Extraction endogène d'une structure de document pour un alignement multilingue

Romain BRIXTEL
Laboratoire GREYC, Université de Caen-Basse-Normandie
Campus II, 14032 Caen Cedex
rbrixtel@info.unicaen.fr

Résumé. Pour des raisons variées, diverses communautés se sont intéressées aux corpus multilingues. Parmi ces corpus, les textes parallèles sont utilisés aussi bien en terminologie, lexicographie ou comme source d'informations pour les systèmes de traduction par l'exemple. L'Union Européenne, qui a entraîné la production de document législatif dans vingtaine de langues, est une des sources de ces textes parallèles. Aussi, avec le Web comme vecteur principal de diffusion de ces textes parallèles, cet objet d'étude est passé à un nouveau statut : celui de document. Cet article décrit un système d'alignement prenant en compte un grand nombre de langues simultanément (> 2) et les caractéristiques structurelles des documents analysés.

Abstract. For many reasons, the multilingual corporas have interested various communities. Among these corporas, the parallel texts are used as well in terminology, lexicography or as a source of informations for example-based translations. The European Union, which involved the production of legislative documents, generates these parrallel texts in more than twenty languages. Also, with the Web as a vector of diffusion, we can wonder if these parallel texts can be treated as documents. This article describes a alignment system taking account a great number of languages (> 2) and the structural characteristics of the analyzed documents.

Mots-clés : alignement multilingue, corpus parrallèles, multitextes, multidocuments, extraction de structures, alignement endogène.

Keywords: multilingual alignment, parallel corpora, multitexts, multidocuments, extraction of structures, endogenous alignment.

1 Introduction

L'alignement est une opération qui consiste à relier des unités qui se correspondent dans des textes parallèles. Des éléments peuvent alors être appariés suivant le grain d'analyse (paragraphes, phrases ou d'autres unités plus fines telles que les mots) sur lequel le système d'alignement s'appuie.

L'intérêt croissant porté à cet axe de recherche est lié à l'augmentation de la production des corpus multilingues regroupant des textes et leurs traductions (textes parallèles ou multitextes), ainsi que leur accessibilité de plus en plus aisée.

Dans cet article, nous présentons tout d'abord un état de l'art décrivant des méthodes d'aligne-

ments à différents niveaux de granularité. Ensuite, nous aborderons la méthode d'alignement multilingue (près d'une vingtaine de langue) implémentée en détaillant les hypothèses utilisées et en présentant les différents alignements générés

2 État de l'art

Deux grains principaux nous intéresse : l'alignement de phrases ainsi que l'alignement d'éléments sous-phrastiques tel que les mots, termes ou expressions.

2.1 Alignement de phrases

Pour l'alignement au grain phrase, deux méthodes principalement utilisées sont à retenir.

(Kay, 1988) part de l'hypothèse suivante : deux phrases sont alignables si les mots qui les composent sont en correspondance. En supposant que certaines phrases sont alignées au début du traitement (habituellement les premières et dernières phrases des documents du multidocument), le processus propose des candidats de phrases à aligner. Aussi, l'algorithme considère leurs places dans le document. On exclut la possibilité d'aligner une phrase se situant au début d'un document avec une autre à une place très eloignée dans l'autre document. Le processus d'alignement est itératif. L'algorithme valide les alignements phrastiques en satisfaisant au maximum les équivalences d'ensembles de mots faites dès le début et en cherchant des phrases équivalentes avec un rang très proche. Chaque nouvelle paire de phrases alignées restreint les domaines d'équivalences entre les mots de langues différentes.

(Gale & Church, 1993) se basent sur la propriété suivante : des phrases sont alignées si leurs longueurs relatives sont équivalentes. La méthode part du principe qu'une phrase longue écrite dans une langue sera traduite par une phrase longue dans une autre langue. La longueur des phrases est alors relative à la longueur moyenne, en nombre de caractères, des phrases de la langue dans laquelle elles sont écrites. De la même façon que la méthode de (Kay, 1988), on admet ici que les phrases alignées ont un rang très proche dans les deux documents analysés. Sont alors envisagés les alignements $1:2$ (une phrase correspond à deux autres), $2:1$, $1:0$ et $0:1$. Ainsi, en l'absence de candidat à l'alignement pour une phrase donnée, la méthode propose de regarder si elle correspond à plusieurs phrases suivant le même critère de longueur, en se limitant aux possibilités d'alignement décrits précédemment. Les possibilités de fusions ou d' omissions de phrases sont inclues par raison pratique en terme de simplification de calcul ; tout en suivant une certaine logique du processus de traduction.

