Multi-alignement vs bi-alignement : à plusieurs, c'est mieux !

Olivier Kraif
LIDILEM, Univ. Grenoble Alpes, BP 25, 38040 Grenoble cedex 9

Résumé. Dans cet article, nous proposons une méthode originale destinée à effectuer l'alignement d'un corpus multi-parallèle, i.e. comportant plus de deux langues, en prenant en compte toutes les langues simultanément (et non en composant une série de bi-alignements indépendants). Pour ce faire, nous nous appuyons sur les réseaux de correspondances lexicales constitués par les transfuges (chaînes identiques) et cognats (mots apparentés), et nous montrons comment divers tuilages des couples de langues permettent d'exploiter au mieux les ressemblances superficielles liées aux relations génétiques interlinguistiques. Nous évaluons notre méthode par rapport à une méthode de bi-alignement classique, et montrons en quoi le multi-alignement permet d'obtenir des résultats à la fois plus précis et plus robustes.

Abstract.

Multi-alignment vs bi-alignment: the more languages the better

In this paper, we propose an original method for performing the alignment of a multi-parallel corpus, ie a parallel corpus involving more than two languages, taking into account all the languages simultaneously (and not by merging a series of independent bi-alignments). To do this, we rely on the networks of lexical correspondences formed by identical chains and cognates (related words), and we show how various tiling of language pairs allow to exploit the surface similarities due to genetic relationships between languages. We evaluate our method compared to a conventional method of bi-alignment, and show how the multi-alignement achieves both more accurate and robust results.

Mots-clés : Alignement multilingue, corpus parallèles, cognats.
Keywords: Multilingual alignment, parallel corpora, cognates.

1 Introduction

Dans une étude pionnière dans le domaine de la désambiguïsation lexicale, Dagan *et al.* (1991) avaient intitulé leur article « *Two Languages Are More Informative Than One* ». Généralisant le même mot d'ordre (« *Five language are better than one* ») pour la désambiguïsation cross-lingue, Lefever *et al.* (2013) ont récemment illustré cette idée sur un corpus multi-parallèle, en bâtissant un système de classification automatique qui incorpore les contextes de traduction de 5 autres langues que la langue cible (l'anglais) : leur expérience montre que pour 4 langues sur 5, les résultats sont meilleurs quand on fait intervenir les contextes de traduction de plusieurs langues. Dans le domaine de l'alignement de corpus parallèle, cependant, il existe une piste qui a encore été assez peu explorée : celle du *multi-alignement*, à savoir l'alignement de plus de deux langues en même temps. Il paraît pourtant légitime de se poser la question suivante : le fait de prendre en compte plusieurs langues *simultanément* permet-il de mieux aligner ? Ce surcroît d'information permet-il de gagner en robustesse ? en précision ?

Notons que les corpus multi-parallèles ne sont pas rares : l'*Acquis communautaire*, qui constitue le socle législatif et réglementaire de l'Union Européenne en est un des exemples les plus représentatifs : il est actuellement distribué en version 3.0, sous le nom de JRC-Acquis Corpus, et concerne 22 langues européennes[1] – pour un total d'environ 636 millions de mots toutes langues confondues. A l'instar de l'UE, de nombreuses organisations internationales (ONU, GIEC, OMC, OMS, OIT, OCDE, etc.) sont pourvoyeuses de rapports et textes multi-parallèles touchant à des domaines très variés (économie, environnement, médecine, diplomatie, droit, technique, etc.). Dans le cadre du projet

[1] cf. http://ipsc.jrc.ec.europa.eu/index.php?id=198

Carmel (Chen et al., 2005), des œuvres littéraires (des récits de voyages) ont été réunies en 4 langues : anglais, espagnol, français, italien. On trouve par ailleurs sur le site de l'OPUS - Open parallel corpus (Tiedemann, 2012) une grande variété de corpus multi-parallèles déjà alignés. Certains de ces corpus sont massivement multilingues, comme le corpus *OpenSubtitles2013* qui compte 59 langues et intègre la plupart des paires de langues impliquées (1 211 paires de langues sur 1 711).

Il est frappant de constater que pour ces corpus massivement multi-parallèles, on utilise toujours des techniques d'alignement bilingue. Pour le JRC-Acquis, tous les alignements ont été effectués 2 à 2, en utilisant l'aligneur Vanilla (Danielsson & Riding, 1997), qui implémente la méthode de Gale & Church (1991), ainsi que l'aligneur Hunalign (Varga et al. 2005). De même les textes du corpus OPUS ont été alignés 2 à 2 grâce à une méthode similaire. Le principal inconvénient de l'alignement 2 à 2 d'un corpus de textes multi-parallèles réside dans le grand nombre de couples à considérer. Par exemple, pour les 22 langues du JRC Acquis, il faut considérer $22 \times 21/2 = 231$ couples différents. D'un point de vue général, pour *n* langues, le nombre de couples impliqués est quadratique : $n \times (n-1)/2$. Cette complexité peut être pénalisante à la fois du point de vue du temps de calcul et de l'espace de stockage des résultats. Quand on a 22 textes parallèles, pourquoi ne pas aligner les 22 langues en même temps, et représenter l'alignement résultant dans une seule structure de données, par exemple un seul fichier au format TMX contenant tous les groupes de phrases équivalents, plutôt que 231 fichiers différents ? La question du multi-alignement constitue donc également un enjeu en termes de complexité : le problème est-il abordable sur un plan calculatoire, et permet-il de faire mieux qu'une complexité quadratique ?

Cet article se propose de donner un début de réponse à ces questions, dans le cadre restreint de l'alignement phrastique. Dans la section 2, nous évoquerons les rares travaux qui ont cherché à sortir du carcan de l'alignement 2 à 2. Nous décrirons ensuite l'architecture d'un aligneur baptisé JAM (pour *Just A Multialigner*) qui s'appuie sur des réseaux de correspondances multilingues. Dans la quatrième section, nous chercherons à identifier ces réseaux de correspondance à travers un petit corpus en 11 langues tiré du corpus Europarl. Nous en déduirons une méthode de tuilage permettant de tirer le meilleur parti de ces correspondances. Enfin, dans la section 5, nous comparerons les performances de JAM avec celles de deux autres aligneurs bilingues ,Vanilla et Yasa (Lamraoui & Langlais, 2013).

