncée samedi par les forces armées américaines par communiqué, sans plus de précisions. Le porte-parole de l’Isaf a confirmé à l’AFP qu’il s’agissait bien du même incident.
«On nous tirait dessus depuis plusieurs maisons de la zone. Un Américain a été tué et plusieurs de nos hommes blessés. La force de la coalition a répondu par un bombardement», a expliqué une source sécuritaire afghane présente pendant l’opération.
…
</TXT>
</DOC>

FIGURE C.9 – Alignement des deux médoides : ”Afghanistan”

C.10 Dixième paire de clusters : ”Paris, Londre et Syrie”

Valeur de comparabilité entre les deux clusters alignés : 0,471830265362955

<DOC>
<DOCID>f5ca7d94c5dc35ad171583a4d615c451</DOCID>
<PUBDATE>Thu Mar 14 08:05:36 CET 2013</PUBDATE>
<CURDATE>Thu Mar 14 08:48:21 CET 2013</CURDATE>
<FEEDURI>http://rss.nytimes.com/services/xml/rss/nyt/World.xml</FEEDURI>
<ITEMURI>http://rendezvous.blogs.nytimes.com/2013/03/13/britain-and-france-push-for-arming-syrian-opposition/?partner=rss&emc=rss</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>IHT Rendezvous: Britain and France Push for Arming Syrian Opposition</TITLE>
<DESC>After two years of an increasingly vicious conflict in Syria, European governments are being urged to lift an arms embargo that some view as helping to keep the Damascus regime in power.</DESC>
<TXT>March 13, 2013, 9:17 am Britain and France Push for Arming Syrian Opposition By HARVEY MORRIS Jim Lopez/Agence France-Presse Getty Images An armed Syrian rebel dodged sniper fire in Aleppo on Monday. LONDON Britain and France are putting pressure on their European partners to lift an embargo on weapons and ammunition to Syrian rebels in a conflict that has claimed 70,000 lives and created one million refugees . David Cameron, the British prime minister, has said his government might veto an extension of the European embargo when it comes up for renewal in May. I hope that we can persuade our European partners, he told a parliamentary committee on Tuesday. But if we cant, then its not out of the question we might have to do things in our own way. His remarks came as Laurent Fabius, the French foreign minister, called for the European Union to rethink a weapons ban that he said favored the Damascus regime, which continued to receive powerful weaponry from Russia and Iran. We understand the idea of not adding weapons to weapons, Mr. Fabius told a parliamentary committee. But that position doesnt work in the face of reality, and that is that the opposition is bombarded by others who are getting weapons while they are not. The stance of the Continents two biggest military powers reflected frustration at the failure to find a diplomatic solution to the two-year civil conflict. However, it has prompted concerns from some European partners about the wisdom of sending weapons to a volatile region that could end up in the hands of anti-Western jihadists. Guido Westerwelle, the German foreign minister, said during a visit to London last week that the decision so far not to lift the embargo was wise and right. We have to avoid a conflagration in the whole region, he said. Mr. Camerons signal that Britain was prepared to go it alone with arms supplies to the rebels came on the eve of an Anglo-Russian strategic dialogue in London on Wednesday that was to include discussion of the Syrian crisis. The British and other Western governments have criticized Moscow for supporting the regime of President Bashar al-Assad and failing to push him toward a negotiated settlement. Ahead of the London talks, however, the Russians have hinted they might be prepared to halt their own weapons supplies to the regime if there were a similar ban on sending arms to the rebels.
…
</TXT>
</DOC>

<DOC>
<DOCID>3c2645d68e38f5ffa175f129c42b6621</DOCID>
<PUBDATE>Fri Mar 15 01:02:00 CET 2013</PUBDATE>
<CURDATE>Sat Mar 16 10:20:18 CET 2013</CURDATE>
<FEEDURI>http://feeds.lefigaro.fr/c/32266/f/438192/index.rss</FEEDURI>
<ITEMURI>http://feeds.lefigaro.fr/c/32266/f/438192/s/2990bb7e/l/0L0Slefigaro0Bfr0Cinternational0C20A130C0A30C140C0A10A0A30E20A130A314ARTFIG0A0A50A60Eparis0Eet0Elondres0Eveulent0Elivrer0Edes0Earmes0Eaux0Erebelles0Esyriens0Bphp/story01.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Paris et Londres veulent livrer des armes aux rebelles syriens</TITLE>
<DESC>VIDÉO - Le président français et son ministre des Affaires étrangères ont marqué la volonté de la France d'enfreindre, s'il le faut, l'embargo européen sur la livraison d'armes à destination des rebelles. Articles en rapport Le plaidoyer des Frères musulmans contre l'égalité des femmes d'Egypte Des milices islamistes apparaissent en Égypte L'ambitieuse Theresa May dérange Cameron Le régime syrien bloque toujours l'aide humanitaire Syrie : armer les rebelles risque d'attiser les divisions</DESC>
<TXT>Réactions (670)
François Hollande a exprimé la volonté de la France de livrer des armes aux rebelles syriens. Crédits photo : THIERRY CHARLIER/AFP
Tweet
Recommander
VIDÉO - Le président français et son ministre des Affaires étrangères ont marqué la volonté de la France d'enfreindre, s'il le faut, l'embargo européen sur la livraison d'armes à destination des rebelles.
Correspondant à Bruxelles
Imposant la question syrienne au sommet, François Hollande a demandé jeudi aux Européens de lever l'embargo sur les armes de guerre en faveur de la révolte contre Bachar el-Assad, sans exclure de passer outre si ses partenaires s'y refusaient. «Nous souhaitons que les Européens lèvent l'embargo (…), c'est ce que je vais dire à mes collègues, a dit le président en arrivant à Bruxelles. La France doit d'abord convaincre ses partenaires. Si d'aventure il devait y avoir un blocage, alors la France prendrait ses responsabilités (…) Mais elle doit aussi prendre ses responsabilités. On ne peut pas laisser un peuple se faire massacrer par un régime qui a démontré qu'il refuse toute discussion politique.»
Paris peut compter sur le soutien résolu de Londres. Dès mardi, le premier ministre David Cameron avait annoncé que faute d'accord européen, il entendait agir «comme bon (lui) semble» pour livrer des armes à la rébellion syrienne. Dans un contraste saisissant avec l'acrimonie franco-britannique du dernier sommet, les deux hommes se sont retrouvés pour un tête-à-tête avant le huis clos à vingt-sept.
Les rebelles réclament des armes antichars
L'objectif affiché est de faire monter la pression politique sur le régime el-Assad pour qu'il accepte enfin le dialogue avec une opposition que la France a été la première à reconnaître. Il ne s'agit «pas d'aller vers une guerre totale» et «la France n'écarte pas une issue politique» à l'avenir, insiste le chef de l'Etat. Mais à l'instant présent, «nous devons considérer que les solutions politiques ont échoué».
…
</TXT>
</DOC>

FIGURE C.10 – Alignement des deux médoides : ”Paris, Londre et Syrie”

D

Dix premières paires de clusters obtenues sur la base du **Tri du pire des cas**

Nous présentons ici dix premières paires de clusters alignés, représentées par leurs médoides, obtenues sur la base du **Tri du pire des cas** des plus faibles comparabilités pour le corpus Flux RSS.