Des méthodes d'ancrages lexicaux préalables sont utilisées en complément pour affiner les résultats. Ainsi, (Debili & Sammouda, 1992), (Simard *et al.*, 1992), (Melamed, 1999) utilisent les similarités graphiques entre les langues (cognates) pour mettre des mots en correspondance. En suivant l'hypothèse de (Kay, 1988), cette technique améliore les alignements phrastiques

Les hypothèses sous-jacentes aux méthodes d'alignement de type (Kay, 1988) et (Gale & Church, 1993), pour un alignement d'une paire de texte au grain phrase, peuvent être résumées de la façon suivante (Véronis, 2000) :

– l'ordre des phrases dans les deux textes est identique ou très proche ;
– les textes contiennent peu de suppressions ou d'adjonctions ;

– les alignements $1 : 1$ sont très largement majoritaires et les rares alignements $m : n$ sont limités a de petites valeurs de m et n (≈ 2).

Même si ces hypothèses paraissent légitimes, les systèmes d'alignement se basant dessus rencontrent de grandes difficultés si les documents comportent des différences structurelles (par exemple : date figurant au début d'un texte et à la fin d'un autre).

2.2 Alignement sous-phrastique

Dans le cadre de l'alignement en phrases, la mise en correspondance de mots n'est pas le but premier mais seulement un amorçage : on se satisfait ici d'un alignement grossier de mots. Si l'ordre des phrases de deux traductions est plus ou moins respecté, il en est autrement dès que l'on se situe au niveau de la phrase. L'insertion d'adverbes ou d'adjectifs dans une expression ou encore la permutation entre deux propositions d'une phrase ne permet pas de faire les mêmes suppositions que pour l'alignement au niveau de la phrase. De plus, les phénomènes d'agglutination et de flexions ébranlent les méthodes statistiques basées sur la comparaison de chaînes de caractères constantes

L'alignement en éléments sous-phrastiques consiste à détecter ces éléments puis à les mettre en correspondance. Or, séparer ces deux étapes est délicat car déterminer la nature exacte des éléments sous-phrastiques en jeu (par exemple : mots, expressions, chunk, propositions) dépend aussi bien de la langue cible que de la langue source (par exemple : "rideau de fer" se traduit par "Eisenvorhang" en allemand et "iron curtain" en anglais). L'utilisation d'outils statistiques est donc extrêmement délicat dans de telles conditions et choisir un de ces outils se révèle être du "cas par cas" en fonction des langues à aligner. De plus, la plupart des expressions sont seulement "semi-figées" et peuvent être altérées par certaines opérations (flexion, insertion d'adjectifs, passivation...).

(Giguet, 2005) considère le problème d'alignement sous-phrastique comme étant celui de l'appariement de suites de mots ayant une répartition similaire à travers les textes analysés. Ce choix de traitement statistique influe sur les résultats. La répartition de mots ayant une graphie identique à travers le document dépend des caractères flexionnel et agglutinant de la langue considérée. Une analyse morpho-endogéne des documents permet alors des alignements satisfaisants entre des documents anglais et grecque ; mais ne résout pas la comparaison entre deux documents de langues ayant un caractère agglutinant différent, tel que l'anglais et le finnois.

(Zimina, 2004) applique une approche textométrique sur des corpus multilingues. Sur un texte parallèle français-anglais-russe, elle montre que lorsque qu'un mot est doté d'un large éventail de sens dans le corpus, la comparaison des fréquences totales des formes graphiques ne constitue pas toujours une bonne indication pour l'appariement (par exemple : la correspondance d'un mot polysémique comme "droit"). L'étude des formes graphiques en contexte (avec leur voisinage, les mots les entourant) permet alors de lever certaines des ambiguïtés résiduelles. Au contact d'autres mots associés sur l'axe syntagmatique, différents composants du sens du mot sont activés et il devient possible d'en tenir compte lors de l'appariement. La répartition d'un mot et la répartition des mots voisins sont utilisés comme éléments de comparaison.

