# Détection de zones parallèles à l'intérieur de multi-documents pour l'alignement multilingue

Charlotte Lecluze, Romain Brixtel, Loïs Rigouste, Emmanuel Giguet, Régis Clouard, Gaël Lejeune, Patrick Constant

# Détection de zones parallèles à l'intérieur de multi-documents pour l'alignement multilingue

Charlotte Lecluze[1], Romain Brixtel[1], Loïs Rigouste, Emmanuel Giguet[1], Régis Clouard[1,3], Gaël Lejeune[1] et Patrick Constant[2]
(1) GREYC - CNRS UMR 6072 - Université de Caen Basse-Normandie, Caen, France
(2) Pertimm, Asnières-sur-Seine, France
(3) EnsiCaen, Ecole Nationale Supérieure d'Ingénieurs de Caen, France
prenom.nom@unicaen.fr, prenom.nom@pertimm.com,
prenom.nom@ensicaen.fr

RÉSUMÉ

Cet article aborde une question centrale de l'alignement automatique, celle du diagnostic de parallélisme des documents à aligner. Les recherches en la matière se sont jusqu'alors concentrées sur l'analyse de documents parallèles par nature : corpus de textes réglementaires, documents techniques ou phrases isolées. Les phénomènes d'inversions et de suppressions/ajouts pouvant exister entre les différentes versions d'un document sont ainsi souvent ignorées. Nous proposons donc une méthode pour diagnostiquer en contexte des zones parallèles à l'intérieur des documents. Cette méthode permet la détection d'inversions ou de suppressions entre les documents à aligner. Elle repose sur l'affranchissement de la notion de mot et de phrase, ainsi que sur la prise en compte de la Mise en Forme Matérielle du texte (MFM). Sa mise en œuvre est basée sur des similitudes de répartition de chaînes de caractères répétées dans les différents documents. Ces répartitions sont représentées sous forme de matrices et l'identification des zones parallèles est effectuée à l'aide de méthodes de traitement d'image.

ABSTRACT

**Parallel areas detection in multi-documents for multilingual alignment**

This article broaches a central issue of the automatic alignment : diagnosing the parallelism of documents. Previous research was concentrated on the analysis of documents which are parallel by nature such as corpus of regulations, technical documents or simple sentences. Inversions and deletions/additions phenomena that may exist between different versions of a document has often been overlooked. To the contrary, we propose a method to diagnose in context the parallel areas allowing the detection of deletions or inversions between documents to align. This original method is based on the freeing from word and sentence as well as the consideration of the text formatting. The implementation is based on the detection of repeated character strings and the identification of parallel segments by image processing.

MOTS-CLÉS : détection et alignement de zones, appariement de N-grammes de caractères, corpus de multidocuments.

KEYWORDS: area detection and alignment, character N-grams matching, multidocuments corpora.

# 1 Introduction

Notre travail se situe dans le domaine de l'alignement, c'est-à-dire de la mise en correspondance d'éléments textuels sémantiquement équivalents entre des documents en relation de traduction. Nous appelons cet ensemble de documents un *multidocument* et chacun des documents qui le composent des *volets* [1]. L'opération traduisante réalisée par le traducteur humain vise à interpréter le sens d'un document donné dans la langue source et à produire un document sémantiquement équivalent dans une ou plusieurs langues cibles. Cette opération peut donner lieu à des modifications dans l'organisation interne des différents volets d'un multidocument. Dans l'état de l'art, cette question a été principalement traitée au niveau microscopique : ordre des mots dans la phrase, permutation ou suppression de phrases. Dans cet article, nous nous intéressons au contraire aux différences au niveau macroscopique. Nous étudions plus particulièrement les phénomènes d'inversion et de suppression/ajout qui rendent *asynchrones* certains documents traduits. Ces documents ne sont pas alignables par des techniques classiques. La figure 1 présente deux cas de traductions dans le même multidocument [2].

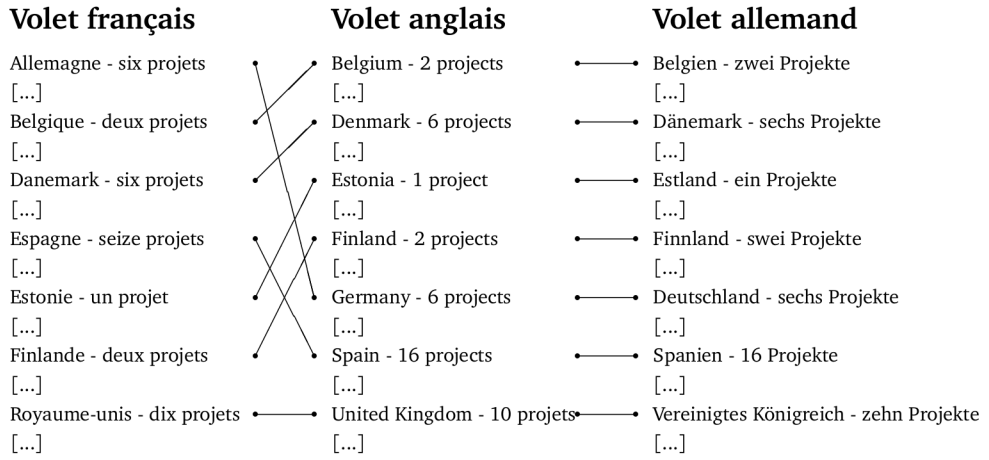


FIGURE 1 – Inversion et maintien de l'ordre entre les différents volets d'un multidocument