D.1 Premières paire de clusters : ”Animal”

Valeur de comparabilité entre les deux clusters alignés : 0,2514	
<DOC> <DOCID>6b4e8896232d8455cd2630d1765f99e8</DOCID> <PUBDATE>Wed Feb 20 15:59:58 CET 2013</PUBDATE> <CURDATE>Wed Feb 20 20:41:27 CET 2013</CURDATE> <FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI> <ITEMURI>http://www.bbc.co.uk/news/world-asia-21519560#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Lost Antarctic penguin found in NZ</TITLE> <DESC>A royal penguin is being cared for at a Wellington Zoo after washing up on the coast of New Zealand, 2,000km (1,240 miles) from its Antarctic home.</DESC> <TXT>Lost Antarctic Royal penguin found in New Zealand Vets said the penguin could make its way home if it recovered Lost penguin Happy Feet 'missing' A royal penguin is being cared for at a New Zealand zoo after being found stranded on a beach 2,000km (1,240 miles) from home its Antarctic. The young male bird, which was dehydrated and starving, is thought to be only the fourth royal penguin to wash up there in more than a century. He is believed to have come from a breeding colony in the sub-Antarctic Macquarie Island. Vets said the bird, dubbed Happy Feet Jr, may have been drifting for a year. Lisa Argilla, a vet at Wellington Zoo, said the penguin had possibly struggled to find enough food or had had problems hunting and had come ashore as he needed to go through his seasonal moulting. He was found on Tora beach, on the coast to the south of Wellington, on Sunday. "It's very weak, doesn't want to stand. It's making very small progress every day but it's still in critical condition," Ms Argilla told the TVNZ channel. She told AFP his kidneys were not functioning properly, adding: "Hopefully we can reverse that, feed him up and bring him back to good health but it's touch and go at the moment." If he recovered, she said, he would be released to make his way home. "They're amazing at navigation so that shouldn't be a problem for him," she said. Last year, an emperor penguin, the original Happy Feet, made headlines when he appeared on New Zealand's shores. He had surgery to remove 3kg (6.6lb) of sand from his stomach, which he is thought to have eaten thinking it was snow, before being released with a tracking device. But he disappeared soon after and was believed to have been eaten. More on This Story</TXT> </DOC>	<DOC> <DOCID>e35912719c9d6731b04887102f207e12</DOCID> <PUBDATE>Fri Feb 22 01:56:03 CET 2013</PUBDATE> <CURDATE>Fri Feb 22 07:14:05 CET 2013</CURDATE> <FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI> <ITEMURI>http://www.romandie.com/news/n.asp?n=Triste_fin_pour_le_manchot_Happy_Feet_Junior40220220130156.asp</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Triste fin pour le manchot Happy Feet Junior</TITLE> <DESC>Un manchot surnommé Happy Feet junior qui avait été rejeté sur une côte néo-zélandaise à 2000 km de sa région d'origine est mort malgré tous les efforts mis en oeuvre pour le sauver, a annoncé vendredi le Zoo de Wellington. Selon sa...</DESC> <TXT>Tweet Triste fin pour le manchot Happy Feet Junior Un manchot surnommé Happy Feet junior qui avait été rejeté sur une côte néo-zélandaise à 2000 km de sa région d'origine est mort malgré tous les efforts mis en oeuvre pour le sauver, a annoncé vendredi le Zoo de Wellington. Selon sa vétérinaire en chef Lisa Argilla, il a succombé à la malnutrition et à un problème rénal. Une équipe de vétérinaires a passé cinq jours à soigner l'oiseau, un jeune manchot royal, qui a dérivé très loin de sa colonie d'origine, l'île subantarctique de Macquarie, après avoir passé environ 12 mois en mer. "A son arrivée, le manchot pesait près de trois kilos de moins que la normale, il n'avait absolument aucune réserve et de ce fait, nous supposons que cela a provoqué chez lui de multiples défaillances viscérales, après l'insuffisance rénale diagnostiquée à son arrivée", a-t-elle dit. Spécialité difficile "La médecine vétérinaire pour les animaux sauvages est une spécialité très difficile, et bien que nous ayions fait le mieux que nous pouvions, malheureusement le manchot n'a pas survécu", a-t-elle ajouté. La découverte de l'oiseau a rappelé l'histoire de Happy Feet, un manchot empereur qui s'était échoué sur une plage près de Wellington en juin 2011 et qui était devenu une attraction mondiale pendant les huit semaines où il avait été soigné au zoo. Le manchot avait finalement été relâché par un navire de recherche néo-zélandais dans l'océan Antarctique, après avoir reçu les visites de personnalités telles que l'acteur et réalisateur britannique Stephen Fry et les meilleurs voeux du Premier ministre néo-zélandais John Key. Cependant, le boîtier de localisation GPS qui avait été accroché à l'oiseau avait cessé de transmettre après quelques jours, faisant craindre que le manchot ait été dévoré par un requin. (ats / 22.02.2013 01h56)</TXT> </DOC>

FIGURE D.1 – Alignement des deux médoides : ”Animal”

D.2 Deuxième paire de clusters : ”Argents”

Valeur de comparabilité entre les deux clusters alignés : 0,2379	
<DOC> <DOCID>b182778847fa2111dcaf5f4ac44a1ac9</DOCID> <PUBDATE>Thu Feb 14 08:57:14 CET 2013</PUBDATE> <CURDATE>Fri Feb 15 07:41:35 CET 2013</CURDATE> <FEEDURI>http://www.globaltimes.cn/DesktopModules/DnnForge%20-%20NewsArticles/Rss.aspx?TabID=99&ModuleID=405&CategoryID=14,49,50,51,52,53,15&MaxCount=100&sortBy=StartDate&sortDirection=DESC</FEEDURI> <ITEMURI>http://www.globaltimes.cn/content/761532.shtml</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Australians lose millions of dollars in online dating scam</TITLE> <DESC>Organized criminals have defrauded West Australian victims out of at least 4.5 million AU dollars ($4. 66 million) over the last six months after starting relationship with vulnerable people online, the Government of Western Australia (WA) revealed on Thursday.</DESC> <TXT>Australians lose millions of dollars in online dating scam Xinhua | 2013-2-14 15:57:14 Print Organized criminals have defrauded West Australian victims out of at least 4.5 million AU dollars ($4. 66 million) over the last six months after starting relationship with vulnerable people online, the Government of Western Australia (WA) revealed on Thursday. Relationship fraud losses totaling about 568,000 AU dollars were reported to WA ScamNet line between August 2012 and January 2013, according to the state's Department of Commerce. Dom Blackshaw from the WA Major Fraud Squad said the police had identified millions of dollars transferred to fraudsters in Nigeria, Ghana and Sierra Leone. "In the frauds identified, some individuals have sent up to 300, 000 AU dollars," Blackshaw said in a statement. "After identifying hundreds of vulnerable individuals repeatedly sending large amounts of money, we began the painstaking task of sending letters to try to alert them to the fact they are being duped." WA's Commissioner for Consumer Protection Anne Driscoll said relationship fraud victims found it difficult to accept that the person they were dating online for months or years was a con artist.</TXT> </DOC>	<DOC> <DOCID>462c0e891997c4aa9a13d69e902b8ddc</DOCID> <PUBDATE>Mon Apr 15 08:44:08 CEST 2013</PUBDATE> <CURDATE>Mon Apr 15 10:03:06 CEST 2013</CURDATE> <FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI> <ITEMURI>http://www.romandie.com/news/n.asp?n=Crocodile_Dundee_a_perdu_des_millions_de_dollars_placen_Suisse_RP_150420130844-16-346456.asp</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>"Crocodile Dundee" a perdu des millions de dollars placés en Suisse</TITLE> <DESC>L'acteur australien Paul Hogan, qui incarnait "Crocodile Dundee" dans le film du même nom, a déposé plainte devant un tribunal américain. Il veut récupérer des millions de dollars déposés sur le compte d'une banque suisse, qui lui ont été...</DESC> <TXT>Tweet "Crocodile Dundee" a perdu des millions de dollars placés en Suisse L'acteur australien Paul Hogan, qui incarnait "Crocodile Dundee" dans le film du même nom, a déposé plainte devant un tribunal américain. Il veut récupérer des millions de dollars déposés sur le compte d'une banque suisse, qui lui ont été dérobés par son conseiller fiscal, selon lui. L'acteur de 73 ans, devenu star grâce au film de 1986, affirme que son conseiller fiscal a fait main basse sur 34 millions de dollars US (plus de 30 millions de francs), précise lundi la presse australienne. Dans des documents déposés auprès d'un tribunal de Californie, le plaignant déclare que Philip Egglishaw "a pris la fuite ou a dépensé" les millions déposés à la Cornèr Banque à Lausanne, selon le "Sydney Morning Herald" et "The Australian". En 2012, Paul Hogan avait trouvé un accord à l'amiable confidentiel avec le fisc australien, aux termes de huit ans de négociation. Les autorités lui réclamaient, ainsi qu'à son collaborateur John Cornell, plus de 150 millions de dollars australiens (148 millions de francs) en impôts non payés depuis les années 1980, intérêts et pénalités de retard. (ats / 15.04.2013 07h58) </TXT> </DOC>