La difficulté principale de l'alignement à ce grain réside dans la difficulté à cerner les éléments que l'on veut aligner, les phénomènes linguistiques sont plus nombreux (sans pour autant que l'un se détache des autres) et sont différents suivant les langues. L'introduction de connaissances linguistiques est alors relativement coûteuse et l'on devient dépendant des langues traitées.

3 Quel grain pour un alignement multilingue ?

Un des points qui émerge de cette vision globale de l'alignement est que les traitements à un grain affectent ceux à d'autres grains. D'une part (Kay, 1988) se base sur ce que l'on peut appeler un alignement grossier au grain mot pour un alignement de phrases, d'autre part un alignement à un grain phrastique peut être un préalable à l'alignement de mots ou de suite de mots ((Zimina, 2004), (Giguet, 2005)).

Les indices majoritairement utilisés [1] proviennent d'une vision de l'alignement comparable à celle abordée sur les problèmes de traitements de flux de caractères ou de séquence de mots. D'autre part, l'alignement a été abordé comme un problème de découpe de textes et de mise en correspondance des morceaux découpés (typiquement, découper le texte en phrases puis trouver les correspondances entre elles, ou trouver la couverture maximale entre les éléments issus de la découpe de deux phrases alignées). La hiérarchie $texte \rightarrow paragraphe \rightarrow phrase \rightarrow expression \rightarrow mot$ est alors la plus utilisée pour définir les grains qui seront traités. L'originalité de la méthode proposée dans cette article est de se placer en parallèle de cette hiérarchie sans pour autant l'ignorer.

Le Web est le plus grand vecteur de diffusion de ces textes, il les affecte alors naturellement. Voir l'alignement autrement qu'en envisageant exclusivement les objets analysés comme des flux de caractères va dans ce sens : l'utilisation de liens hypertextes, d'images, de tableaux ou d'autres applications de mise en forme matérielle (MFM) telles que des marques de graisse et d'emphase, affecte le statut des traductions disponibles. En considérant ces traductions comme des documents, les multitextes décrits dans la littérature peuvent être abordés comme des "multidocuments".

Très peu de techniques prennent en compte l'aspect réellement multilingue des multidocuments (qui se résument dans la littérature majoritairement à des multidocuments de deux documents, ou bi-documents) qui est l'essence même des multidocuments. On peut se demander si à trop s'attacher à peu de langues, nous n'obtenons pas des méthodes de traitement ad hoc : le multilinguisme est une force qui permet l'abstraction alors que se cantonner à un nombre de langues réduit oriente les méthodes de traitement que l'on peut appliquer. (Simard, 2000) montre que l'ajout de langues rends plus fiable les alignements générés automatiquement, mais sa méthode se résume plus à l'utilisation conjointe de plusieurs bi-documents qu'un multidocument dans son ensemble.

Le nombre de langues peut agir comme un filtre suffisamment fort pour contraindre l'apparition de résultats. Aussi, même si l'on veut tendre vers un idéal multilingue maximal, il est toujours possible d'exclure une langue d'un multidocument du traitement. Le fait de remarquer qu'une langue ou qu'un type de document est "récalcitrant" à une hypothèse que l'on a formulé peut nous apprendre beaucoup sur le document/la langue (ou le groupe de langues/documents) exclu et la façon d'aborder le multidocument.

Collectant et diffusant des communiqués de presse de l'Union Européenne sous format électronique à travers leur site Web, Europa [2] permet de récupérer des multidocuments comprenant des documents écrits dans plus de vingt langues différentes. A partir d'un corpus d'étude [3] extrait de

[1] Longueur des phrases, contenu des phrases (principalement : mots ou suite de mots), position des phrases ou d'éléments sous-phrastiques (mots ou suite de mots), distance graphique d'éléments sous-phrastiques (cognates), patron linguistique d'extraction propre à une langue ou à une famille de langue, etc.

[2] `http://europa.eu/`

[3] 63 multidocuments de plus de 16 documents/langues chacun.

ce site, nous nous orientons vers un alignement prenant en compte simultanément le maximum de langues en parallèle à la hiérarchie $texte \rightarrow paragraphe \rightarrow phrase \rightarrow expression \rightarrow mot$. Il faut alors trouver les invariants qui peuvent être détectés dans tous les documents pour définir les bases d'un treillis d'informations robuste.