À gauche, la figure 1 présente un bi-document **asynchrone avec inversion massive** de plusieurs zones de textes entre les volets français et anglais (et par conséquent allemand) du même multidocument. À droite, l'alignement entre les volets allemand et anglais témoigne d'une traduction **synchrone**, tout est là et dans le même ordre. Dans le premier cas, nous considérons qu'il existe deux zones parallèles, correspondant dans chaque langue au volet dans son ensemble. Dans le second cas, le contenu sémantique est le même, mais l'ordre des différentes zones de texte n'est pas conservé, la traduction est **asynchrone** Nous considérons alors qu'il existe plusieurs zones entre lesquelles on recherche le parallélisme. Dans la figure 1, les paragraphes sont triés par ordre alphabétique ce qui provoque des différences d'ordre selon les langues. Des cas de suppression de zones de textes entre deux volets peuvent également être observés. Ces

1. Un bi-document est ainsi un cas particulier de multidocument contenant deux volets.
2. Communiqué de presse IP/05/1157 de l'Union Européenne. Les paragraphes sont triés par ordre alphabétique. Nous utilisons les [...] pour symboliser le contenu d'un paragraphe dont nous conservons ici seulement le début.

phénomènes d'inversion et de suppression constituent les principaux obstacles aux méthodes d'alignement qui reposent sur une hypothèse de parallélisme. Notre méthode s'attaque au problème de la détection en contexte du parallélisme, suivant la hiérarchie de grain : **document → zone → segment → N-grammes de caractères** [3]. Notre objectif est de rechercher et d'aligner les zones qui maximisent le parallélisme à l'intérieur de chaque bi-document : les *multizones*. Pour ce faire, nous proposons des outils de diagnostic de parallélisme afin de déterminer si un bi-document est synchrone, asynchrone avec suppression ou inversion. Notre méthode comprend quatre étapes :

**Appariement :** recherche des correspondances multilingues de N-grammes de caractères à partir d'une collection de multidocuments.

**Calcul de similarité :** comparaison des segments de textes de niveaux supérieurs et représentation sous forme de matrices de points (*dotplots*).

**Analyse :** détection des zones similaires entre deux volets, les *bizones*.

**Diagnostic :** qualification de la nature du parallélisme entre deux documents.

Plusieurs courants existent dans le domaine de l'alignement. Ils se distinguent notamment par le grain qu'ils proposent d'aligner : mots, phrases, paragraphes ou sections. Nous consacrons donc la section 2 à un rappel des principales méthodes d'alignement proposées à ce jour, avant de présenter le positionnement de nos travaux dans la section 3. Dans la section 4, nous décrivons les quatre étapes de notre méthode. Enfin, dans la section 5, nous présentons nos résultats en matière de diagnostic de parallélisme et d'alignement automatique de zones à partir de ces matrices.

## 2 Contexte

Les méthodes d'alignement sous-phrastique appliquées à des phrases alignées, si diverses soient-elles, trouvent toutes leur limite dans le fait qu'elles présupposent la disponibilité de corpus préalablement alignés en phrases (HANSARD, BTEC...). De tels corpus sont cependant peu nombreux et couvrent peu de langues. Des corpus où le grain aligné est plus gros (souvent le document) sont en revanche disponibles pour un très grand nombre de langues. Ils laissent réellement envisager la pratique d'opérations de rétro-ingénierie massives et peu supervisées sur ces documents issus du travail du traducteur humain. Ces pratiques permettent d'extraire des informations linguistiques et des ressources lexicales pouvant être utiles tant aux traducteurs, qu'aux lexicographes, aux linguistes ou aux terminologues.

Plusieurs méthodes d'alignement sous-phrastique à partir de documents non préalablement alignés en phrases ont été proposées (Simard *et al.*, 1993; Church, 1993; Dagan *et al.*, 1993). Les auteurs établissent un lien entre la similitude de graphie et la similitude de sens (cognats). Ces cognats servent de point d'ancrage pour un alignement de texte. Néanmoins, si ces similitudes sont fréquentes au sein d'une même famille de langues, elles s'avèrent plus rares entre les langues de familles différentes. Le problème existe également pour des langues ne partageant pas le même système d'écriture. En outre, ces méthodes reposent sur l'hypothèse centrale de parallélisme (Melamed, 1997; Simard & Plamondon, 1996). Or, cette hypothèse réduit considérablement le champ d'application des méthodes d'alignement sous-phrastique.

3. Nous utilisons N de façon générique, sa valeur n'étant pas prédéfinie.

À travers le système K-vec, Fung & Church (1994) ont proposé une méthode d'alignement de documents basée sur une similarité de répartition de mots. L'idée de K-vec est de découper chacun des deux volets en portions égales (*K-segments*) et d'affecter à chaque mot de chaque texte, un vecteur avec K dimensions (K-vec). K-vec fait l'hypothèse que si deux mots sont traductions l'un de l'autre, ils apparaissent dans les mêmes segments que deux mots qui ne le sont pas. K-vec semble être le premier système sans présupposé sur la présence de cognats ou les limites de phrases. Cependant, les systèmes reposant sur la similitude de répartition de mots se heurtent à la nature flexionnelle de certaines langues, un même mot pouvant alors recouvrir plusieurs formes selon sa fonction dans la phrase. En outre, K-vec suppose la linéarité de la traduction entre les volets, ce qui n'est pas toujours le cas, notamment sur les paires de textes en langues asiatiques ou en langues de la famille indo-européenne traitées par les auteurs. Enfin, des phénomènes d'ajouts/suppressions peuvent également interférer. Pour de meilleurs résultats, Fung & Mckeown (1994) ont implémenté **D**ynamique de K-vec (DK-vec) qui produit un petit dictionnaire dont les entrées sont utilisées comme des ancres pour l'alignement.

(Bourdaillet & Ganascia, 2007) abordent la question de l'alignement monolingue de textes comprenant des *déplacements*. Plus précisément l'étude porte sur les différentes versions laissées par un écrivain d'une de ses œuvres, c'est-à-dire les brouillons successifs. Aligner en monolingue ces réécritures revient à calculer une *distance d'édition avec déplacements*. En effet, les trois opérateurs de la distance de Levenshtein (insertions, suppressions et remplacements) ne suffisent alors pas à décrire les phénomènes potentiellement observables. Ces travaux constituent une amorce de recherche sur la question d'une méthode d'alignement prenant en charge les déplacements de portions de texte entre deux volets d'un bi-document. Cependant, la tâche se trouve grandement simplifiée par son contexte monolingue. L'hypothèse qu'une même graphie recouvre le même sens dans les deux versions est directement exploitable et la multiplication des hapax simplifie la tâche.