2 Etat de l'art

En ce qui concerne l'alignement phrastique, force est de constater que depuis les travaux pionniers du début des années 1990, peu de choses ont bougé : les modèles superficiels faisant intervenir les longueurs de phrases (Gale & Church, 1991, Brown et al. 1991) ont largement fait leur preuve, complétés par des méthodes intégrant le repérage de ressemblances de surface, n-grammes, transfuges (i.e. les chaînes identiques, souvent des nombres et des entités nommées) ou cognats (Simard, Foster & Isabelle, 1992 ; Mc Enery & Oakes, 1995 ; Kraif, 2001). Nombre de ces algorithmes s'appuient sur le cercle vertueux énoncé par Kay et Röschensein dès 1988 (Kay & Röscheisen, 1993) : aligner au niveau des phrases pour aligner au niveau des mots, et aligner au niveau des mots pour aligner au niveau des phrases. Davis et al. (1995), dans le souci de tenir compte des ruptures de parallélisme fréquentes dans les traductions réelles, ont proposé de combiner ces différents types d'indices pour les intégrer dans un même cadre algorithmique. La campagne d'évaluation Arcade 2 (Chiao *et al.* 2006), a montré qu'il était possible d'appliquer ces méthodes, avec une certaine robustesse, à des couples de langues éloignées ne partageant pas le même alphabet (le français avec le russe, le chinois, le japonais ou l'arabe).

En complément des modèles de surface basés sur les longueurs, certaines méthodes s'appuient sur les alignements au niveau lexical, ce qui peut s'avérer nécessaire dans le cas de traductions « bruitées » s'écartant du parallélisme. Ce type de modèle requiert dans certains cas un lexique bilingue externe, comme dans Li et al. (2010), qui utilisent ce lexique à la fois pour extraire des points d'ancrage fiables destinés à réduire l'espace de recherche, et pour calculer l'alignement final. Moore (2002) obtient des résultats à la fois robustes et précis pour des alignements 1-1, en combinant les longueurs de phrases et l'alignement lexical, les alignements obtenus dans une première passe étant utilisés pour entraîner le modèle 1 d'IBM (Brown et al., 1993) en vue d'affiner l'alignement des phrases à partir de l'alignement des mots. Braune & Frazer (2010) améliorent cette dernière méthode en proposant de regrouper les phrases non alignées avec les alignements 1-1, afin de pouvoir constituer des alignements 1-n et augmenter le rappel. Plus récemment, Lamraoui & Langlais (2013), insistant sur le fait que l'alignement phrastique était un champ de recherche encore ouvert et méritant de nouvelles investigations, ont pourtant montré que leur aligneur Yasa, dont l'architecture très simple s'appuie sur un préalignement basé sur les cognats et un alignement phrastique « classique » incorporant longueurs de phrases et densité de cognat, était capable de dépasser les systèmes état de l'art plus complexes tels que BMA (Moore, 2002) et Hunalign (Varga et al., 2005), et ceci sur différents genres de textes.

Enfin, s'affranchissant de la contrainte de parallélisme, notons que de nombreux travaux ont porté sur l'alignement à travers des corpus comparables, au niveau phrastique (Munteanu et al., 2004) ou sub-phrastique (Hewavitharana & Vogel, 2011).

Le succès de ces méthodes explique peut-être le fait que le multi-alignement ait été si peu exploré. Dans le cas de trois langues, une étude asses complète a été effectuée par Simard (1999), avec un article dont le titre fait écho à l'article précédemment cité : « *Text-Translation Alignment: Three Languages Are Better Than Two* ». Il y présente une méthode d'alignement ternaire, nommée *Trial*, basée sur la réitération de la méthode bilingue. Étant donné 3 textes A, B, C, on aligne d'abord A avec B, puis C avec le bi-texte AB (le calcul du coût d'un appariement entre une phrase *c* et une bi-phrase (*a,b)* étant une simple combinaison linéaire des coûts d'appariement entre *c* et *a*, et *c* et *b)*. La méthode présentée par Simard n'a pas pour but d'économiser le temps de calcul, puisque tous les alignements bilingues AB, BC et AC sont calculés préalablement, afin de choisir la paire de langue optimale, qui sera ensuite réalignée avec la langue restante. En outre, les trois alignements bilingues permettent de dégager des points d'ancrage pour l'alignement ternaire, lorsqu'ils sont concordants (i.e. quand pour trois phrase a,b et c on a les appariements (a,b) (b,c) et (c,a)). Ce que montre Simard, ce n'est donc pas une réduction du calcul, mais une amélioration (certes modeste, avec 1% de F-mesure en plus) de la qualité de l'alignement final. Ainsi, quand un couple de langues est défaillant (p.ex. parce qu'on a trop peu de mots apparentés pour guider l'alignement des phrases), une troisième langue peut apporter une information complémentaire et suppléer à cette défaillance.

Dans la perspective de l'alignement sous-phrastique, la méthode d'alignement par échantillonnage proposée par Lardilleux (2010) présente l'originalité d'aligner simultanément plus de deux langues, toutes les unités des phrases alignées pouvant être fusionnées dans un même contexte d'occurrence sans langue définie – la méthode étant dite « alingue ». Des tirages aléatoires d'échantillon du corpus permettent de regrouper les unités qui partagent les mêmes distributions, ces regroupements pouvant aussi bien réunir des unités dans une même langue (i.e. des expressions polylexicales) que des équivalents traductionnels. Cette méthode est intéressante du point de vue de l'économie des traitements, l'alignement simultané du corpus multi-parallèle étant moins coûteux que l'alignement des langues deux à deux, mais elle n'est pas directement transférable à la problématique de l'alignement phrastique, et ne s'appuie pas, comme dans *Trial*, sur un renforcement des alignements par triangulation.