4 La MFM : un invariant de structure multilingue ?

Les documents extraits d'Europa se présentent sous la forme de documents XHTML [4]. De nombreuses traces non-textuelles peuvent y être repérées via la MFM telles que les tableaux, les séparations horizontales (sauts de ligne via la balise
 ou les traits horizontaux via la balise <hr/>), l'application de graisse et d'italique, voir des liens hypertextes [5]. L'avantage de ces marques est qu'elles mettent en valeur des zones à l'intérieur des documents. Une phrase ayant une MFM particulière, en plus d'être mise en valeur dans son intégralité, met aussi en avant chaque élément dont elle est composée (chunks, propositions, mots ou caractères). Une approche consiste à se demander si ces marques peuvent scinder le document en plusieurs parties. Par exemple, en détectant une ligne de séparation dans un document il est possible de considérer deux zones, celle avant la ligne et celle située après. Si cette séparation est présente dans tous les documents du multidocument, alors nous pouvons mettre en relation les zones ainsi dégagées.

Il parait douteux d'effectuer une dichotomie sur un document suivant l'apparition d'un élément sous-phrastique mis en valeur (par exemple : un mot mis en gras). Au grain sous-phrastique, sa position dans la phrase est affectée par les règles de construction propre à la langue du document. A ce grain, ces indices sont des marques caractérisant les grains dans lesquels ils sont inclus. Comme les cognates, ils sont détectés pour marquer ces grains (les phrases pour les cognates) et les aligner ; ils ne sont pas utilisés pour diviser le document.

Si le grain phrase n'est pas approprié, nous considérons un grain supérieur. Un alinéa [6] peut ainsi être mis en gras afin qu'il soit assimilé à un titre de partie. La mise en évidence de passage peut être une finalité en soit ou un moyen d'organiser le discours. Par l'utilisation de titres, de résumés ou de passages ayant une MFM différente de celle appliquée au corps de texte, le document est découpé en différents segments.

L'alignement peut être vu comme l'extraction d'équivalences sémantiques. Considérer cet usage de la MFM comme un vecteur de sens préservé dans le processus de traduction nous amène à exploiter ces marques en tant qu'invariant entre les documents de différentes langues.

Cette segmentation à un grain plus élevé que la phrase permet de restreindre les espaces de recherche [7] dans le cadre de la recherche d'équivalences sémantiques entre les documents d'un multidocument. Par exemple, l'alignement phrastique suppose que chaque phrase est alignable avec une autre ayant une position très proche dans le texte. Dégager des zones permettrait de cibler nos recherches pour dégager les équivalences : une phrase appartenant à une zone est potentiellement alignable avec les phrases des autres documents appartenant à la même zone du texte. Et si cette méthode parait viable dans le cadre de l'alignement phrastique, elle peut

[4] eXtensible HyperText Markup Language, recommandation w3c `http://www.w3.org/MarkUp/#recommendations`

[5] Les liens hypertextes ne sont pas des marques de MFM mais leurs mises en relief en font parti.

[6] `http://atilf.atilf.fr/` — Texte compris entre deux retours à la ligne.

[7] « Diviser pour régner ».

aussi l'être dans les grains contenus dans les zones de recherches détectées (les grains contenus peuvent être des paragraphes, des mots, des expressions...).

4.1 Caractérisation des alinéas par la MFM

Nous partons d'un découpage en alinéas des documents. Sur des documents XHTML, ces alinéas peuvent être détectés grâce aux balises de bloc (telles que <p>, <h1>, <h2>,) en opposition aux balises en ligne (comme , ,). D'autres méthodes comme celles utilisant les techniques de reconnaissances de texte [8] peuvent être utilisées pour arriver à cet objectif. Seules les MFM appliquées sur un alinéa dans son ensemble nous serviront à diviser les documents en parties.