Enfin, plusieurs de ces auteurs ont proposé de reporter sur des matrices de points (*dotplots*) les appariements ainsi révélés (Church & Helfman, 1993). Le problème de l'alignement est ainsi transformé en un problème de traitement d'image (Chang & Chen, 1997). Notons que des hypothèses similaires ont été exploitées pour la détection de plagiat (Brixtel *et al.*, 2010).

# 3 Positionnement

Les travaux que nous présentons se situent dans la lignée des travaux d'alignement de documents précédemment cités. Nous utilisons également des *dotplots* pour diagnostiquer si la traduction est globalement littérale entre deux volets. Nos travaux se distinguent néanmoins par la granularité choisie pour amorcer le traitement des documents. Comme Cromières (2006) et Mcnamee & Mayfield (2004), nous nous intéressons aux N-grammes de caractères répétés, mieux à même de révéler des similitudes à la fois monolingues et multilingues (Lecluze, 2011). Nous étendons cependant la portée de la méthode **en l'appliquant aussi bien au contenu textuel qu'à la Mise en Forme Matérielle (MFM)** (Brixtel, 2011).

Le corpus que nous utilisons est constitué de communiqués de presse de l'Union Européenne [4] au format HTML. Nous présentons des résultats sur 6 couples de langues : français-espagnol (fr,es), français-grec (fr,el), français-finnois (fr,fi), français-anglais (fr,en), français-allemand (fr,de) et

4. Ces communiqués sont disponibles sur le site Europa, le portail de l'Union Européenne : `http://europa.eu/`.

français-danois (fr,da). Ces couples représentent un échantillon de familles de langues proches et éloignées. Nous supposons en effet que plus les langues sont génétiquement éloignées, plus il sera difficile de les rapprocher d'un point de vue strictement lexical. Nous introduisons également le grec pour attester que notre méthode est robuste à l'absence de similitudes de graphie.

# 4 Méthodes d'appariement et de détection de parallélisme

## 4.1 Appariement multilingue de N-grammes de caractères

Notre travail se situe dans la lignée de ceux de Cromières (2006), nous procédons à une recherche de N-grammes de caractères répétés en contexte, des *populations*. Les populations sont déduites d'un tableau de suffixes. Elles sont obtenues en calculant des motifs sans trou tels que décrits par (Ukkonen, 2009) [5]. Ces chaînes possèdent les caractéristiques suivantes :

**répétées :** les chaînes ont un effectif de 2 ou plus ;

**maximales :** les chaînes ne peuvent être étendues à gauche ou à droite sans perdre une occurrence.

L'intérêt de ces chaînes est double : révéler des facteurs communs monolingues au delà des mots graphiques et mettre en évidence des correspondances multilingues.

| Langue | Mots graphiques signifiant « transport » et leur effectif |
|---|---|
| fr | transports (3), transport (3) |
| es | transporte (5), transportes (1) |
| el | μεταφορών (3), μεταφοράς (1), μεταφορές (1), μεταφορά (1) |

Table 1 – Liste des mots graphiques signifiant « transport » dans un échantillon de textes en français, espagnol et grec ainsi que leurs effectifs entre parenthèses.

Ici, comme en témoigne le tableau 1, les écarts d'effectifs entre des mots alignés dans un échantillon sont déjà considérables. Or si l'on s'intéresse désormais aux répétitions de chaînes de caractères, il existe dans chaque langue une sous-chaîne commune à l'ensemble des équivalents sémantiques de « transport ».

| Langue | Chaînes de caractères répétées signifiant « transport » | Effectif |
|---|---|---|
| fr | transport- (3+3) | 6 |
| es | transporte- (5+1) | 6 |
| el | μεταφορ- (3+1+1+1) | 6 |

Table 2 – Chaînes de caractères répétées maximales communes aux mots signifiant « transport » dans le même échantillon de textes (fr, es et el) et leur effectif respectif.

Notre méthode consiste à obtenir de façon endogène et indépendante des langues une série de points d'ancrage entre deux volets : des **appariements**. Un appariement est une correspondance sémantique fortement généralisée telle qu'on en trouve par exemple dans un dictionnaire. Par

5. Les outils permettant le calcul de ces chaînes sont disponibles ici : `https://code.google.com/p/py-rstr-max/`

extension, l'appariement en tant que méthode est la mise en correspondance de chaînes de caractères répétées entre des multidocuments : des **populations**. Nous utilisons la similitude de répartition de ces chaînes (effectifs et positions dans la collection).

Ainsi, nous calculons les appariements entre chaînes de caractères de langues différentes, en prenant en compte des similitudes de répartitions sur l'ensemble des bi-documents de la collection. Les collections que nous constituons sont composées de 40 multidocuments chacune. Un exemple de répartition pour deux N-grammes de caractères est donné sur le tableau 3.

| langue | N-gramme | effectif corpus | effectif par multidocument | | | |
|---|---|---|---|---|---|---|
| | | | $doc_0$ | $doc_1$ | [...] | $doc_{199}$ |
| el | '_αερολιμέν' | (23) | 4 | 2 | [...] | 3 |
| fr | 'aéroports' | (21) | 4 | 2 | [...] | 2 |

TABLE 3 – Exemple de répartitions de deux N-grammes de caractères grec et français. Les espaces blancs sont représentés par le caractère « _ ».