3 Cadre algorithmique pour un multi-aligneur

Les méthodes bilingues telles que celles de Gale & Church sont difficilement généralisables au cas de *n* langues, la complexité des algorithmes de programmation dynamique mis en œuvre étant exponentielle en $O(t^n)$, pour des textes de taille *t*. Le système *trial*, dans la mesure où il implique de pré-calculer tous les alignements 2 par 2, nous semble également assez lourd sur le plan algorithmique lorsqu'un grand nombre de langues est mis en jeu. Il n'a d'ailleurs jamais été étendu au-delà de trois langues, à notre connaissance. D'autres méthodes peu coûteuses sont envisageables, comme l'alignement par transitivité : si A est aligné avec B et B est aligné avec C, alors on peut calculer rapidement, par transitivité, un alignement entre A et C. Mais lorsque l'on prend la clôture transitive des alignements, on obtient en général des alignements plus grossiers, regroupant plusieurs phrases, ce qui aboutit à une baisse de la précision. Par ailleurs cette méthode ne tire pas parti du principe de triangulation : tout repose sur une seule langue pivot, et si l'alignement au niveau d'un couple est faible, cette faiblesse sera propagée vers la troisième langue par le jeu de la transitivité, au lieu d'être éventuellement compensée par la prise en considération d'un autre couple plus solide.

3.1 Cognats et multi-alignement

Pour des corpus multi-parallèles tels que ceux de l'Union Européenne, il apparaît que la parenté linguistique entre les différents groupes de langues impliqués (langues romanes, langues germaniques, langues slaves, langues baltes, langues finno-ougriennes, pour ne citer que les principaux groupes) doit pouvoir jouer un rôle prépondérant dans le multi-alignement. Afin d'explorer cette hypothèse, nous avons téléchargé la transcription de la session du 17 janvier 2000 du parlement européen, tiré du corpus Europarl3 (cf. http://opus.lingfil.uu.se/Europarl3.php), qui contient des versions alignées dans les langues suivantes : allemand, anglais, danois, espagnol, français, finnois, grec, italien, portugais, néerlandais, suédois (on utilisera désormais les codes ISO, par ordre alphabétique : DA, DE, EN, ES, EL, FI, FR, IT, NL, PT, SV). La partie française comporte environ 30 000 tokens (mots graphiques, ponctuations, nombres, etc.). Nous avons manuellement révisé les alignements fournis pour tous les couples impliquant le français afin d'avoir une référence fiable (la plupart des alignements fournis étaient de bonne qualité à part pour le couple FR-NL qui a nécessité un peu plus de révisions).

La première tâche a consisté à mesurer le degré de proximité graphique des formes alignées entre toutes les langues prises deux à deux, afin d'évaluer jusqu'à quel point la parenté génétique peut se traduire en un critère automatiquement exploitable (l'identification des candidats cognats). Pour chaque couple de phrases, nous avons compté les candidats cognats en retenant toutes les paires de mots d'au moins 7 caractères pour laquelle la sous-chaîne commune maximale (SCM, cf. Kraif, 2001) correspond à au moins 80 % des caractères de la chaîne la plus courte des deux chaînes comparées. Avec de tels critères, plutôt sélectifs, on trouve de très nombreux cognats avec un minimum de bruit. Par exemple, pour les langues DA, DE, EN, on trouve les paires suivantes : *periodiske↔Periodischen, Schroedter↔Schroedterin, Parlaments↔Parliament, Regionalpolitik↔Regional, regionaler↔regional, Europa-Parlamentets↔Europaparlamentets, Europæiske↔Europäischen, Kommission↔Commission*, etc.

Les résultats, cumulant le nombre de chaînes identiques (hormis les nombres et les noms commençant par une majuscule) et le nombre de cognats identifiés avec les critères précédents, sont présentés dans le tableau 1 :

	DA	**DE**	**EN**	**ES**	**FI**	**FR**	**IT**	**NL**	**PT**	**SV**	**Total**
DA		1 114	1 202	705	458	1 984	1 041	1 019	479	2 325	**14 327**
DE	1 114		863	448	397	735	747	722	376	925	**10 327**
EN	1 202	863		1 968	527	2 367	2 225	1 174	1 493	1 256	**17 075**
ES	705	448	1 968		222	1 829	2 234	638	3 750	764	**16 558**
FI	458	397	527	222		292	481	197	174	617	**7 365**
FR	1 984	735	2 367	1 829	292		2 120	936	1 350	851	**16 464**
IT	1 041	747	2 225	2 234	481	2 120		978	1 935	354	**16 115**
NL	1 019	722	1 174	638	197	936	978		489	893	**11 046**
PT	479	376	1 493	3 750	174	1 350	1 935	489		579	**14 625**
SV	2 325	925	1 256	764	617	851	354	893	579		**12 564**
Total	**14 327**	**10 327**	**17 075**	**16 558**	**7 365**	**16 464**	**16 115**	**11 046**	**14 625**	**12 564**	**136 466**

TABLEAU 1 : Nombre de transfuges et cognats identifiés dans les bi-phrases par couples de langues (grec exclu)

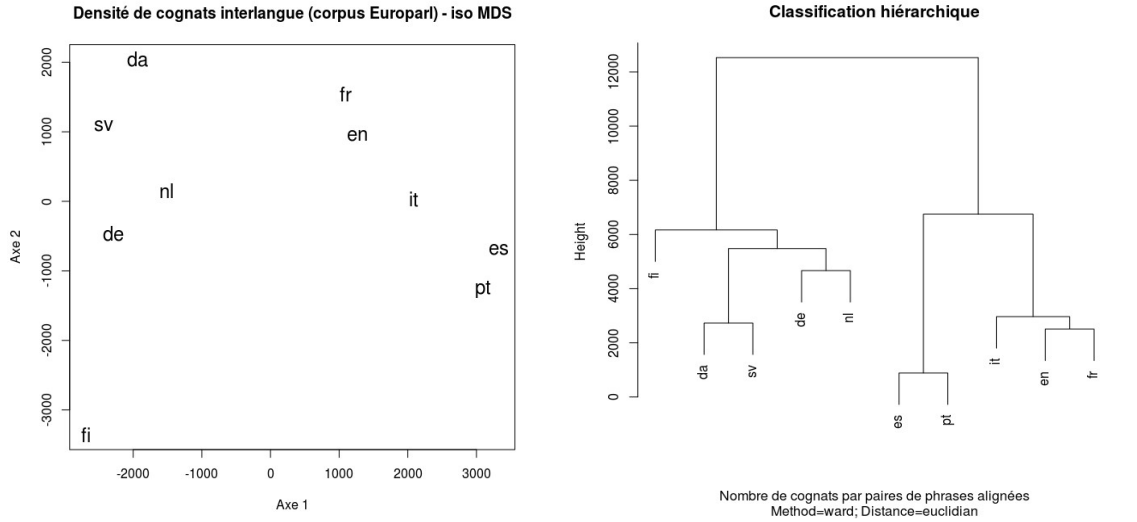

FIGURE 1 : Analyse multivariée du nombre de cognats par paires de phrases alignés (iso MDS et hclust)