FIGURE D.2 – Alignement des deux médoides : ”Argents”

D.3 Troisième paire de clusters : ”Tennis”

Valeur de comparabilité entre les deux clusters alignés : 0,2282	
<DOC> <DOCID>5b184e8d15849a9c599615323c7a4130</DOCID> <PUBDATE>Sat Apr 06 22:44:49 CEST 2013</PUBDATE> <CURDATE>Sun Apr 07 09:48:08 CEST 2013</CURDATE> <FEEDURI>http://www.france24.com/en/monde/rss</FEEDURI> <ITEMURI>http://www.france24.com/en/20130406-argentina-france-davis-cup?ns_campaign=editorial&ns_source=RSS_public&ns_mchannel=RSS&ns_fee=0&ns_linkname=20130406_argentina_france_davis_cup</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>TENNIS: Argentina pull ahead of France in Davis Cup quarter-final</TITLE> <DESC>Argentina on Saturday took a 2-1 lead in their Davis Cup quarter-final against France in Buenos Aires after coming back from a set down to win a doubles match.</DESC> <TXT>Argentina pull ahead of France in Davis Cup quarter-final AFP Argentina on Saturday took a 2-1 lead in their Davis Cup quarter-final against France in Buenos Aires after coming back from a set down to win a doubles match. By News Wires (text) Argentina came from a set down to upset favourites France in the doubles and take a 2-1 lead in their Davis Cup quarter-final at Parque Roca on Saturday. David Nalbandian and Horacio Zeballos beat Julien Benneteau and Michael Llodra 3-6 7-6 7-5 6-3 after a remarkable comeback from 4-1 down in the third set. Zeballos, who had a poor opening set, lifted his game to complement experienced Davis Cup campaigner Nalbandian, a former world number one, and stun the French pair who won only three of the last 14 games in windy conditions. World number eight Jo-Wilfried Tsonga, who won the opening point for France when he beat Carlos Berlocq in five sets on Friday, is scheduled to meet Juan Monaco in Sunday's first singles and will be trying to save the tie for France. Gilles Simon, who lost Friday's second singles to Monaco, is due to face Berlocq in the fifth rubber. But Simon, chosen in preference to Richard Gasquet who has ankle trouble, complained of back pain during the second set on Friday and could be a doubt for Sunday. (REUTERS)</TXT> </DOC>	<DOC> <DOCID>f77b0fa20fdd9760f579475572967d00</DOCID> <PUBDATE>Sat Apr 06 21:11:05 CEST 2013</PUBDATE> <CURDATE>Sat Apr 06 23:01:40 CEST 2013</CURDATE> <FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI> <ITEMURI>http://www.romandie.com/news/n.asp?n=Le_piege_se_referme_pour_les_Mousquetaires51060420132111.asp</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Le piège se referme pour les Mousquetaires</TITLE> <DESC>Le dimanche s'annonce infernal pour l'équipe de France de Coupe Davis ! Les "Mousquetaires" sont en effet menés 2-1 à Buenos Aires devant l'Argentine après leur défaite en double. Julien Benneteau et...</DESC> <TXT>Tweet Le piège se referme pour les Mousquetaires Le dimanche s'annonce infernal pour l'équipe de France de Coupe Davis ! Les "Mousquetaires" sont en effet menés 2-1 à Buenos Aires devant l'Argentine après leur défaite en double. Julien Benneteau et Michael Llodra se sont inclinés en quatre sets, 3-6 7-6 7-5 6-3, devant David Nalbandian et Horacio Zeballos. Les Français ont cultivé deux travers au cours de cette rencontre: une infériorité très marquée en retour de service, notamment de la part de Llodra, et un manque de maîtrise nerveuse. Les Français ont mené 5-2 dans le troisième set avant de perdre le fil de leur tennis au moment même où il devait porter l'estocade. Remarquable joueur de double, David Nalbandian fut bien sûr l'homme fort de ce double. Avec sa main exceptionnelle, le joueur de Cordoba a réussi des prouesses et placé l'Argentine sur une voie royale. Richard Gasquet forfait et Gilles Simon touché au dos, les Français ont besoin d'un véritable miracle dimanche au Parque Roca pour espérer gagner ce quart de finale. A Astana aussi, rien n'est dit. Mené 2-0 après les simples de vendredi, le Kazahkstan a obtenu un sursis avec le succès en trois sets en double du duo Schukin/Golubev devant Stepanek/Hajek. L'absence de Tomas Berdych peut inciter les Kazakhs à croire à un impossible retour dimanche. (Sport Information / 06.04.2013 21h11) </TXT> </DOC>

FIGURE D.3 – Alignement des deux médoides : ”Tennis”

D.4 Quatrième paire de clusters : ”Films”

Valeur de comparabilité entre les deux clusters alignés : 0,2242

<DOC>

<DOCID>24095c554a8654d98983d41bd30af660</DOCID>

<PUBDATE>Sat Feb 16 22:24:01 CET 2013</PUBDATE>

<CURDATE>Sun Feb 17 12:02:40 CET 2013</CURDATE>

<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>

<ITEMURI>http://www.bbc.co.uk/news/world-europe-21485735#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>Romanian film wins top Berlin prize</TITLE>

<DESC>The Romanian film Child's Pose picks up the Golden Bear for best film at the Berlin film festival, while an unemployed Bosnian Roma wins best actor.</DESC>

<TXT>Romanian film Child's Pose wins Golden Bear in Berlin Director Calin Peter Netzer said he was a "little bit speechless" by his film's win The Romanian film Child's Pose has picked up the coveted Golden Bear prize for best film at the 63rd Berlin film festival. The film, directed by Calin Peter Netzer, tells the story of a wealthy mother who uses her connections to try and stop her son from going to jail. The film was a favourite among the 19 contenders. Mr Netzer said he was "a little bit speechless" by the win. An unemployed Roma from Bosnia-Hercegovina won the best actor award. Nazif Mujic re-enacted his family's real-life struggle to get vital medical treatment in the low-budget An Episode In the Life of an Iron Picker, which also picked up the runner-up Silver Bear award. Calin Peter Netzer's film is a tale of corruption and guilt in modern Romania. It follows a rich and controlling mother, played by Luminita Gheorghiu, as she bribes witnesses into giving false statements to save her son from jail after he accidentally runs down and kills a boy. US filmmaker David Gordon Green won best director at the festival for his comic road movie Prince Avalanche, while best actress went to Chile's Paulina Garcia for her role as a Santiago divorcee in Gloria. Share this page</TXT>

</DOC>

<DOC>

<DOCID>c1451396450ad38ba47e5c35d456d06d</DOCID>

<PUBDATE>Sat Feb 16 20:43:05 CET 2013</PUBDATE>

<CURDATE>Sun Feb 17 11:52:34 CET 2013</CURDATE>

<FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI>

<ITEMURI>http://www.romandie.com/news/n.asp?n=_Berlinale__l_Ours_d_or_attribue_au_drame_roumain_Child_s_Pose40160220132043.asp</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>Berlinale : l'Ours d'or attribué au drame roumain Child's Pose</TITLE>

<DESC>BERLIN - Le drame roumain Child's Pose de Calin Peter Netzer, racontant l'histoire d'une mère de famille aisée qui cherche à protéger son fils responsable d'un accident de la route mortel, a remporté samedi l'Ours d'or récompensant le...</DESC>

<TXT>Tweet Berlinale : l'Ours d'or attribué au drame roumain Child's Pose

BERLIN - Le drame roumain Child's Pose de Calin Peter Netzer, racontant l'histoire d'une mère de famille aisée qui cherche à protéger son fils responsable d'un accident de la route mortel, a remporté samedi l'Ours d'or récompensant le meilleur film à la Berlinale.