La première étape consiste à détecter la MFM utilisée pour chaque alinéa afin de permettre une segmentation du document. Or, une MFM n'est pas systématiquement appliquée tout le long d'un alinéa, comme nous pouvons le remarquer sur l'exemple ci-dessous [9] :

[...] Ir iekl,autas TRIPS *un* PVO *prasi-bas, [...]*

Les acronymes "TRIPS" et "PVO" ont été mis en valeur dans cet alinéa en modifiant la MFM qui a été attribué en majorité. Si nous considérons les MFM appliquées à un alinéa seulement si elles sont présentes à tout endroit du texte, nous ne caractérisons pas correctement cet alinéa. Pour pallier ces problèmes, nous considérons la MFM d'un alinéa de la façon suivante :

– soit mfm_{deb} l'ensemble des MFM appliquées au premier mot de l'alinéa ;
– soit mfm_{fin} l'ensemble des MFM appliquée au dernier mot de l'alinéa ;
– la MFM caractérisant l'alinéa est définie par $\text{mfm}_{al} = \text{mfm}_{deb} \cup \text{mfm}_{fin}$.

Nous évitons les problèmes cités ci-dessus ainsi que ceux pouvant être provoqués par l'usage de lettrine ou encore ceux provoqués par un usage important d'une MFM dans un alinéa, comme par exemple dans le cas d'une citation en italique et occupant une grande partie de l'alinéa dans lequel elle est située.

4.2 Découpage du document par la MFM

La figure (FIG. 1) présente une visualisation des documents estonien (et), italien (it) et néerlandais (nl) [10] segmentés en alinéas . La segmentation est faite sur 20 langues même si nous n'en montrons ici que 3. Un alinéa est représenté par l'identifiant de sa MFM.

Chaque numérotation des identifiants est interne à chaque langue. Par exemple, un document peut être écrit en majorité en italique là où un autre est ecrit en romain : l'invariant considéré n'est donc pas un invariant de forme mais de structure.

De cette visualisation, nous pouvons pour un document donné dégager deux types d'alinéas : les alinéas nombreux et contigus (contenant le corps du texte) et les autres (révélateurs d'une

[8] OCR

[9] Extrait du document letton IP\05\1659 `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/1659&format=HTML&aged=1&language=LV&guiLanguage=en`

[10] Extrait du multidocument IP\05\817. `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/817&format=HTML&aged=1&language=IT&guiLanguage=en`

id.	extrait de l'alinéa avec sa MFM
a	Bruxelles 30 giugno 2005
b	**2006 - […] professionale**
c	***La Commissione […] lavoro.***
a	Lavorare in un […] di 10 anni.
a	Vladimír Spidla […] lavorare".
a	Su un bilancio […] professionale.
a	Nel 2006 […] commissari.

alignement au grain document		
et		1 1 2 3 1 1 1
it		a b c a a a a
nl		$\alpha\ \alpha\ \beta\ \gamma\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha$

FIG. 1 – Alinéas du document it et identifiants de MFM d'alinéa pour les documents et, it et nl

structure). En faisant cette distinction, nous regroupons automatiquement les suites contigues d'alinéas de même MFM. Les alinéas mis en valeur sont alors ceux qui brisent ces suites, que nous alignons aussi entre eux (alinéas "b" et "c" pour le document it - FIG. 2).

id.	extrait de l'alinéa avec sa MFM
a	Bruxelles 30 giugno 2005
b	**2006 - […] professionale**
c	***La Commissione […] lavoro.***
a	Lavorare in un […] di 10 anni.
a	Vladimír Spidla […] lavorare".
a	Su un bilancio […] professionale.
a	Nel 2006 […] commissari.

alignement 1										
et		1 1		2		3		1 1 1		
it		a		b		c		a a a a		
nl		$\alpha\ \alpha$		β		γ		$\alpha\ \alpha\ \alpha\ \alpha\ \alpha$		

FIG. 2 – Découpage du document italien pour l'alignement 1

Cependant, considérer seulement les suites d'alinéas comme précédemment (FIG. 2) ne révèle pas toutes les équivalences. Si nous alignons les titres entre eux et les sous-parties entre elles de façon indépendante, nous obtenons un découpage qui restreint simplement les espaces de recherche. Nous pouvons proposer manuellement les équivalences suivantes (FIG. 3) qui ne sont pas dévoilées par l'alignement automatique opéré précédemment.