Quand on considère les valeurs marginales, on constate que certaines langues cumulent beaucoup plus de cognats que d'autres : elles occupent en quelque sorte une position plus centrale au sein de ces différentes familles linguistiques, position qui leur confère en moyenne une plus grande ressemblance avec un plus grand nombre de langues – c'est notamment le cas du français et de l'anglais. Par ailleurs, ces densités de mots ressemblants font apparaître les affinités entre langue et groupe de langues. La figure ci-dessous montre les résultats de deux méthodes multivariées (classification hiérarchique ascendante obtenue avec la fonction *hclust()* de 'R', et échelonnement multidimensionnel avec iso MDS) en interprétant la matrice du tableau 1 comme une matrice de similarité (arbitrairement, on remplit la diagonale de la matrice par la valeur 4000 qui correspond approximativement à la borne supérieure des similarités observées). On note que les aspects génétiques sont étroitement corrélés aux similarités qui ressortent de ces données.

3.2 L'aligneur JAM

Pour utiliser ces cognats comme point d'appui pour le multi-alignement, on se dote d'une structure de données complexe que l'on nomme treillis de multi-alignement (cf. le tableau 2, qui contient les numéros de phrases appariées : chaque ligne correspond à un nœud du treillis, et les arcs sont les liens de successions propres à chaque langue). Cette table permet de mettre en correspondance les appariements simples entre phrases des différents textes (ici représentées par leur numéro), ces appariements pouvant être fragmentaires (lignes incomplètes) et l'ordre des lignes étant fixé par un chaînage propre à chaque langue : p.ex. pour NL le successeur de la ligne (_;2;2;2;2;2;2;2;2;2;2;2) est la ligne (_;7;_;6;_;7;7;_;_;9;7;_). Par construction, tout appariement d'une phrase avec une autre est transitif : à l'intérieur d'une même ligne, toutes les phrases sont alignées entre elle. Pour un groupe de langues donné (p.ex. S=EN-NL-SV) une étape de l'alignement consiste à : 1/ considérer tous les candidats cognats définis pour ce groupe de langues Co(S) 2/ identifier les intervalles à aligner, entre deux points consécutifs issus des étapes précédentes (pour les langues considérées p.ex. ici P_i=(2;2;2) et P_{i+1}=(7;10;9) 3/ à l'intérieur de chaque intervalle, sélectionner les cognats qui apparaissent avec la même fréquence pour chaque langue et ajouter les nouveaux points correspondant aux appariements de ces cognats 4/ Revenir à 2 sauf si stabilité 5/ Enfin, à l'issue de ces itérations, effectuer une étape de complétion : pour tous les couples de langues, lancer l'algorithme d'alignement de Gale & Church en ne calculant que les chemins qui passent par les points précédemment obtenus. Dans l'exécution de cette étape, les couples de langues sont ordonnés par complétude décroissante, c'est-à-dire que l'on traite en priorité les couples qui ont produit de nombreux points. Au fur et à mesure des alignements successifs, tous les appariements 1:1 qui correspondent à des points jugés *équilibrés*[2] et *cohérents* (i.e. sans croisement avec des points existants), sont ajoutés dans le treillis – et pourront être utilisés comme point d'ancrage pour les couples de langues non encore alignés.

Lors de chacune de ces étapes, pour tout ajout d'un nouveau point au treillis, trois cas de figure sont à considérer : soit il existe plusieurs lignes du treillis de multi-alignement liées à des phrases du nouveau point, et ces lignes doivent alors être fusionnées (si c'est possible) ; soit il existe une seule ligne liées à ces phrases, et celle-ci doit être complétée ; soit une nouvelle ligne doit être créée.

Ces étapes peuvent être exécutées pour n'importe quelle combinaison de groupes de langues, p.ex. Comb= {DA-DE-FI-NL-SV-EN, EN-ES-EL-FR-GR-IT-PT}. Notons qu'il est important que ces groupes possèdent des intersections non vides afin de relier toutes les langues au final : comme nous le verrons dans la section 3.3, un *tuilage* approprié est déterminant pour les résultats.

DA	DE	EL	GR	EN	ES	FI	FR	IT	NL	PT	SV
		1	1	1			1	1		1	
	2	2	2	2	2	2	2	2	2	2	2
	3	3	3		3	3	3	3		3	
	4				4	4	4	4		4	
	5			4	5	5	5	5		5	
	6			5	6	6	6	6		6	
	7			6		7	7		9	7	
	8			7	8	8	8	8	10	8	9
	10			8	9	10	9	9	11	9	
11	11			9		11	10	10	12		13
…	…	…		…	…	…	…	…	…	…	…

TABLEAU 2 : Extrait du treillis de multi-alignement obtenu pour le début de notre corpus, contenant les numéros de phrases appariées (dans ce multi-alignement, ainsi que dans tous les suivants, nous incluons le grec EL, ainsi que sa variante translittérée, qui sera désormais notée GR)

Calculer l'ensemble Co(S) des candidats cognats propres à une combinaison donnée peut se révéler assez coûteux : pour n textes avec un vocabulaire moyen de v formes, on aura environ $v^2*n*(n-1)/2$ couples de formes à comparer. Pour éviter cela, on commence par amorcer l'alignement avec les transfuges (les chaînes identiques) : calculer l'ensemble de transfuges Tr(S) est assez simple, en construisant initialement un hachage associant pour chaque forme

[2] Par point « équilibré », nous entendons que toutes les longueurs des phrases, prises deux à deux, sont comparables : on impose que le rapport de la plus courte sur la plus longue, en longueur relative, doit être compris entre 0.75 et 1.

la liste des différentes langues comportant cette forme. Si les combinaisons de Comb sont connues à l'avance, alors il est aisé d'alimenter Tr(S) dès la construction de ce hachage. À chaque ajout d'un nouveau point dans le treillis de multi-alignement, on compare alors les phrases alignées 2-à-2, et l'on en extrait les candidats cognats, cette fois en se basant sur la recherche de SCM (cf. section 3.1).