Le jury présidé par le réalisateur chinois Wong Kar Wai a également distingué David Gordon Green, meilleur réalisateur pour sa comédie Prince Avalanche. L'acteur bosniaque Nazif Mujic, qui joue dans le drame de Danis Tanovic An Episode in the Life of an Iron Picker sur les Roms, et l'actrice chilienne Paulina Garcia, qui campe une quasi-sexagénaire pleine de vie dans Gloria, ont été désignés meilleurs interprètes.

Child's Pose est un film minimaliste, tourné en majeure partie dans des appartements privés. L'actrice Luminita Gheorghiu incarne une mère ultra-possessive, Cornelia, qui se sert de ses relations et de son argent pour éviter la prison à son fils immature, Barbu, coupable d'avoir écrasé un adolescent d'une famille modeste à la suite d'un excès de vitesse.

Outre les lieux de tournage très confinés, la caméra ne quitte presque jamais les deux principaux protagonistes, renforçant le côté étouffant de la relation, inspirée par celle que le réalisateur entretenait avec sa propre mère, avait avoué ce dernier devant la presse.

Je suis un peu sans voix. Je veux remercier le jury pour ce prix formidable, a déclaré Calin Peter Netzer, très ému.

Le film bosniaque An Episode in the Life of an Iron Picker, raconte, lui, l'histoire vraie et désespérée d'un couple de Roms, à travers lequel Danis Tanovic a voulu attirer l'attention sur le sort qu'il juge scandaleux de cette minorité.

Il s'agit d'un docu-fiction avec des non professionnels qui interprètent leurs propres rôles. Outre l'acteur Nazif Mujic, ce long-métrage a également reçu l'Ours d'argent Grand Prix du Jury.

Je ne m'attendais vraiment pas à ça (...) Parfois la colère peut déboucher sur de bonnes choses. Pas souvent, mais c'est le cas cette fois, a commenté Danis Tanovic.

La récompense pour la meilleure actrice est allée à la favorite chilienne Paulina Garcia. Surtout connue pour ses rôles à la télévision, elle interprète le personnage Gloria, éponyme de ce film de Sebastian Lelio, une presque sexagénaire déterminée à être heureuse en dépit des coups durs de la vie.

L'Ours d'argent du meilleur réalisateur a été attribué au jeune David Gordon Green (37 ans) pour son film Prince Avalanche, qui dépeint Alvin (Paul Rudd) et Lance (Emile Hirsch), dont le métier monotone consiste à tracer des lignes jaunes sur des routes apparemment sans fin, dans les paysages sauvages du Texas. Avec ce trophée, le jury a récompensé la seule vraie comédie en course parmi les 19 films de la sélection officielle.

(©AFP / 16 février 2013 20h41)

</TXT>

</DOC>

FIGURE D.4 – Alignement des deux médoides : ”Films”

D.5 Cinquième paire de clusters : ”Président français”

Valeur de comparabilité entre les deux clusters alignés : 0,2220	
<DOC> <DOCID>db218632feb94ca8261dd665d0deb20d</DOCID> <PUBDATE>Fri Mar 15 20:20:16 CET 2013</PUBDATE> <CURDATE>Sat Mar 16 10:30:02 CET 2013</CURDATE> <FEEDURI>http://www.france24.com/en/monde/rss</FEEDURI> <ITEMURI>http://www.france24.com/en/20130315-carla-bruni-calls-french-president-hollande-penguin-song-sarkozy?ns_campaign=editorial&ns_source=RSS_public&ns_mchannel=RSS&ns_fee=0&ns_linkname=20130315_carla_bruni_calls_french_president_hollande</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>FRANCE: Is Hollande the ‘penguin’ in Carla Bruni's new song?</TITLE> <DESC>The lyrics of a new song by France’s former first lady Carla Bruni-Sarkozy, about a “penguin” who “takes on the airs of a king,” were widely interpreted as a swipe at French President François Hollande when they were made public on Friday.</DESC> <TXT>Is Hollande the penguin in Carla Bruni's new song? AFP file photo The lyrics of a new song by Frances former first lady Carla Bruni-Sarkozy, about a penguin who takes on the airs of a king, were widely interpreted as a swipe at French President Franois Hollande when they were made public on Friday. By News Wires (text) Carla Bruni has reportedly taken a swipe at French President Francois Hollande, depicting him as a bumbling buffoon with no manners in a song that features on the former first lady's new album. The lyrics of "The Penguin" were released Friday and immediately interpreted as an attack on the man who succeeded her husband, Nicolas Sarkozy, as France's leader. The former supermodel sings: "He takes on the airs of a king/but I know, the penguin/doesn't have the manners of a lord. "Hey penguin!/If one day you cross my path again/I will teach you, penguin/I will teach you to kiss my hand." French media saw the lyrics as a reference to Hollande's frosty treatment of his outgoing predecessor on the day when he took over as president, notably declining to accompany Sarkozy to the car that carried him away from the presidential Elysee palace. In French, describing someone as a penguin implies they are both clumsy and a little ridiculous in the manner of a clown or a buffoon. "The Penguin" is one of the tracks on "Little French Songs", Bruni's fourth album, which is due to be released on April 1. It also includes a song, "Chez Keith et Anita" (At Keith and Anita's place), in which Bruni, a former girlfriend of Mick Jagger, depicts the drug-fuelled lifestyle of Rolling Stones guitarist Keith Richards and his longtime girlfriend Anita Pallenberg. The new album is Bruni's first since 2008's "Comme Si De Rien N'Etait" (Simply). Her musical career was put on hold while Sarkozy was in office. (AFP)</TXT> </DOC>	<DOC> <DOCID>45cd8b3ed67231784d61040fac28df3c</DOCID> <PUBDATE>Fri Mar 15 18:03:00 CET 2013</PUBDATE> <CURDATE>Sat Mar 16 10:13:19 CET 2013</CURDATE> <FEEDURI>http://www.lavenir.net/rss.aspx?foto=1&intro=1&section=info&info=1642237c-66b9-4e8a-a8c1-288d61fefe7e</FEEDURI> <ITEMURI>http://www.lavenir.net/article/detail.aspx?articleid=DMF20130315_00282758</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Carla Bruni chante «le pingouin», en référence à François Hollande?</TITLE> <DESC>Carla Bruni sort les griffes. La chanteuse et ancienne Première dame s'en prend à François Hollande, qu'elle qualifie de "pingouin" solitaire, sans manières ni personnalité, dans une chanson de son prochain album.</DESC> <TXT>Politique française Carla Bruni sort les griffes. La chanteuse et ancienne Première dame s'en prend à François Hollande, qu'elle qualifie de "pingouin" solitaire, sans manières ni personnalité, dans une chanson de son prochain album. Le nouvel album de Carla Bruni-Sarkozy, "Little French Songs", publié par Barclay, filiale d'Universal est attendu le 1er avril. La radio RTL a publié ce vendredi des extraits d'une chanson. "Il prend son petit air souverain mais je le connais, moi, le pingouin n'a pas de manière de châtelain. Eh, le pingouin, si un jour tu recroises mon chemin, je t'apprendrai, le pingouin, je t'apprendrai à me faire le baise main (...) Ni laid ni beau, le pingouin, ni haut ni bas, ni froid ni chaud, le pingouin, ni oui ni non (...) Tiens le pingouin t'as l'air tout seul dans ton jardin", selon ces paroles qui semblent viser François Hollande. Selon les médias français, cette chanson rappelle la passation de pouvoir, le 15 mai 2012, à l'Elysée entre Nicolas Sarkozy et François Hollande. Le chef de file des députés de droite Christian Jacob, proche de Nicolas Sarkozy, a jugé vendredi que la chanson était un "raccourci un peu rapide", tout en concédant que "les rimes sont jolies et la chanson est bien écrite". "Little french songs", du nom de l'un des titres où l'ancienne mannequin rend hommage à tous les grands auteurs compositeurs français, sera le quatrième album de la chanteuse et ex-mannequin, âgée de 45 ans, qui écrit elle-même les textes de ses chansons et en compose souvent la musique. Son troisième album, "Comme si de rien n'était", était sorti en juillet 2008, durant la présidence de Nicolas Sarkozy. Les gains avaient été reversés à la Fondation de France. 13 </TXT> </DOC>

FIGURE D.5 – Alignement des deux médoides : ”Président français”