Afin de limiter l'explosion combinatoire induite par une comparaison exhaustive de toutes les chaînes répétées maximales, nous comparons simplement les chaînes d'effectifs proches. Nous utilisons une distance L1 normalisée. Cela consiste à faire pour deux N-grammes de caractères ($s_1$ et $s_2$) de deux langues différentes, le rapport entre la somme des différences d'effectifs par document et la somme des effectifs des deux N-grammes dans la collection de bi-documents dans ces langues soit :

$$distance(s_1,s_2) = \frac{\sum_{doc} |effectif(s_1,doc) - effectif(s_2,doc)|}{effectif\_corpus(s_1) + effectif\_corpus(s_2)}$$

Ce calcul de distance permet de produire des appariements de populations de N-grammes de caractères avec une distance située dans $[0,1]$. Une distance de 0 signifie que deux N-grammes ont des répartitions identiques dans le corpus. C'est à partir des distances entre N-grammes que nous calculons des similarités entre les segments les contenant. Cette distance permet de calculer des correspondances fortement généralisées dans une collection de multidocuments ou *multizones*. Elle rend le traitement insensible aux différences d'ordres entre les volets et aux suppressions locales de zones de textes. Nous donnons quelques exemples d'appariements ainsi calculés dans la figure 2.

Les appariements obtenus lors de la première étape du processus servent à calculer la similarité entre des paires de segments d'une taille arbitraire fixée à 1% de la taille des volets d'origine. Dans notre hiérarchie de grain, ces segments correspondent au grain inférieur aux zones que nous cherchons à construire.

## 4.2 Calcul de la similarité entre les segments d'un bi-document

Tout d'abord, nous introduisons quelques définitions sur les segments que nous traitons. Soit $\Sigma$ un alphabet. Un *document* est un élément de $\Sigma^*$. Un *segment* est une sous-partie d'un document que nous exprimons relativement à la taille du document. Ainsi, le segment (d,(3,4),(0,1)) est l'ensemble des caractères débutant à la position 3,4 et de longueur 0,1. Ces positions sont exprimées en pourcentage par rapport à la taille $|d|$ du document. La segmentation obtenue pour un document $d_1$ est notée $S_1 = (s_1^1, \ldots, s_n^1)$. Nous segmentons nos documents en 200 segments correspondant à 1% de texte. Ces segments se chevauchent donc, et pour la même segmentation

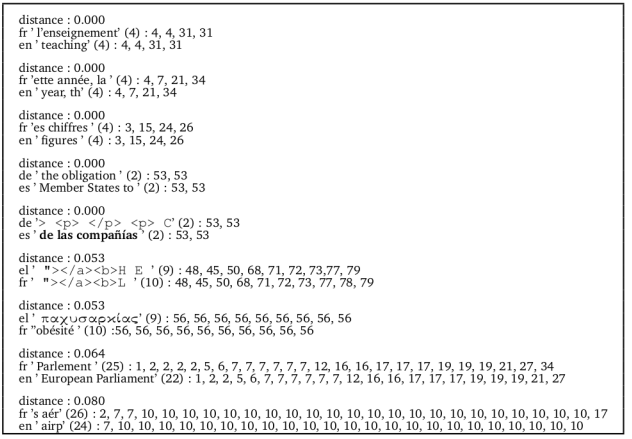


FIGURE 2 – Appariements de populations de chaînes de caractères répétées dans la collection. Chaque groupe de 3 lignes présente : ligne 1, la distance qui a été calculée entre deux chaînes de caractères sur la collection, lignes 2 et 3, respectivement pour la chaîne 1 et la chaîne 2 : la langue, la **'chaîne'**, son (effectif dans la collection) et la liste des identifiants de multidocument dans lesquels elle apparaît.

appliquée sur deux documents $d_1$ et $d_2$ d'un bi-document, $S_1 = (s_1^1, \ldots, s_n^1)$ et $S_2 = (s_1^2, \ldots, s_n^2)$ nous obtenons une matrice de similarité $\mathscr{M}^{(d_1,d_2)}$ de taille $n \times n$.
C'est en fonction de la répartition des segments similaires sur toute la matrice $\mathscr{M}^{(d_1,d_2)}$ que nous jugeons du parallélisme entre deux documents $d_1$ et $d_2$. Pour ce faire, nous définissons la similarité entre deux segments $i \in S_1$ et $j \in S_2$ via la fonction suivante : $\mathscr{M}^{(d_1,d_2)}_{(i,j)} = \frac{nb\_liens(i,j)}{max\_liens(i)}$
$nb\_liens(i,j)$ représente le nombre d'appariements ayant une distance inférieure à 0,1 mettant en jeu des N-grammes de caractères inclus dans les segments $i$ et $j$. $max\_liens(i)$ représente le nombre de liens maximum entre le segment $i$ et tous les segments de $S_2$ (Tableau 4).

| *Segments($S_2$)* | [0] | [0.05] | [0.1] | [0.15] | [0.2] | [...] | [0.75] | [0.8] | [0.85] | [0.90] | [0.95] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Nombre de liens* | 14 | 3 | 0 | 0 | 0 | ... | 0 | 0 | 2 | 0 | 0 |

TABLE 4 – Illustration de $max\_liens(i)$, $max\_liens$ vaut ici 14, le maximum sur la ligne

Dans la mesure où nous ne supposons pas de parallélisme initial, nous considérons l'ensemble des liens possibles entre les occurrences des N-grammes appariés sans nous focaliser sur un espace de recherche précis.

## 4.3 Représentation en deux dimensions de la similarité de deux volets

Les figures 3, 4 et 5 donnent des exemples de matrices $\mathscr{M}^{(d_1,d_2)}_{(i,j)}$ représentées en image en niveau de gris. Une similarité maximale est représentée par un pixel noir. Plus un pixel est clair, plus les segments associés sont différents selon notre fonction de similarité. Ainsi, quand deux documents sont traduits de façon globalement littérale, une diagonale se dessine de l'angle supérieur gauche à l'angle inférieur droit de la matrice (figure 3). Une diagonale cassée signifie l'existence d'inversions dans l'ordre de la traduction (figure 4).