id.	extrait de l'alinéa avec sa MFM
a	Bruxelles 30 giugno 2005
b	**2006 - […] professionale**
c	***La Commissione […] lavoro.***
a	Lavorare in un […] di 10 anni.
a	Vladimír Spidla […] lavorare".
a	Su un bilancio […] professionale.
a	Nel 2006 […] commissari.

alignement 2									
et		1 1		2		3 1 1 1			
it		a		b		c a a a a			
nl		$\alpha\ \alpha$		β		$\gamma\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha$			

FIG. 3 – Découpage du document italien pour l'alignement 2

2, b et β représentent des titres dans les trois documents. L'équivalence montrée en (FIG. 3) dévoile le fait que les parties qui suivent chaque titre peuvent être aussi alignées. Afin de multiplier les alignements possibles sans pour autant perdre l'alignement simple vu pour l'alignement 1

(FIG. 2), nous nous orientons vers une méthode d'extraction de structure des documents du multidocument.

4.3 Structuration du document par la MFM

Le but ici n'est pas de tenter la détection de la structure logique des documents, mais de réussir à calculer une représentation de structure. Nous ne cherchons pas à savoir si un élément détecté est une signature ou un résumé même si cette étape peut servir de base à d'éventuelles applications allant dans ce sens. Nous cherchons simplement à rendre les documents comparables entre eux.

Dans l'exemple précédant 1, a et α représentent la MFM du corps de texte, les autres alinéas conditionnent alors la structure. Pour différencier ces alinéas, nous utilisons l'ordre d'apparition de ceux-ci dans le document. Dans le but de représenter la structure sous forme d'arbre, nous considérons que descendre dans la hiérarchie arborescente du document s'effectue suivant l'ordre de lecture du document en fonction des alinéas ayant une MFM différente de celle représentant le corps de texte. Aussi, deux alinéas ne représentant pas le corps de texte et ayant la même MFM sont à la même profondeur de l'arbre, au même niveau de hiérarchie. Suivant ces remarques, nous établissons que :

– Chaque alinéa ayant une MFM représentant le corps de texte
  - possède un niveau de hiérarchie dépendant des alinéas précédemment rencontrés ;
  - n'augmente pas le niveau de hiérarchie des prochains alinéas.
– Chaque alinéa ayant une MFM différente de celle représentant le corps de texte
  - possède un niveau de hiérarchie dépendant des alinéas précédemment rencontrés. Ce niveau est alors fixé pour tous les alinéas ayant la même MFM ;
  - augmente le niveau de hiérarchie des prochains alinéas.

Le tableau (FIG. 4) montre l'application de la méthode sur le document italien.

modification de niveau de hierarchie		**+1**	**+1**				
niveau de hierarchie	1	1	2	3	3	3	3
alinéas	a	b	c	a	a	a	a

FIG. 4 – Calcul des niveaux de hierarchie pour le document italien

Afin d'illustrer la mise en correspondance des niveaux de hiérarchie de deux alinéas ayant deux MFM identiques, le tableau (FIG. 5) présente l'application de la méthode de hiérarchisation sur un document ayant une structure plus complexe [11].

modification de niveau de hierarchie		**+1**	**+1**				**+1**		**+1**			
niveau de hierarchie (! :fixe les niveaux *, * :niveau fixé)	1	1	2	3	3	3	3**!**	4	3*	4	4	4
alinéas	A	B	C	A	A	A	D	A	D	A	A	A

FIG. 5 – Calcul des niveaux de hierarchie pour le document français (IP\05\606)

[11] Document français du multidocument IP\05\606. `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/606&format=HTML&aged=1&language=FR&guiLanguage=en`

Il est possible alors de générer les arborescences à partir de la hiérarchisation qui a été calculée. La racine (niveau 0) représente pour chaque arbre le grain englobant les alinéas : le document.

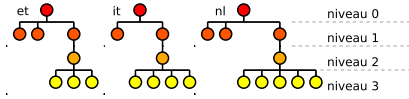


FIG. 6 – Structures générées - documents estonien, italien et néerlandais

Pour comparer les arbres entre eux, nous reprenons les critères utilisés pour l'alignement 2 (FIG. 3). Les suites connexes de feuilles sont regroupées pour permettre une comparaison entre les documents de structures équivalentes.