D.6 Sixième paire de clusters : ”Vin”

Valeur de comparabilité entre les deux clusters alignés : 0,2168

<DOC>

<DOCID>dcd73bf7c0a9be7bc5bd73c42b78f950</DOCID>

<PUBDATE>Sat Mar 09 01:01:00 CET 2013</PUBDATE>

<CURDATE>Sat Mar 09 07:15:52 CET 2013</CURDATE>

<FEEDURI>http://www.thetimes.co.uk/tto/news/world/rss</FEEDURI>

<ITEMURI>http://www.thetimes.co.uk/tto/life/celebrity/article3709217.ece</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>Brangelina joins Beckham among the celebrity wine-makers</TITLE>

<DESC>Angelina Jolie, Brad Pitt and David Beckham have made France their second home, and although Gérard Depardieu has left the...</DESC>

<TXT>Welcome to your preview of The Times Subscribe now Brangelina joins Beckham among the celebrity wine-makers Brad Pitt and Angelina Jolie: Were choosing France for the long term. The life is good Ian West/PA Send 1 of 1 Brad Pitt and Angelina Jolie: Were choosing France for the long term. The life is good Ian West/PA David Chazan Paris Published at 12:01AM, March 9 2013 Angelina Jolie, Brad Pitt and David Beckham have made France their second home, and although Grard Depardieu has left the country, all four share a very Gallic hobby: winemaking. Next week Pitt and Jolie are to launch their first vintage, produced at their 1,200-acre Chteau Miraval estate in the south of France, which they bought four years ago. Their passion for wine is shared by David and Victoria Beckham, who bought a vineyard in Napa Valley, California, at about the same time. Beckhams new club, Paris Saint-Germain, has played down reports that it was planning to market a red wine Subscribe now To see the full article you need to subscribe Subscribe What Ive learnt: Jonny Lee Miller Published 1 minute ago Actor Jonny Lee Miller, 40, found fame playing Sick Boy in Trainspotting in 1996 the same year in which he married his first wife, Angelina Jolie Pendleton enters new cycle of life Published at 12:01AM, March 9 2013 Olympic gold medallist Victoria Pendleton talks to Robert Crampton about how she is coping with life after the Games Why they all love Westwoods Cocotte dress Published at 12:01AM, March 9 2013 The designers popular creation, which is from her premium Gold Label and sells for upwards of 1,500, is all blowsy, English sex appeal My Week Justin Bieber* Published at 12:01AM, March 9 2013 Justin baby, says my manager. Whats up with you these days? Do not mess with me, I say. Or I will unleash the Beliebers. Sponsored Editorial</TXT>

</DOC>

<DOC>

<DOCID>55dfc9eee51c584f0d542ab4b17b70cd</DOCID>

<PUBDATE>Fri Feb 15 19:15:09 CET 2013</PUBDATE>

<CURDATE>Sat Feb 16 08:59:49 CET 2013</CURDATE>

<FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI>

<ITEMURI>http://www.romandie.com/news/n.asp?n=Angelina_et_Brad_rejoignent_Depardieu_et_les_autres_dans_les_vignes_RP_150220131915-14-321586.asp</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>Angelina et Brad rejoignent Depardieu et les autres dans les vignes</TITLE>

<DESC>"Mis en bouteille par Jolie-Pitt & Perrin": Angelina Jolie et Brad Pitt vont rejoindre la longue liste des célébrités internationales qui investissent dans le vin. Ils lanceront en mars leur vin rosé, produit dans le sud de la France. Le...</DESC>

<TXT>Tweet Angelina et Brad rejoignent Depardieu et les autres dans les vignes

"Mis en bouteille par Jolie-Pitt & Perrin": Angelina Jolie et Brad Pitt vont rejoindre la longue liste des célébrités internationales qui investissent dans le vin. Ils lanceront en mars leur vin rosé, produit dans le sud de la France.

Le couple d'acteurs s'est associé à une famille française de vignerons, les Perrin, pour élaborer l'étiquette du "Miraval Côtes de Provence", du nom de leur propriété acquise en 2008 pour quelque 40 millions d'euros. Près de la ville de Brignoles, la bâtisse du XVIIe siècle, qui leur sert de résidence d'été, est entourée de 500 hectares, dont 50 ou 60 hectares de vignes.

Le couple au top du glamour rejoint de nombreuses autres stars qui ont investi dans les vignes, pour une question d'image, par intérêt financier ou par passion. Le plus connu des acteurs-vignerons est Gérard Depardieu, un passionné qui a plusieurs vignobles dans le Bordelais, l'Anjou et le Languedoc.

Toujours dans le sud de la France, l'ex-champion allemand de Formule 1 Michael Schumacher possède un vignoble à Saint-Raphaël, classé Côte de Provence. Le footballeur David Beckham est lui aussi propriétaire d'un domaine dans le sud de la France, alors que le chanteur Sting produit un Chianti dans sa propriété en Toscane.

Image ou réelle implication?

"Les stars ne mettent que rarement la main à la pâte, par manque de compétences techniques et de temps, la quasi-totalité d'entre elles ne vivant pas sur le domaine. Si certaines ne font donc que vendre leur image, d'autres, en revanche, sont complètement impliquées dans leur business", indique le site mesvignes.com.

Qu'en sera-t-il pour les "Brangelina"? Ils veulent s'impliquer, selon leur associé Marc Perrin, propriétaire de plusieurs vignobles. "Ils veulent être fiers du vin produit sur leur propriété. Ils recherchent vraiment l'excellence", ajoute-t-il, après les avoir rencontrés.

Marc Perrin parle de 150'000 bouteilles de rosé mises sur le marché pour la première année. Il y aura également, plus tard, du rouge et du blanc. En France, le rosé coûtera "autour de 15 euros" et se vendra "chez les bons cavistes et dans les bons restos".

(ats / 15.02.2013 16h19)

</TXT>

</DOC>

FIGURE D.6 – Alignement des deux médoides : ”Vin”

D.7 Septième paire de clusters : ”Pilules”