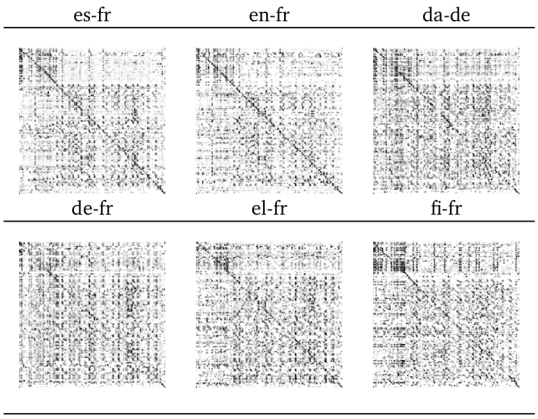


FIGURE 3 – Cas de volets synchrones (communiqué de presse IP/05/1156). L'ordre de traduction est globalement conservé, on observe une diagonale au centre de la matrice.

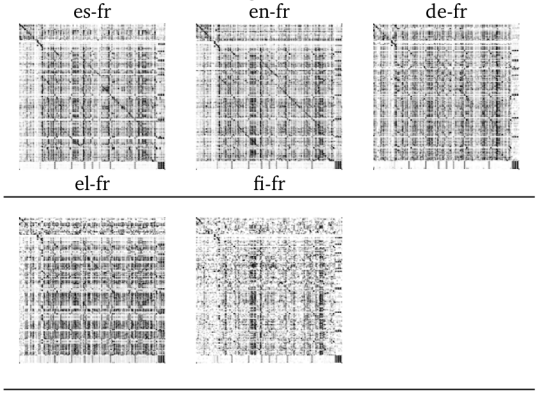


FIGURE 4 – Cas de volets avec inversion (communiqué IP/05/1157). L'ordre de la traduction est massivement inversé. La diagonale est « éclatée » en plusieurs segments de droites. (Pas de matrice da-fr, puisque le bi-document da-fr est synchrone).

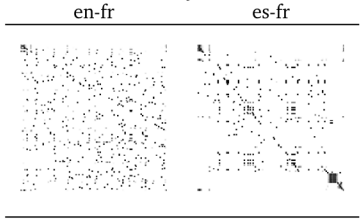


FIGURE 5 – Volets avec suppression (communiqués de presse IP/05/473 et IP/05/1558). On observe plusieurs segments de droites à l'intérieur de la matrice ayant des angles différents de celui de la première diagonale.

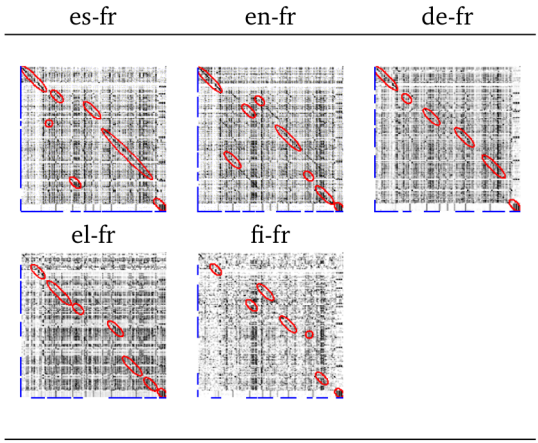


FIGURE 6 – Détection de segments de droites sur des cas de volets avec inversion. Il n'y a pas de matrice da-fr puisque le bi-document da-fr est synchrone.

Nous pouvons constater que pour un même jeu de paramètres pour tous les couples de langues, les résultats se dégradent à mesure de l'éloignement génétique des langues. La visibilité des droites et des segments de droite dans les figures 3 et 4 se dégrade entre la ligne 1 et la ligne 2. Le couple français-grec montre que la méthode s'accommode de différences morphologiques (pas de nécessité d'avoir des cognats). Les résultats sur ce couple sont proches de ceux du couple français-espagnol. En revanche, les résultats sur le couple français-finnois sont moins nets. L'analyse au niveau des caractères n'a pas permis de pallier totalement les différences sur le concept de mot entre ces deux langues. La différence de richesse lexicale entre ces langues joue pour beaucoup dans nos résultats. Le finnois fait un faible usage de la synonymie, comparativement au français par exemple, ce qui donne lieu à des différences de distributions conséquentes entre des unités pourtant sémantiquement équivalents. Celles-ci sont difficiles à appréhender au grain document et à calculer à partir d'une collection. Cette hypothèse devra être vérifiée en comparant des langues plus proches ou de façon plus générale en comparant davantage de couples. Ainsi, les différents phénomènes linguistiques interférant dans les résultats pourront être apprécier plus finement. La recherche des bons paramètres pour chaque couple de langue pourra mettre en évidence des similarités/dissimilarités entre famille de langue. À partir de ces résultats, nous nous focalisons sur l'analyse de l'image complète représentant ces matrices afin de détecter si deux documents sont parallèles ou s'ils contiennent d'éventuelles inversions ou suppressions de zones de textes.

## 4.4 Détection automatique des segments de droites

L'objectif de cette étape est double : mettre graphiquement en évidence les segments de droites et récupérer les informations propres à ces segments (positions, longueurs...). Nous calculons ces segments au moyen de la transformée de Hough. Ces informations seront utilisées pour l'étape de diagnostic suivante. Nous présentons dans les figures 6 et 7 des exemples de détection sur les cas de bi-documents asynchrones précédemment exposés.

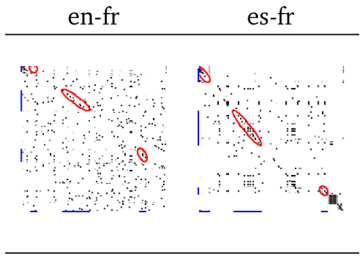


FIGURE 7 – Détection de segments de droites sur des cas de volets avec suppression.