L'algorithme complet de JAM est décrit dans la figure ci -dessous :

```
T←{Premier,Dernier};  # on initialise le treillis avec le premier et le dernier point du multi-texte
Comb←{ensemble des combinaisons de langues considérées}
Pour chaque combinaison de langues S={L1-L2-…-LK} de l'ensemble Comb
    CO(S) ←{ensemble des transfuges identifiés pour S}
    # 1 - itérations
    Pour chaque cognat ou transfuge C de CO(S)
        Pour chaque couple de points (Pi,Pi+1) résultant de la suite ordonnée des points de T définis pour les langues de S
        Si C a n occurrences correspondant aux phrases occL,1, occL,2, … occL,n dans l'intervalle [Pi,Pi+1] pour chaque langue L de S
            Pour j=1…n
                Si le nouveau point (occL1,j, occL2,j, … ,occLK,j) est équilibré et cohérent
                    T←T U (occL1,j, occL2,j, … ,occLK,j)
                    Pour toutes les paires de mots (MLx,j MLy,j) des phrases occLx,j, occLy,j alignées du nouveau point
                        Si longueur(MLx,j) > 6 et  SCM(MLx,j MLy,j) >=0.8*min(longueur(MLx,j),longueur(MLy,j))
                            associer le même identifiant de cognat à MLx,j et à MLy,j
                        Fin Si
                    Fin Pour
                Fin Si
            Fin Pour
        Fin Si
    Fin Pour
    # 2 - complétion
    Ordonner tous les couples de langue (Li,Lj) par ordre décroissant en fonction du nombre de bi-points obtenus.
    Pour chaque couple de langue (Li,Lj) de cette liste ordonnée
        Appliquer l'algorithme de Gale & Church entre chaque point existant dans T
        Pour tout bi-point (Numi,Numj) de type 1 :1
            Si  (Numi,Numj) est équilibré et cohérent
                T←T U (Numi,Numj)
            Fin Si
        Fin Pour
    Fin Pour
```

FIGURE 2 : Algorithme itératif d'appariement des transfuges et cognats

Afin de garantir le maximum de précision, nous appliquons les deux critères suivants dans la sélection des points :

- *redondance* : dans un premier cycle d'itérations, on ne tient compte que des points contenant au moins *minMatchNumber* appariements de cognats (ou transfuges). Après stabilité, on réitère en décrémentant cette valeur. Dans les résultats qui suivent, on a testé *minMatchNumber*=2 puis 1.
- *parallélisme* : à chaque ajout d'un nouveau point *P*, on considère les deux points existants P_{inf} et P_{sup} qui encadrent ce point (pour les langues considérées dans ce point). On peut alors calculer la longueur des intervalles entre P_{inf} et *P* (notons Inf_L) et entre *P* et P_{sup} (notons Sup_L) pour chaque langue *L*. La vérification de parallélisme se fait alors pour chaque couple de langues L1, L2 : si déviation(Inf_{L1},Inf_{L2})[3] > *maxDiffInterval* et/ou déviation(Sup_{L1},Sup_{L2}) > *maxDiffInterval* alors les deux coordonnées du point correspondant à L1 et L2 sont rejetées. On peut appliquer cette contrainte sur les deux intervalles en même temps (« parallélisme fort »), ce qui n'autorise aucun décrochement dans l'alignement des points, ou bien d'un côté seulement (demi-parallélisme), ce qui peut permettre des sauts.

Nous avons effectué un premier test en utilisant un jeu de combinaisons de langues simple, prenant le français comme pivot (on notera FR-pivot) : $Comb_{FR\text{-}pivot}$={EN-FR, FR-IT, ES-FR, FR-PT, DA-FR, FR-NL, FR-SV, DE-FR, FR-FI, FR-GR, FR-EL}. Comme le montre la première colonne du tableau 3, on obtient une précision excellente mais un rappel assez faible. En examinant les points obtenus à ce stade, on obtient encore de nombreux « trous », comme l'illustre l'exemple du tableau 2 : l'alignement n'est pas complet car certaines langues se trouvent isolées, comme DA, EL, NL ou SV. Pour tenter d'éliminer ces trous, nous appliquons alors l'algorithme de complétion finale en lançant l'algorithme d'alignement de Gale & Church (1991) pour toutes les langues prises deux à deux. Notons que cet algorithme livre des appariements groupés de type 1:2, 2:1, 2:2 tandis que notre multi-alignement n'enregistre que des

[3] Pour deux intervalles Int_{L1} et Int_{L2}, on a : déviation(Int_{L1},Int_{L2})=2x | Int_{L1}-Int_{L2} | / (Int_{L1}+Int_{L2}). Le seuil *maxDiffInterval* prend des valeurs entre 0.5 et 0.25 suivant la taille de l'intervalle.

correspondances 1:1 sans fusion ni croisement. Pour rester dans le cadre imposé par notre structure de données, seuls les appariements 1:1 sont retenus (ce qui explique que le rappel ne puisse être optimal).

On obtient alors les résultats de la deuxième colonne du tableau 3. Notons que le coût de cet algorithme est modéré, étant donné l'étroitesse de l'espace de recherche : pour obtenir les résultats précédents, l'algorithme de Gale & Church a été lancé 6 662 fois sur des intervalles d'une longueur moyenne de 4 phrases environ, l'intervalle le plus grand ayant une longueur de 75 phrases[4].