Valeur de comparabilité entre les deux clusters alignés : 0,2143	
<DOC> <DOCID>76ad3f54b3ffcbba02b21aceaf34501e</DOCID> <PUBDATE>Fri Apr 05 17:51:36 CEST 2013</PUBDATE> <CURDATE>Sat Apr 06 23:14:59 CEST 2013</CURDATE> <FEEDURI>http://www.france24.com/en/monde/rss</FEEDURI> <ITEMURI>http://www.france24.com/en/20130405-us-judge-morning-after-pill-prescription?ns_campaign=editorial&ns_source=RSS_public&ns_mchannel=RSS&ns_fee=0&ns_linkname=20130405_us_judge_morning_after_pill_prescription</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>USA: US judge orders ‘morning-after’ pill available without prescription</TITLE> <DESC>The ‘morning-after’ emergency contraceptive pill must be made available to girls of all ages without prescription, a US federal judge ordered on Friday as part of a lawsuit brought by reproductive-rights groups.</DESC> <TXT>- healthcare reform - USA US judge orders morning-after pill available without prescription The morning-after emergency contraceptive pill must be made available to girls of all ages without prescription, a US federal judge ordered on Friday as part of a lawsuit brought by reproductive-rights groups. By News Wires (text) A federal judge on Friday ordered the U.S. Food and Drug Administration to make the morning-after emergency contraception pill available without a prescription to all girls of reproductive age. The ruling by U.S. District Judge Edward Korman in Brooklyn, New York, comes in a lawsuit brought by reproductive-rights groups that had sought to remove age and other restrictions on emergency contraception. Currently, only women age 17 or older can obtain emergency contraception pills without a prescription. Point-of-sale restrictions require that all women present identification to a pharmacist before obtaining the drug. In his ruling, Korman said the FDAs rejection of requests to remove age restrictions to obtain the pill had been arbitrary, capricious and unreasonable. Nancy Northup, president of the Center for Reproductive Rights, hailed the ruling. Women all over the country will no longer face arbitrary delays and barriers just to get emergency contraception, she said. FDA spokeswoman Erica Jefferson declined to comment on the ruling, saying it was an ongoing legal matter. (REUTERS)</TXT> </DOC>	<DOC> <DOCID>754766a106080059a34ea096930a67a8</DOCID> <PUBDATE>Fri Mar 01 15:50:08 CET 2013</PUBDATE> <CURDATE>Fri Mar 01 18:35:21 CET 2013</CURDATE> <FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI> <ITEMURI>http://www.romandie.com/news/n.asp?n=_L_Academie_de_medecine_pour_la_restriction_de_l_usage_des_pilules_combinees_apres_35_40_ans69010320131550.asp</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>L'Académie de médecine pour la restriction de l'usage des pilules combinées après 35-40 ans</TITLE> <DESC>PARIS - Pour limiter les risques associés aux pilules contraceptives combinées, l'Académie de médecine préconise notamment de réduire leur usage chez les plus de 35-40 ans et chez les femmes obèses dans un rapport rendu public...</DESC> <TXT>Tweet L'Académie de médecine pour la restriction de l'usage des pilules combinées après 35-40 ans PARIS - Pour limiter les risques associés aux pilules contraceptives combinées, l'Académie de médecine préconise notamment de réduire leur usage chez les plus de 35-40 ans et chez les femmes obèses dans un rapport rendu public vendredi. Dans ce rapport signé par cinq médecins, l'Académie relève que dans l'immédiat, les méthodes contraceptives existantes doivent toutes rester disponibles mais qu'elles doivent faire l'objet d'une stricte surveillance, avec un renforcement du dépistage des facteurs de risque et une information des femmes. La prise de position de l'Académie intervient alors que les pilules combinées de 3e et 4e génération sont dans le collimateur des autorités sanitaires depuis le début de l'année. Ces dernières demandent aux médecins de ne plus les prescrire en premier recours en raison d'un risque accru de thrombose (caillot) veineuse par rapport aux pilules de 2e génération. Dans leur rapport, les médecins parmi lesquels figure l'hématologue Jacqueline Conard, relèvent qu'après la grossesse, la contraception oestroprogestative est la cause la plus fréquente de thrombose veineuse chez les femmes en âge de procréer et que ces thromboses peuvent se compliquer d'embolie pulmonaire, un accident responsable de 10 décès par millions d'utilisatrices de contraception par an. Mais il existe également des facteurs aggravants pour la thrombose veineuse, comme un âge supérieur à 35 ans, une obésité ou des antécédents familiaux de thrombose veineuse qui conduisent les auteurs du rapport à proposer une restriction de l'usage des pilules combinées après 35-40 ans et à déconseiller sa prescription aux femmes obèses. Les pilules combinées sont par ailleurs contre-indiquées chez les femmes ayant déjà présenté des thromboses veineuses ou ayant une thrombophilie biologique connue ou encore une pathologie associée à un risque de thrombose. En l'absence d'antécédent personnel de thrombose ou de thrombophilie, l'Académie préconise dans ses règles de prescription, de donner la pilule de 2e génération en premier recours, comme le recommande le ministère de la santé et d'informer les femmes au mieux par un document écrit, précisant les risques. Les médecins sont également invités à respecter les contre-indications en ce qui concerne les risques artériels (infarctus et accident vasculaire cérébral) lorsqu'ils prescrivent une pilule combinée. Ces risques artériels, beaucoup plus rares, sont augmentés de la même manière quelle que soit la pilule oestroprogestative, et sont observés essentiellement chez les fumeuses chez qui la contraception oestroprogestative est contre-indiquée après 35 ans ou en cas d'hypertension non contrôlée ou de migraines avec aura, précise le rapport. Les risques s'observent aussi bien avec les pilules qu'avec les patchs ou anneaux vaginaux contenant des oestroprogestatifs. En cas de contre-indication, une contraception non hormonale comme un stérilet, un diaphragme ou un préservatif doit être proposée. L'autre solution est une pilule ne contenant qu'un progestatif, mais sa tolérance est souvent médiocre (acné, absence de règles ou saignements), note l'Académie de médecine. </TXT> </DOC>

FIGURE D.7 – Alignement des deux médoides : ”Pilules”

D.8 Huitième paire de clusters : ”Milliadaire américain”

Valeur de comparabilité entre les deux clusters alignés : 0,2011

<DOC>

<DOCID>bd20047f3363eed9a4c343e7330aefda</DOCID>

<PUBDATE>Fri Apr 12 06:53:18 CEST 2013</PUBDATE>

<CURDATE>Fri Apr 12 09:31:00 CEST 2013</CURDATE>

<FEEDURI>http://www.abc.net.au/news/feed/52278/rss.xml</FEEDURI>

<ITEMURI>http://www.abc.net.au/news/2013-04-12/us-billionaire-wins-fake-wine-trial/4625762</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>US billionaire wins fake wine trial</TITLE>

<DESC>A New York jury has backed a US billionaire who complained that a cache of wine he bought costing thousands of dollars a bottle was phony.</DESC>

<TXT>US billionaire wins fake wine trial Posted April 12, 2013 14:53:18 Map: United States A New York jury has backed a US billionaire who complained that a cache of wine he bought costing thousands of dollars a bottle was phony. Bill Koch, brother to the politically connected siblings who help fund the US conservative Tea Party movement, sued wealthy businessman Eric Greenberg over what he said was a bad batch of bottles in a $3.3 million purchase he made at auction. A federal court jury has agreed and awarded compensation to Mr Koch, according to his lawyer. John Hueston said his client would receive damages of about $336,800 and another $24,600 in additional damages for findings of wilful misconduct. The legal battle started in 2007 following an auction in which Mr Greenberg consigned part of his 70,000-bottle cellar for sale. Mr Koch bought 2,669 bottles. They had extraordinary labels like a Chateau Latour 1928, worth $2,700, a Chateau Latour 1864, several Cheval Blanc 1921s, and a Chateau Petrus magnum from 1921 that was bought for $27,900. Mr Koch testified he had been looking for "the best of the best" and thought he was getting it. Some bottles were said to have come from "English royalty". But 24 bottles turned out to be fake, Mr Koch said. 'Sour grapes' The scam of pouring cheap booze into prestigious bottles is an acknowledged problem in the wine world. Former major dealer Rudy Kurniawan is set to go on trial later this year in New York for allegedly running a fake wine factory in his home. In a sign of the stakes in these rarified circles, Mr Koch and Mr Greenberg threw considerable legal resources at the trial. Both men attended, each sitting flanked by half a dozen lawyers and assistants, while the jury handled evidence, including bottles of wine that the amateur might assume would cost a fortune - but could be worthless. Mr Greenberg's lawyer, Arthur Shartsis, said Mr Koch's case was just sour grapes and that his own client had never knowingly served a legally questionable wine. "Mr Greenberg didn't believe those bottles were fake," he said in his closing arguments to the jury. Mr Shartsis sought to shift the blame from his client to the auction house Zachys. "In the contract, Zachys had complete responsibility," he said. "Somebody made a mistake, that is not a fraud." But Mr Hueston argued that Mr Greenberg went ahead with his consignment to Zachys in full knowledge that the bottles included fakes - and in fact hoping that Mr Koch would snap them up. AFP</TXT>

</DOC>

<DOC>

<DOCID>22b816f87f3b68a1312975a38727c09e</DOCID>

<PUBDATE>Sat Apr 13 07:39:00 CEST 2013</PUBDATE>

<CURDATE>Sat Apr 13 15:37:38 CEST 2013</CURDATE>

<FEEDURI>http://www.lavenir.net/rss.aspx?foto=1&intro=1&section=info&info=1642237c-66b9-4e8a-a8c1-288d61fefe7e</FEEDURI>

<ITEMURI>http://www.lavenir.net/article/detail.aspx?articleid=DMF20130413_00295914</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>Grands crus contrefaits: un milliardaire américain obtient 12 millions de dollars de dommages</TITLE>

<DESC>La justice new-yorkaise a accordé vendredi 12 millions de dollars de dommages-intérêts à un milliardaire américain qui avait acheté à prix d’or de très grands vins qui se sont avérés contrefaits.</DESC>

<TXT>Source: afp

Le milliardaire floridien William Koch, à gauche, et son avocat John Hueston ont fait du procès une croisade pour assainir le marché du négoce du vin.