## 4.5 Diagnostic de parallélisme entre des documents traduits

Les informations recueillies grâce au traitement d'image sont utilisées pour établir automatiquement un diagnostic de parallélisme. Nous définissons **quatre diagnostics** :

**volets synchrones** : pas d'inversion, ni de suppression de détectée (figure 3).

**matrices indéfinies** : $\frac{lt}{ld}$ <0,2 où $lt$ est la longueur totale des segments de droite et $ld$ la longueur de la diagonale.

**volets asynchrones avec inversion** : les coordonnées des segments (en rouge) dans une des deux dimensions ne se suivent pas (figures 4 et 6).

**volets asynchrones avec suppression** : la différence de longueur des projections sur les axes (en bleu) est supérieure à 2,5% (figures 5 et 7).

## 4.6 Alignement de zones : retour aux textes

Les informations recueillies grâce au traitement d'image sont également utilisées pour permettre le retour aux textes. Nous présentons ici deux exemples : le premier correspond à la matrice en-fr de la figure 7, le second à la matrice en-fr de la figure 6, ainsi qu'à la figure 1.

**Alignement de zones sur un bi-document asynchrone avec suppression**. Le tableau 5 illustre un cas de suppression dans un des deux volets, le volet en français, correspondant à environ un tiers de la taille du volet en anglais (2120 caractères). Si la suppression a bien été diagnostiquée, l'alignement de zones n'est lui que partiellement correct. La multizone 2 correspond exactement à l'attendu. Le cœur des multizones 1 et 3 sont corrects mais les contours restent mal définis.

**Alignement de zones sur un bi-document asynchrone avec inversion**. Le tableau 6 illustre un cas de différences d'ordre entre les zones de textes de deux volets. L'ordre des présentations des projets listés par pays respecte l'ordre alphabétique des noms des pays concernés. Tous les segments de droites de la matrice ont été mis en évidence, l'alignement de zones découlant des segments est globalement correct.

| | *IP/05/473* | |
|---|---|---|
| | *fr* | *en* |
| *Multizone 1* | rtations de textiles chinois  </h1>    M. Peter Mandelson, commissaire responsable du commerce, a annoncé ce jour qu'il avait décidé de demander à la Commi | <document celex="IP-05-473" lang="en"> <palign="right"> IP/05/473    Brussels, 24 April 2005  <h1> <a name="Heading4"></a>  European Commission launch |
| *Multizone 2* | les de sauvegarde. Elle entamera parallèlement des consultations immédiates avec la Chine pour tenter de dégager une solution satisfaisante.     Peter Mandelson a déclaré : «Nous venons de recevoir les statistiques d'importation des États membres pour le premier trimestre 2005. Elles sont très préoccupantes pour plusieurs catégories de produits textiles et d'habillement. Face à cette situation, l'Europe ne peut rester les bras croisés et assister à la disparition de son industrie. Notre enquête me permettra de décider s'il convient que l'UE adopte des mesures de sauvegarde. Il faudrait certes laisser les exportations chinoises croître à un rythme normal à la suite | the EU should impose special safeguard measures. In parallel, it will launch immediate consultations with China in an attempt to find a satisfactory solution.     Peter Mandelson said : "Member States have finally made available the import statistics for the first quarter of 2005. In several categories of textile and clothing imports they do give cause for serious concern. Based on these facts, Europe cannot stand by and watch its industry disappear. Our investigation will enable me to decide whether the EU should introduce safeguard measures. Chinese exports should, of course, be allowed to grow at a normal speed following the removal of quotas. But we must also extend protection to European industry if it is faced with a rui |
| *Multizone 3* | ssi une action. Les données d'importation concernant un certain nombre d'autres catégories semblent préoccupantes, mais exigent une analyse plus approfondie, actuellem | he global trade in textiles on 1 January 2005. This clause allows for short-term protective measures until the end of 2008.    Next Steps    These investigations will last for a maximum of 60 days, of which the first 21 will be used to take submissions from parties. The Commission will make a thorough assessment of market impact in the affected product categories. During this period, the Commission will also hold informal consultat |

TABLE 5 – Alignement de zones entre les volets fr et en du communiqué IP/05/473 avec suppression détectée à l'aide de notre méthode.

# 5 Évaluation

## 5.1 Évaluation 1

Le corpus que nous utilisons est constitué de 213 multidocuments. Pour chacun de ces multidocuments nous étudions 6 couples de langues (fr,es), (fr,el), (fr,fi), (fr,en), (fr,de) et (da,de). Évaluer le diagnostic des bi-documents (indéfini, synchrone, asynchrone avec inversion ou asynchrone avec suppression) n'est pas une tâche triviale, en effet il n'existe pas de références pour évaluer la détection de multizones. Nous présentons les résultats obtenus à partir d'une référence établie pour trois collections « tout venant » d'une part et trois collections thématiques d'autre part constituées à partir de notre corpus (Tableau 7). Les expériences réalisées sur ce premier corpus ont confirmé que :

– l'utilisation de la MFM améliore le taux de décision de +15% (+10% sur les couples de langues proches et +20% sur les couples de langues éloignées) ;
– les similitudes de répartition de chaînes de caractères répétées permettent d'aligner des documents, y compris dans des langues éloignées avec un taux de décision toutefois plus faible sur les langues éloignées : -11% ;
– exploiter une collection de multidocuments thématiquement proches contribue faiblement à l'amélioration : +3% de précision sur les documents synchrones.