	Sans complétion finale		Avec complétion finale	
Couple de langues	*P*	*R*	*P*	*R*
DA-FR	96,84	24,35	97,47	80,42
DE-FR	98,36	30,36	97,81	81,52
EL-FR	97,33	15,07	98,18	83,59
EN-FR	98,53	64,39	98,56	87,31
ES-FR	99,40	50,67	99,53	87,10
FI-FR	99,13	12,46	98,82	91,80
FR-GR	97,74	17,89	98,06	83,45
FR-IT	98,51	49,89	98,60	83,91
FR-NL	96,27	32,32	96,46	65,80
FR-PT	98,99	53,51	99,16	90,68
FR-SV	98,53	26,72	99,01	79,66

TABLEAU 3 : Résultats de JAM pour les combinaisons FR-pivot

3.3 Tuilage des couples de langues

Cherchons maintenant une combinaison de langues qui soit optimale, sans s'appuyer *a priori* sur le français. Une piste consiste à chercher le meilleur *tuilage* des alignements. Par tuilage on entend un ensemble de combinaisons de langues tel que chaque langue apparaît dans au moins une combinaison et que chaque combinaison possède au moins une langue en commun avec une autre combinaison. Parmi les tuilages possibles, on cherchera le tuilage qui met en jeu les groupes des langues plus fortement associées (d'après les données du tableau 1), afin qu'il puisse s'appuyer les uns sur les autres, de manière complémentaire, pour former un tout plus solide.

L'anglais étant la langue obtenant la plus importante valeur marginale (cf. tableau 1), on peut supposer qu'un tuilage basé sur l'anglais comme pivot donnera de bons résultats : on notera cette combinaison $Comb_{EN\text{-pivot}}$={EN-FR, EN-IT, EN-ES, EN-PT, EN-SV, DA-EN, EN-NL, DE-EN, EN-FI, EN-GR, EN-EL}. Notons que l'ordre d'application des couples peut jouer un rôle : on traitera d'abord les meilleurs couples, qui forment un réseau d'associations plus denses et plus sûres.

Une autre stratégie consiste à s'appuyer sur des triplets plutôt que sur des couples, la triangulation des langues permettant peut-être d'améliorer la précision des résultats. Nous avons cherché à identifier les meilleurs triplets potentiels en retenant, pour chaque langue, ses deux langues les plus fortement associées en termes de densité de cognats. Comme on peut supposer que la validité d'une combinaison n'est pas dépendante d'un texte en particulier, nous avons pris un autre corpus, tiré du JRC Acquis[5], pour calculer les densités de cognats à travers les phrases alignées. En ordonnant les triplets de façon décroissante en fonction du nombre global de cognats obtenus pour la première langue de chaque triplet, on obtient : $Comb_{TRI}$={ EN-FR-ES, FR-IT-EN, ES-PT-EN, PT-ES-IT, NL-EN-FR, SV-DA-NL, DE-DA-NL, FI-EN-SV, EL-GR-EN}. Enfin, à titre de *baseline*, nous avons testé également un tuilage « aléatoire » basé sur l'ordre alphabétique des codes de langue : $Comb_A$={DA-DE, DE-EL, EL-EN, EN-ES, ES-FR, FR-FI, FI-GR, GR-IT, IT-NL, NL-PL, PT-SV}.

Les résultats du tableau 4 montrent qu'au final les différences sont peu significatives entre ces différents tuilages : si on observe des différences marquées avant complétion au niveau du rappel (p.ex. R=16,75% pour la *baseline* contre 34,33 % pour *FR-Pivot*), la complétion permet un certain rattrapage qui uniformise les résultats à 2 point de F-mesure

[4] Si on utilisait directement l'algorithme de Gale & Church (1991) sur l'intégralité des textes pour les 66 couples en présence, chaque texte faisant environ 1 000 phrases, on aurait une complexité bien supérieure.

[5] Décision du 27 février 2002, ref. Celex=42002D0234, cf. http://optima.jrc.it/Acquis/index_2.2.html.

prêt. À ce stade, *FR-Pivot* semble être la meilleure combinaison – ce résultat étant toutefois à prendre avec précaution et peut être dû à des spécificités de notre corpus : il faudrait des études sur un corpus plus grand et plus varié pour tirer des conclusions sur ce plan.

	$Comb_{FR\text{-}pivot}$		$Comb_{EN\text{-}PIVOT}$		$Comb_{A}$		$Comb_{TRI}$	
Couple de langues	*P %*	*R %*	*P %*	*R %*	*P %*	*R %*	*P %*	*R %*
DA-FR	97,47	80,42	97,60	80,72	97,53	78,43	96,88	77,14
DE-FR	97,81	81,52	97,51	79,59	98,24	79,19	95,04	75,84
EL-FR	98,18	83,59	98,16	82,56	98,62	81,32	98,38	81,63
EN-FR	98,56	87,31	98,44	87,42	99,13	85,29	99,14	86,14
ES-FR	99,53	87,10	97,08	81,68	98,81	85,26	97,47	82,91
FI-FR	98,82	91,80	94,03	80,87	98,80	90,27	99,39	89,73
FR-GR	98,06	83,45	98,15	82,42	97,76	81,39	98,25	81,49
FR-IT	98,60	83,91	97,94	81,97	98,28	80,35	98,84	82,61
FR-NL	96,46	65,80	95,51	64,82	95,18	61,86	96,88	66,61
FR-PT	99,16	90,68	97,80	87,61	97,44	83,55	98,76	87,06
FR-SV	99,01	79,66	97,74	77,57	95,51	72,18	98,46	76,47
Moyenne	**98,33**	**83,21**	**97,27**	**80,66**	**97,76**	**79,92**	**97,95**	**80,69**

TABLEAU 4 : Résultats comparés pour différents tuilages (après complétion finale)

4 Comparaison avec des méthodes de bi-alignement

Reste à déterminer, à l'issue de ces différentes observations, si le recours au multi-alignement présente vraiment un intérêt par rapport à l'alignement binaire : c'est la question centrale à laquelle il nous faut maintenant tenter de donner une réponse. Pour ce faire, nous avons comparé JAM avec l'aligneur Vanilla (Danielsson & Riding, 1997), basé sur l'algorithme de Gale & Church (1991), et l'aligneur Yasa (Lamraoui & Langlais, 2013) qui a obtenu des résultats au niveau de l'état de l'art en conjuguant longueurs de phrases et densité de cognats. Pour Jam, nous avons utilisé $Comb_{FR\text{-}PIVOT}$. en appliquant la contrainte de « parallélisme fort » dans le filtrage des points. Dans le tableau 5, nous ajoutons les résultats de JAM+GC, pour les bi-alignements obtenus par l'application de l'algorithme de Gale & Church à l'issue de l'étape finale de complétion :