Associated Press/Reporters

ADVERTISING - ne pas effacer

La justice new-yorkaise a accordé vendredi 12 millions de dollars de dommages-intérêts à un milliardaire américain qui avait acheté à prix d’or de très grands vins qui se sont avérés contrefaits.

John Hueston, avocat de Bill Koch, cadre dans une entreprise spécialisée dans l’énergie en Floride, a confirmé le montant des dommages à l’AFP, en indiquant que son client était très content du résultat de son procès.

La décision intervient à l’issue d’un procès acharné durant lequel Bill Koch, qui appartient à la famille qui a aidé à collecter des fonds pour le mouvement conservateur Tea Party, a accusé l’homme d’affaires multicartes californien Éric Greenberg de fraude, affirmant qu’il lui avait vendu pour 3,5 millions de dollars de faux grands crus, lors d’une vente aux enchères.

Le plaignant, qui avait voulu faire de son procès une croisade pour assainir le marché du négoce du vin, était enchanté par le jugement.

«Il y avait un code du silence dans ce satané négoce du vin, et maintenant, il a été brisé», s’est-il réjoui selon ses propos rapportés dans la presse américaine.

La bataille juridique a démarré en 2007 après une vente aux enchères au cours de laquelle Éric Greenberg avait mis en vente 70.000 bouteilles de vin. Bill Koch avait acheté 2.669 de ces bouteilles. Mais 24 d’entre elles, acquises pour 355.000 dollars, s’étaient avérées contrefaites.

5

</TXT>

</DOC>

FIGURE D.8 – Alignement des deux médoides : ”Milliadaire américain”

D.9 Neuvième paire de clusters : ”Paris”

Valeur de comparabilité entre les deux clusters alignés : 0,2005	
<DOC> <DOCID>7ead0a78dfd196708371951d30853819</DOCID> <PUBDATE>Fri Apr 05 15:13:41 CEST 2013</PUBDATE> <CURDATE>Sat Apr 06 22:55:37 CEST 2013</CURDATE> <FEEDURI>http://v1.theglobeandmail.com/generated/rss/BN/International.xml</FEEDURI> <ITEMURI>http://feedproxy.google.com/~r/TheGlobeAndMail-International/~3/9SSZ6el0EDg/</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Paris using small flock of sheep to mow city lawns</TITLE> <DESC>The initiative, which started this week, sees four sheep from an island in Brittany put to work munching the bountiful grass of the Paris Archives</DESC> <TXT>Will the future see flocks of sheep baaing beneath the Eiffel Tower and bleating by Notre Dame cathedral? Paris is enlisting a few hungry sheep to keep the citys grass trim, replacing gas-guzzling lawnmowers. The initiative, which started this week, sees four sheep from an island in Brittany put to work munching the bountiful grass of the Paris Archives. The eco-experiment, which could expand around the capital from October, follows on from a stint last year by two goats that the Louvre hired to mow the lawn of Paris famed Tuileries gardens. Already, private companies have hundreds of operational sheep mowing lawns of big companies around Paris. (It) efficient... and cheap, says Paris City Halls Fabienne Giboudeaux. I can imagine this very easily in London. And New York... even Tokyo.</TXT> </DOC>	<DOC> <DOCID>9feb3b84a590191f835a75ad9811498b</DOCID> <PUBDATE>Tue Feb 12 19:48:06 CET 2013</PUBDATE> <CURDATE>Tue Feb 12 22:49:18 CET 2013</CURDATE> <FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI> <ITEMURI>http://www.romandie.com/news/n.asp?n=_Paris_lance_la_renovation_thermique_de_200_ecoles65120220131948.asp</ITEMURI> <AUTHOR>[]</AUTHOR> <TITLE>Paris lance la rénovation thermique de 200 écoles</TITLE> <DESC>PARIS - La Ville de Paris va lancer la rénovation thermique de 200 écoles, un plan d'une ampleur inégalé devant lui permettre d'atteindre l'objectif de 600 écoles rénovées d'ici 2020 inscrit dans son Plan Climat, s'est félicité mardi le...</DESC> <TXT>Tweet Paris lance la rénovation thermique de 200 écoles PARIS - La Ville de Paris va lancer la rénovation thermique de 200 écoles, un plan d'une ampleur inégalé devant lui permettre d'atteindre l'objectif de 600 écoles rénovées d'ici 2020 inscrit dans son Plan Climat, s'est félicité mardi le groupe EELV au Conseil de Paris. Le Conseil de Paris a adopté mardi une délibération autorisant la Ville à passer les appels d'offres pour réaliser ces travaux. Une première tranche de 100 écoles a déjà été mise en oeuvre en 2011. Paris est la première collectivité à lancer un projet de rénovation thermique de cette ampleur, a souligné René Dutrey (EELV), adjoint au Maire en charge du développement durable. Le groupe EELV a encouragé la Ville à aller plus loin. Cette initiative ambitieuse devra se poursuivre dès le début de la prochaine mandature par une nouvelle tranche de 300 écoles, a souligné Sylvain Garel, coprésident du groupe au Conseil de Paris. Par ailleurs, le volontarisme de notre Ville doit désormais s'étendre à d'autres équipements publics tels que les gymnases, les musées et les piscines, a-t-il estimé. (©AFP / 12 février 2013 19h45) </TXT> </DOC>

FIGURE D.9 – Alignement des deux médoides : ”Paris”

D.10 Dixième paire de clusters : ”Oscars”

Valeur de comparabilité entre les deux clusters alignés : 0,1966

<DOC>

<DOCID>8b9a2ef5b66f6743f6ccf66b1fa0ca8d</DOCID>

<PUBDATE>Sun Feb 24 08:10:15 CET 2013</PUBDATE>

<CURDATE>Sun Feb 24 13:25:33 CET 2013</CURDATE>

<FEEDURI>http://feeds.bbci.co.uk/news/world/rss.xml</FEEDURI>

<ITEMURI>http://www.bbc.co.uk/news/entertainment-arts-21563890#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>Twilight named top turkey at Razzies</TITLE>

<DESC>The final Twilight instalment sweeps the board at the Razzie awards in Hollywood, which highlight the year's worst films.</DESC>

<TXT>Twilight named top turkey at Razzie Awards Twilight and Snow White and the Huntsman's Kristen Stewart was worst actress Oscar sets the stage for big night The final Twilight film has swept the board at Hollywood's Razzie awards, which highlight the year's worst films. Twilight Saga: Breaking Dawn Part 2 was given seven Golden Raspberry titles including worst picture and worst actress for Kristen Stewart. Stewart, who is presenting an award at the Oscars on Sunday, was not present at the spoof award ceremony. Instead, a cardboard cut-out of the actress was brought in to the press conference to "accept" her trophy. The five Twilight films have made a total of $3bn (1.9bn) at the box office. Razzies founder John Wilson said the worst thing about the franchise was that "people take it so seriously". "I believe that rather than 40 million girls who bought tickets, it was four million girls who bought 10 tickets each," he added. "That makes me feel better about the American public." The film also "won" the worst director prize for Bill Condon, worst supporting actor for Taylor Lautner, worst re-make, rip-off or sequel, worst screen ensemble and worst screen couple for Lautner and Mackenzie Foy. Stewart owes her worst actress prize to her performances in two films - Twilight and Snow White and the Huntsman. Adam Sandler (left) has won several Razzies and "triumphed" again The Razzies are described by the organisers as "saluting the worst that Hollywood has to offer each year". The prize for each category is a gold-coloured raspberry trophy. Winners rarely turn up to claim their prizes at the Los Angeles ceremony, and Stewart was no exception. 'Female Rambo' Pop star Rihanna was named worst supporting actress for her debut film role in Battleship, based on the board game of the same name. The singer has just 68 lines of dialogue in the action movie, including lines such as "Kentucky fried chicken!", "Boom", "Yeah" and "Sucker's really jumping around". John Wilson described Rihanna as "Rambert - a female Rambo". "She's following in the footsteps of Mariah Carey, Jennifer Lopez and Madonna - singers who really can't act but get paid millions of dollars anyway." Razzies veteran Adam Sandler was named worst actor for his role in That's My Boy. The film also won worst script. Last year, the event was moved from its traditional Oscars-eve slot to April Fools' Day. But voters of the Golden Raspberry Award Foundation said they preferred handing their prizes out the night before the Oscars, so the event was shifted back to that slot for 2013. More on This Story</TXT>