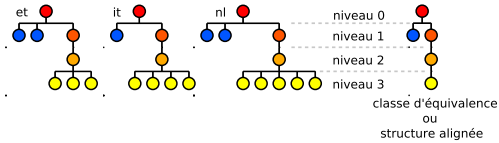


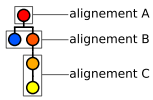
FIG. 7 – Alignement des structures

En regroupant ainsi les structures, nous obtenons une classe d'équivalence de structures (ou structure alignée). Chaque noeud et chaque feuille contiennent des suites d'alinéas alignées. Les documents sont aussi regroupés dans le même noeud (FIG. 8 alignement A). Ceci va avec l'idée que créer un multidocument équivaut à aligner des documents entre eux (ou à aligner au grain document). D'autres alignements peuvent être révélés dans les relations frère-frère (FIG. 8 alignement B) ou père-fils (FIG. 8 alignement C).

L'alignement 1 (FIG. 2) peut être retrouvés en considérant tous les alignements contenus dans chaque noeud et chaque feuille de la structure alignée. L'alignement 2 (FIG. 3) est récupérable via l'alignement C (FIG. 8) et les alignements issus des autres noeuds.

5 Perspectives

En traitant ainsi un multidocument, nous souhaitons obtenir qu'une seule classe d'équivalence afin d'avoir un alignement sur toutes les langues du multidocument. Sur le corpus de 63 multidocuments traités, 27 multidocuments possèdent une seule classe d'équivalence, 16 sont divisés en deux classes (dont une représente 1 ou 2 documents). Les autres multidocuments sont divisés en plusieurs classes (< 6) ; une classe d'équivalence regroupe une majorité absolue des documents et les autres contiennent 1 ou 2 documents. Ces multidocuments sont composés de documents ayant des structures différentes : certains possèdent des annexes ou des titres là où les documents de la classe majoritaire n'en ont pas.



alignement A	et	it	nl
alignement B	1 1 2	a b	$\alpha\ \alpha\ \beta$
alignement C	3 1 1 1	c a a a a	$\gamma\ \alpha\ \alpha\ \alpha\ \alpha$

FIG. 8 – Exemple d'alignements pouvant être détectés automatiquement

Dans cette continuité, nous nous attacherons à établir des politiques de comparaisons de classes d'équivalence pour aligner les documents ayant des structures différentes. Une approche consiste à utiliser les cognates pour aligner les suites d'alinéas les contenant. Il est possible d'utiliser les langues proches (où la présence de cognates est forte) se situant dans des classes différentes pour les comparer. Une autre voie consiste à utiliser les longueurs des suites d'alinéas (de façon analogue à (Gale & Church, 1993)) pour envisager des alignements $m : n$ (où on pourra se demander si $m, n \geq 2$) entre ces suites (par exemple si des annexes ou des titres sont ajoutés). Enfin, nous pouvons calculer les distances d'éditions des arbres de structure.

Cette approche apporte des indices supplémentaires pour affiner les méthodes déjà existantes et avoir une nouvelle vue sur l'alignement.

Références

DEBILI F. & SAMMOUDA E. (1992). Appariements de phrases de textes bilingues français-anglais et français-arabes. In *Actes de COLING-92*, p. 528–524, Nantes : COLING-92.

GALE W. & CHURCH K. (1993). Indentifying word correspondences in parallel text. In *Fourth DARPA Speech and Natural Language Workshop*, p. 152–157.

GIGUET E. (2005). Multi-grained alignment of parallel texts with endogenous ressources. Borovets, Bulgaria : Modern Approaches in Translation Technologies Workshop.

KAY M. (1988). *Text-translation alignment*. Rapport interne, Xerox Palo Alto Research Center.

MELAMED D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, p. 107–130.

SIMARD M. (2000). Three languages are better than two. In J. VERONIS, Ed., *Parrallel Text Processing*.

SIMARD M., FOSTER G. & ISABELLE P. (1992). Using cognates to align sentences in bilingual corpora. In *proceedings of TMI-92*.

VÉRONIS J. (2000). Alignement de corpus parallèle. In J.-M. PIERREL, Ed., *Ingienerie des langues*, p. 151–171.

ZIMINA M. (2004). Topographie bitextuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles. Université de la Sorbonne Nouvelle.