	Vanilla		JAM $Comb_{FR\text{-}PIVOT}$		JAM $Comb_{FR\text{-}PIVOT}$ + *GC*		YASA	
Couple de langues	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*
DA-FR	0,94	0,93	0,99	0,82	0,96	0,95	0,97	0,94
DE-FR	0,94	0,95	0,98	0,80	0,95	0,95	0,95	0,94
EL-FR	0,09	0,12	0,98	0,82	0,94	0,96	0,96	0,94
EN-FR	0,98	0,98	0,99	0,88	0,98	0,98	0,98	0,98
ES-FR	0,90	0,92	0,99	0,85	0,97	0,98	0,98	0,96
FI-FR	0,96	0,97	0,98	0,90	0,94	0,96	0,99	0,99
FR-GR	0,94	0,95	0,98	0,81	0,93	0,95	0,96	0,94
FR-IT	0,95	0,96	0,98	0,82	0,96	0,97	0,97	0,97
FR-NL	0,80	0,80	0,97	0,63	0,90	0,90	0,95	0,90
FR-PT	0,96	0,97	1,00	0,91	0,98	0,99	0,99	0,98
FR-SV	0,87	0,89	1,00	0,76	0,91	0,92	0,97	0,94
Moyenne (hors EL)	0,92	0,93	0,99	0,82	0,94	0,95	0,97	0,95

TABLEAU 5 : Résultats comparés de Vanilla, de JAM, de JAM+GC (avec l'application a posteriori de l'algorithme de Gale & Church entre les points d'ancrage), et de YASA

Les résultats de Vanilla pour le grec (EL) sont mauvais, mais nous pensons qu'il s'agit d'un problème de prise en compte du codage UTF-8 par Vanilla, et nous n'avons pas intégré ces résultats dans la moyenne (à priori, l'algorithme

de Gale & Church ne s'appuyant que sur les longueurs de phrases, EL et GR devraient être identiques). Nous avons donc calculé les moyennes sans cette ligne (en gris).

Vanilla obtient une F-mesure globale de 92,9 %, tandis que JAM obtient 95,07%, contre 96,3% pour YASA. Les résultats sont serrés, mais YASA s'en sort mieux globalement, et semble plus robuste, notamment pour les alignements avec le suédois et le néerlandais qui ont posé des problèmes à JAM. Pour NL, notamment, un point déviant issu de la phase 1 n'a pas été éliminé par les contraintes de parallélisme, et ce point a localement dégradé une partie de l'alignement. L'architecture multilingue permet donc d'améliorer l'alignement basé sur les longueurs (Gale & Church), mais sans toutefois surpasser un algorithme bilingue qui combine de façon optimale densité de cognats et longueurs. Un paramétrage plus fin de JAM permet d'améliorer les résultats, notamment en utilisant un tuilage basé sur les meilleures paires de langue d'abord : avec $Comb_{Max}$={ES-PT, EN-FR, DA-SV, ES-IT, EN-IT, SV-EN, NL-EN, DE-DA, FI-SV}, on obtient une F-mesure de 96,2%. Cependant, la recherche des paramètres optimaux ayant été effectuée sur ce même corpus, nous ne pouvons en tenir compte, d'autant que YASA n'a pas bénéficié d'un tel réglage. Par ailleurs, au plan du temps d'exécution, JAM (qui est implémenté en Perl) est encore assez lent : il prend 230 s. pour aligner tous les couples, tandis que YASA prend seulement 82 s. pour la même opération.

Il faut noter que les résultats de JAM (deuxième colonne) ne sont pas vraiment comparables avec ceux des autres colonnes, puisqu'il s'agit d'un multi-alignement ne comportant que des correspondances 1:1. Un multi-alignement construit à partir de regroupements de type 1:2, 2:1, 2:2 etc. serait *ipso facto* beaucoup moins précis. En effet, si on applique la propriété de transitivité sur des alignements binaires complets, on peut obtenir des regroupements très larges : il suffit qu'un alignement pour un couple de langues chevauche deux groupes de phrases différents pour d'autres couples pour que ceux-ci fusionnent, et ainsi de suite. Nous en avons fait l'essai en prenant la clôture transitive de nos alignements de référence avec le français, et nous obtenons des groupes élargis qui peuvent compter jusqu'à 13 phrases pour un seul groupe. Le tableau 6 en donne un échantillon pour le début du corpus :

DA	DE	EL	EN	ES	FI	FR	GR	IT	NL	PT	SV
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.4
8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1	8.1 8.2 8.3	8.1 8.2	8.1 8.2
8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.2	8.4 8.5	8.3 8.4	8.3 8.4
9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1
10.1 11.1 11.2	10.1 10.2 11.1 11.2	10.1 11.1 11.2	10.1 10.2 11.1 11.2	10.1 11.1 11.2 11.3	10.1 11.1	10.1 11.1 11.2	10.1 11.1 11.2	10.1 11.1	10.1 10.2 11.1 11.2	10.1 11.1	10.1 10.2 11.1
12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1
13.1 13.2	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1
13.3	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2
14.1 14.2	14.1 14.2 14.3	14.1 14.2	14.1 14.2	14.1 14.2	14.1 14.2	14.1 14.2	14.1 14.2	14.1	14.1	14.1 14.2	14.1 14.2 14.3
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..

TABLEAU 6 : Groupes obtenus par clôture transitive des 11 alignements de référence avec le français

4.1 Robustesse vis-à-vis des ruptures de parallélisme

Nous terminerons cette comparaison par une évaluation de la robustesse comparée de ces approches vis-à-vis des ruptures dans le parallélisme des traductions. Pour ce faire, nous avons créé artificiellement des « trous » dans la version française du corpus, en éliminant de façon aléatoire des blocs de phrases. Dans une première expérimentation, nous avons supprimé aléatoirement un bloc d'une seule phrase, en réitérant respectivement 10 fois, 50 fois et 100 fois. Nous avons alors lancé JAM et JAM+GC avec les combinaisons $Comb_{EN\text{-}PIVOT}$, $Comb_{FR\text{-}PIVOT}$ *et* $Comb_{TRI}$, en relâchant la contrainte de « parallélisme fort » (pour tenir compte des ruptures dans le chemin d'alignement). Comme nous nous y attendions, c'est la combinaison $Comb_{TRI}$ avec les triplets qui est la plus robuste, l'alignement à 3 langues étant moins sensible aux écarts locaux que l'alignement 2-à-2.