</DOC>

<DOC>

<DOCID>fe9a5d0133455e5860078de2e37b5134</DOCID>

<PUBDATE>Sun Feb 24 05:00:02 CET 2013</PUBDATE>

<CURDATE>Sun Feb 24 13:13:16 CET 2013</CURDATE>

<FEEDURI>http://www.romandie.com/rss/flux.xml</FEEDURI>

<ITEMURI>http://www.romandie.com/news/n.asp?n=Happiness_Therapy_et_Amour_vainqueurs_des_Spirit_Awards68240220130500.asp</ITEMURI>

<AUTHOR>[]</AUTHOR>

<TITLE>"Happiness Therapy" et "Amour" vainqueurs des Spirit Awards</TITLE>

<DESC>"Happiness Therapy" a raflé samedi les principales récompenses des Spirit Awards, les "Oscars" du cinéma indépendant, notamment pour son actrice Jennifer Lawrence. "Amour", de Michael Haneke, a remporté pour la France le trophée du meilleur...</DESC>

<TXT>Tweet

"Happiness Therapy" et "Amour" vainqueurs des Spirit Awards

"Happiness Therapy" a raflé samedi les principales récompenses des Spirit Awards, les "Oscars" du cinéma indépendant, notamment pour son actrice Jennifer Lawrence. "Amour", de Michael Haneke, a remporté pour la France le trophée du meilleur film étranger.

"Happiness Therapy", une comédie romantique sur la rencontre entre un homme bi-polaire essayant de reconstruire sa vie après une douloureuse séparation (Bradley Cooper) et une jeune veuve instable (Jennifer Lawrence) est reparti avec les trophées de meilleur film, réalisateur et scénario pour David O'Russell, et de meilleur actrice.

Jennifer Lawrence, 22 ans, ajoute ainsi un trophée à sa collection, après son Golden Globe et son prix du Syndicat américain des acteurs (SAG). Elle se pose désormais en favorite pour l'Oscar de la meilleure actrice, qui sera remis dimanche.

John Hawkes meilleur acteur

Seule catégorie ayant échappé au film, le trophée du meilleur acteur, qui n'est pas allé à Bradley Cooper mais à John Hawkes, pour son rôle de paralytique découvrant la sexualité dans "The Sessions".

Ce film, signé Ben Lewin, a également valu à Helen Hunt le trophée du meilleur second rôle féminin - une catégorie dans laquelle elle concourt également aux Oscars dimanche. Le prix du second rôle masculin est allée à Matthew McConaughey pour son rôle de patron de club de strip-tease masculin dans "Magic Mike" de Steven Soderbergh.

Encore "Amour"

Enfin, le trophée du meilleur film étranger a été remporté par "Amour" de Michael Haneke, déjà couronné la veille à la cérémonie des Césars à Paris. Le réalisateur autrichien s'est réjoui de recevoir un prix "remis par un public jeune, quelque chose d'important pour un film qui traite d'un sujet aussi sérieux".

Le film était en lice pour la France, contrairement aux Oscars où il concourt sous les couleurs autrichiennes pour le film étranger, et dans quatre autres catégories, dont meilleur film et meilleure actrice pour Emmanuelle Riva.

(ats / 24.02.2013 05h00)

</TXT></DOC>

FIGURE D.10 – Alignement des deux médoides : ”Oscars”

Résumé

Les corpus comparables thématiques regroupent des textes issus d¡¯un même thème et rédigés dans plusieurs langues, fortement similaires mais ne comprenant pas de traductions mutuelles. Par rapport aux corpus parallèles qui regroupent des paires de traductions, les corpus comparables présentent trois avantages: premièrement, ce sont des ressources riches et larges : en volume et en période couverte; deuxièmement, les corpus comparables fournissent des ressources linguistiques originales et thématiques. Enfin, ils sont moins coûteux à développer que les corpus parallèles.

Avec le développement considérable du WEB, une matière première très abondante est exploitable pour la construction de corpus comparables. En contre-partie, la qualité des corpus comparables est essentielle pour leur utilisation dans différents domaines tels que la traduction automatique ou assistée, l¡¯extraction de terminologies bilingues, la recherche d¡¯information multilingue, etc.

L¡¯objectif de ce travail de thèse est de développer une approche méthodologique et un outillage informatique pour fournir une assistance à la construction des corpus comparables bilingues et thématiques de ? bonne qualité ?, à partir du WEB et à la demande.

Nous présentons tout d¡¯abord la notion de mesure de comparabilité qui associe deux espaces linguistiques et, à partir d¡¯une mesure quantitative de comparabilité de référence, nous proposons deux variantes, qualifiées de comparabilité thématique, que nous évaluons suivant un protocole basé sur la dégradation progressive d¡¯un corpus parallèle. Nous proposons ensuite une nouvelle méthode pour améliorer le co-clustering et la co-classification de documents bilingues, ainsi que l¡¯alignement des clusters comparables. Celle-ci fusionne des similarités natives définies dans chacun des espaces linguistiques avec des similarités induites par la mesure de comparabilité utilisée. Enfin, nous proposons une démarche intégrée basée sur les contributions précédemment évoquées afin d¡¯assister la construction, à partir du WEB, de corpus comparables bilingues thématiques de qualité. Cette démarche comprend une étape de validation manuelle pour garantir la qualité de l¡¯alignement des clusters comparables. En jouant sur le seuil de comparabilité d¡¯alignement, différents corpus comparables associés à des niveaux de comparabilité variables peuvent être fournis en fonction des besoins spécifiés. Les expérimentations que nous avons menées sur des Flux RSS issus de grands quotidiens internationaux apparaissent pertinentes et prometteuses.

Mots-clés: Corpus comparables thématiques, Mesures de comparabilité, Co-clustering et co-classification de documents bilingues, Alignement des clusters, Constitution des corpus comparables

Abstract

Thematic comparable corpora regroup texts from a same topic and written in several languages, highly similar but without mutual translations. Comparing with parallel corpora which regroup pairs of translations, comparable corpora have three advantages: firstly, they are rich and big resources jointly in volume and in covered period; secondly, comparable corpora provide original language and thematic resources. Finally, they are less expensive to develop than parallel corpus.

With the considerable development of the WEB, an abundant raw material is exploitable for the construction of comparable corpora. However, the quality of comparable corpus is essential for their use in various fields such as automatic or assisted translation, bilingual terminology extraction, multilingual information retrieval, etc.

The objective of this thesis work is to develop a methodological approach and a software toolkit to offer assistance in the construction of thematic bilingual comparable corpora from the WEB and on demand.

We first introduce the general concept of comparability that maps two linguistic spaces and then, from a referenced quantitative comparability measure, we propose two variants that we qualify as thematic comparability measures. We evaluate these quantitative measures following a protocol based on the gradual degradation of a parallel corpus. Then, a new method to improve the co-clustering and co-classification of bilingual documents, as well as the alignment of comparable clusters, is developed. This approach merges native similarities defined in each language space with the similarity that is induced by a comparability measure. Finally, we propose an integrated approach, based on the above mentioned contributions, in order to assist the construction from the WEB, of thematic bilingual comparable corpora of ?good quality?. This procedure comprises a step of manual validation to ensure the quality of the comparable clusters alignment. Tuning the alignment comparability threshold, thematic comparable corpora with various comparability levels can be provided according to some specified requirements. The experiments that we have conducted on RSS feeds collected from major international newspapers appear relevant and promising.

Keywords: Thematic comparable corpora, Comparability measures, Co-clustering and co-classification of bilingual documents, Clusters alignment, Constitution of comparable corpora

n d'ordre : 000000000

Université de Bretagne Sud

Centre d'Enseignement et de Recherche ENSIBS - rue Yves Mainguy - 56000 VANNES

Téléphone : + 33(0)2 97 01 70 70 Fax : + 33(0)2 97 01 70 70