À titre comparatif, on peut préciser que nos résultats sont de 2 à 7% meilleurs qu'une *baseline* prenant comme hypothèse que tous les documents parallèles sont synchrones. Ainsi, le système s'avère très précis et assez pertinent pour les documents synchrones. Nos résultats sur les documents asynchrones (10% du corpus) sont moins satisfaisants.

| | IP/05/1157 | |
|---|---|---|
| | *fr* | *en* |
| *Multizone 1* | Bruxelles, le 19 septembre 2005  <h1> <a name="Heading4" id="Heading4"></a> Environnement : la Commission subventionne 89 projets d'innovation dans 17 pays pour un montant de 71 millions d'euros  </h1>  <i>La Commission européenne a approuvé le financement de 89 projets innovants dans le domaine de l'environnement dans 17 pays, au titre du programme LIFE-Environnement 2005. [...] Pour plus de détails concernant chaque projet, consulter le site suivant :<br /> <a href="http://europa.eu.int/comm/environment/life/project/index.htm">http://europa.eu.int/comm/environment/life/project/index.htm</a>   ANNEXE    Résumé des projets | /a> Environment : Commission supports 89 innovation projects in 17 countries with €71 million  </h1>  <i> The European Commission has approved funding for 89 environmental innovation projects in 17 countries under the LIFE-Environment programme 2005. [...] More information<br /> See the annex for a summary of the 88 projects funded under LIFE-Environment. More detailed information on each project is available at :   <a href="http://europa.e |
| *Multizone 2* | r appliquera une stratégie intégrée pour réduire la pollution agricole diffuse, dans le sens de la directive cadre sur l'eau <a href= "i05_1157.frr.html #_Ref111348773">1</a>.   Le second [...] Le second projet concerne le prétraitement de la laine dans la production de fil. L'objectif principal est de supprimer les émissions de composés organohalogénés absorbables (AOX) et de réduire sensiblement l'utilisation de produits chimiques dans le processus de nettoyage, grâce un procédé durable de prétraitement par plasma.   Un projet porte sur la gestion des déchets e | ht"> ANNEX    Overview of LIFE-Environment projects 2005 by country     Belgium – 2 projects [...] Denmark – 6 projects [...] Estonia – 1 project [...] the fermentation of manure, processing of bio-gas into |
| *Multizone 3* | er les tôles laminées à froid. Un nouveau procédé basé sur la technologie sous vide à haute pression et n'utilisant pas de produits chimiques sera employé.    Belgique – deux projets [...] Danemark – six projets [...] Espagne – seize projets    Trois projets portent sur la  gestion des eaux . Le premier permettra de définir un modèle d | tronic equipment, in line with EU legislation <sup><a name="fnB2" href= "fn2" id="fnB2">[2]) </a></sup>, with a particular emphasis on rural areas.   The second targets households, schools and day-care centres in Helsinki, with a view to increasing awareness and ensuring the amount of waste produced does not exceed 2003 levels.   France – 11 projects [...] The sixth will substitute lead with o |
| *Multizone 4* | s variétés d'amandiers capables de résister à de telles conditions.   Le troisième projet vise à définir un système de gestion durable de la viticulture de montagne, en vue de réduire les incidences de cette activité sur le paysage, les sols et les ressources en eau.   Quatre projets traitent des technologies propres. [...] Le sixième projet démontrera qu'il est techniquement et économiquement possible d'appliquer un nouveau procédé à haute capacité pour séparer les alliages métalliques à pureté élevée (plus de 90%). Utilisé pour extraire le fer, l'aluminium et les métaux lourds contenus dans les véhicules hors d | to reduce diffuse pollution from agriculture, in support of the Water Framework Directive<a href= "i05_1157.enr.html #_Ref111348773">1</a>.   The second [...] The second concerns the pre-treatment of wool in yarn production. The main goal is the elimination of emissions of absorbable organic halides (AOX) and a significant decrease in the use of chemicals in the cleaning process, through a sustainable plasma pre-treatment process.   One project addresses waste management</ |
| *Multizone 5* | nouvelle technologie recourant à la fermentation du lisier, à la transformation du biogaz en énergie et en chaleur « écologiques » et à la séparation intégrale des composants recyclables et non recyclables.   Finlande – deux projets [...] France – onze projets [...] Le quatrième projet vise à démontrer qu'il est techniquement possible de recourir à la technologie des ultrasons pour réduire la production de boues résiduaires dans les stations d'épuration des eaux usé | ng of cold rolled plates. A new chemical-free process will be used, based on high-pressure vacuum technology.    Greece – 4 projects [...] Hungary – 1 project    The project, covering water management, assesses the scale of arsenic contamination in groundwater in the southern part of Hungary. It will develop a pilot management plan, incorporating a new arsenic removal technology.   Ireland – 2 projects [...] Italy – 15 projects [...] Netherlands – 7 projects [...] Portugal – 2 projects [...] Romania – 1 project [...] Spain – 16 projects [...] The third aims at defining |
| *Multizone 6* | ernier projet français concerne la gestion de la qualité de l'air. Il vise à mettre au point un échantillonneur d'air basé sur une nouvelle méthode de surveillance des pollens dans l'air. Au lieu de quantifier les grains de pollens selon leur morphologie, cette méthode reposera sur la mesure en ligne de l'antigénéité/l'allergénéité.   Grèce – quatre projets [...] Hongrie – un projet [...] Irlande – deux projets [...] Italie – quinze projets [...] Luxembourg – un projet [...] Pays-Bas – sept projets [...] Portugal – deux projets [...] Roumanie – un projet [...] Royaume-Uni – dix projets [...] Le quatrième projet vise à réduire l'élimination des déchets hospitaliers non stériles dans les | g a mountain viticulture sustainable management system in order to reduce the environmental impacts of this activity on landscape, soil and water resources.   Four projects deal with clean technologies. [...] The last project will demonstrate the technical and economic feasibility of a new high-capacity process to separate high purity metal alloys (>90%). Used for the separation of iron, aluminium and heavy metals from |
| *Multizone 7* | s incidences environnementales des activités économiques. Le premier vise à démontrer l'efficacité du recyclage de l'eau au moyen d'un nouveau réacteur de digestion aérobie des eaux usées.   Le second projet concerne l'exploitation des friches industrielles pour la culture de biomasse à des fins énergétiques, la réhabilitation des terres endommagées et la production de chaleur et d'énergie à partir de sources d'énergie renouvelables. [...] Suède – deux projets [...] Directive 2002/95/CE du Parlement européen et du Conseil du 27 janvier 2003 relative à la limitation de l'utilisation de certaines substances dangereuses dans | re-use.   A fourth project aims to reduce the disposal of non-sterile clinical waste in landfill sites and promote its use as a raw material for recycled products.   Two projects seek to mitigate the environmental impact of economic activities. One will demonstrate the effectiveness of water recycling using a new reactor for aerobic digestion of wastewater.   A second aims to re-use brownfield sites to grow biomass energy crops, restore damaged land, and generate heat and power from renewable energy sources.[...] Council Directive 1999/13/EC of 11 March 1999 on the limitation of em |

TABLE 6 – Alignement de zones entre les volets fr et en du communiqué IP/05/1157 présentant une différence d'ordre des zones détectées à l'aide de notre méthode.

| #(Moyenne) | da | de | el | en | es | fi | fr |
|---|---|---|---|---|---|---|---|
| #caractères ($10^3$) | 8,8±15,6 | 9,1±14,2 | 9,7±15,7 | 8,5±15,4 | 9,4±1,5 | 8,9±15,5 | 9,6±15,6 |
| #mots | 979±1213 | 960±1124 | 1105±1245 | 997±1213 | 1148±1257 | 791±1167 | 1138±1262 |
| #paragraphes | 16,4±14,5 | 16,7±14,7 | 16,4±14,9 | 16,5±14,9 | 16,5±14,8 | 16,1±14,7 | 17,1±15,2 |

TABLE 7 – Description des 1491 documents du corpus (nombre de caractères, de mots et de paragraphes moyen ± écart-type)

### 5.2 Évaluation 2

Afin de proposer une autre évaluation, nous avons fabriqué deux jeux de test, chacun contenant 240 bi-documents dans les 6 couples de notre corpus. Après modifications, le premier jeu contient 60 bi-documents asynchrones avec inversion et le second 60 bi-documents asynchrones avec suppression.

Sur le premier jeu, nous avons obtenu un rappel de 25,4% et une précision de 41,6% ($F_1 - mesure = 31,6$). Sur le second, les résultats étaient meilleurs puisque nous avons obtenu un rappel de 57,1% avec une précision de 43,4% ($F_1 - mesure = 49,3$).

## 6 Conclusion

Les travaux que nous avons présentés dans cet article témoignent de la possibilité d'établir un alignement brut de documents traduits, grâce à un appariement de N-grammes de caractères répétés dans une collection de multidocuments. Ces unités sont révélées sans utilisation de ressources externes et de façon indépendante des langues en présence. Cela permet d'amorcer sans présupposé de parallélisme un alignement lexical de mots. L'étape de détection automatique des zones parallèles qui revient à un problème de traitement d'image s'avère au même titre que l'alignement plus difficile à mesure que les langues mises en confrontation s'éloignent. Ces travaux témoignent d'une part que l'alignement de corpus parallèles n'est pas un sujet clos et d'autre part que la combinaison d'indices utilisée permet bien de les exploiter sans présupposé de parallélisme. Ces travaux recèlent des perspectives tant opératoires que de recherche. Des perspectives opératoires en ce qui concerne l'établissement automatique des paramètres de création des matrices et de diagnostic des bi-documents, ainsi qu'au niveau de la détection des segments de droites des images qu'il faut affiner. En terme de recherche, ces travaux offrent des perspectives pour le contrôle a posteriori de traduction mais également pour le traitement de corpus comparables.

## Remerciements

Merci aux relecteurs, particulièrement pour la suggestion de fabriquer un nouveau jeu de données.

# Références


Bourdaillet J. & Ganascia J. (2007). Alignment of noisy unstructured data. In *IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data, January 6-12*, Hyderabad, India.

Brixtel R. (2011). *Alignement endogène de documents, une approche multilingue et multi-échelle*. PhD thesis, Université de Caen/Basse-Normandie.

Brixtel R., Fontaine M., Lesner B., Bazin C. & Robbes R. (2010). Language-Independent Clone Detection Applied to Plagiarism Detection. In *SCAM 2010* : IEEE Computer Society.

Chang J. S. & Chen M. H. (1997). An alignment method for noisy parallel corpora based on image processing techniques. In *Proceedings of the eighth conference on European chapter of the ACL*, p. 297–304, Spain.

Church K. W. (1993). Char_align : a program for aligning parallel texts at the character level. In *Proceedings of the ACL 93*, p. 1–8, Ohio.

Church K. W. & Helfman J. I. (1993). Dotplot : A program for exploring Self-Similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics*, **2**(2), 153–174.

Cromières F. (2006). Sub-sentential alignment using substring co-occurrence counts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, p. 13–18, Australia.

Dagan I., Church K. W. & Gale W. A. (1993). Robust bilingual word alignment for machine aided translation. *In proceedings of the workshop on very large corpora*, **1**, 1—-8.

Fung P. & Church K. W. (1994). K-vec : a new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, p. 1096–1102, Japan.

Fung P. & Mckeown K. (1994). Aligning noisy parallel corpora across language groups : Word pair feature matching by dynamic time warping. *In Proceedings of the first conference of the AMTA*, p. 81–88.

Lecluze C. (2011). Recherche d'une granularité optimale pour l'alignement multilingue : N-grammes de caractères ou n-grammes de mots ? In *JeTou, Journées d'études Toulousaines*, France.

Mcnamee P. & Mayfield J. (2004). Character N-Gram tokenization for european language text retrieval. *Information Retrieval*, **7**, 73–97. ACM ID : 961313.

Melamed I. D. (1997). A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the ACL*, p. 305–312, Spain : Association for Computational Linguistics.

Simard M., Foster G. F. & Isabelle P. (1993). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research : distributed computing - Volume 2*, p. 1071–1082, Canada.

Simard M. & Plamondon P. (1996). Bilingual sentence alignment : Balancing robustness and accuracy. *In proceedings of the 2nd conference of the AMTA*, **13**, 59–80.

Ukkonen E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theorie in Computer Science*, **410**(43), 4341–4349.