Nous avons également paramétré YASA afin d'éviter une dégradation brutale des résultats, en élargissant l'espace de recherche[6]. On obtient les résultats suivants :

Nb. blocs supprimés	Vanilla P	JAM P	JAM+GC P	YASA P	Vanilla R	JAM R	JAM+GC R	YASA R	Vanilla F	JAM F	JAM+GC F	YASA F
10	0,89	**0,97**	0,93	0,95	0,91	0,81	**0,95**	0,95	0,90	0,88	0,90	**0,95**
50	0,77	**0,97**	0,89	0,92	0,84	0,80	**0,95**	0,94	0,80	0,88	0,92	**0,93**
100	0,62	**0,93**	0,80	0,88	0,80	0,74	0,90	**0,92**	0,66	0,83	0,84	**0,90**

TABLEAU 7 : Résultats comparés de Vanilla, JAM ($Comb_{TRI}$ + GC) et YASA pour le corpus français dégradé (blocs de taille 1)

Ces résultats montrent que la précision de JAM se maintient à un niveau élevé et se dégrade légèrement pour 100 phrases supprimées – ceci étant dû à l'étape de complétion, basée sur les alignements de GC, qui gèrent mal ce cas de figure (les probabilités de suppression étant à priori très faibles). En effet, avant l'étape de complétion, JAM obtient une précision de 98% pour un rappel de 24 %. Globalement, pour le bi-alignement complet, si JAM+GC résiste mieux à la dégradation du corpus que Vanilla, YASA obtient à nouveau de meilleures performances, avec une F-mesure au-dessus de 90 %, malgré une diminution de 5 points.

Dans une deuxième expérimentation, nous avons étudié l'effet de la taille des blocs supprimés : cette fois nous ne supprimons qu'un seul bloc, comportant respectivement 10, 50, 100, 200 et 300 phrases.

Taille du bloc supprimé	Vanilla P	JAM+GC P	YASA P	Vanilla R	JAM+GC R	YASA R	Vanilla F	JAM+GC F	YASA F
10	0,82	0,93	**0,94**	0,85	**0,95**	0,93	0,83	**0,94**	0,94
50	0,45	**0,94**	0,87	0,50	**0,95**	0,82	0,47	**0,95**	0,84
100	0,54	**0,94**	0,94	0,62	0,93	**0,94**	0,57	**0,94**	0,94
200	0,25	**0,94**	0,92	0,33	**0,94**	0,93	0,28	**0,94**	0,93
300	0,06	**0,93**	0,84	0,08	**0,93**	0,87	0,07	**0,93**	0,85

TABLEAU 8 : Evolution des résultats en fonction de la taille des blocs supprimés

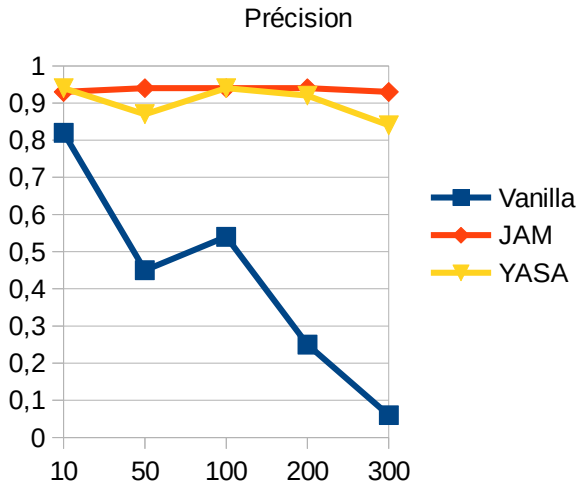

FIGURE 3 : *Evolution de la précision en fonction de la taille des blocs supprimés*

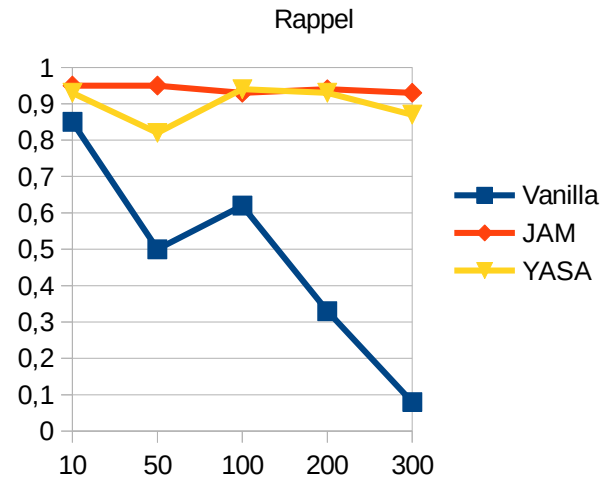

FIGURE 4: *Evolution du rappel en fonction de la taille des blocs supprimés*

Cette fois c'est JAM+GC qui se maintient à un niveau élevé, avec une F-mesure presque inchangée. YASA semble plus instable par rapport à ce type de rupture, étant donné la dégradation assez nette observée pour la suppression de 50

[6] Avec l'option -b 300, pour fixer à 300 phrases le rayon de recherche autour des points d'ancrage.

phrases (à l'instar de Vanilla). Par ailleurs, notons que le paramétrage spécifique de YASA a un coût algorithmique assez important : cette fois l'alignement de tous les couples est effectué en 600 s., alors que rien n'a changé du côté de JAM sur ce plan, qui prend un peu plus de 200 s.

5 Conclusion

Dans cette étude, nous avons proposé un cadre algorithmique pour la mise en œuvre d'une véritable méthode de multi-alignement, destinée à aligner simultanément toutes les versions d'un corpus multi-parallèle (i.e. comportant plus de deux langues). Notre algorithme repose sur une structure de données spécifique : un treillis de multi-alignement, où l'on définit pour chaque langue un chaînage spécifique des appariements partiels avec d'autres langues. L'algorithme itératif proposé permet de tirer parti dans un premier temps des transfuges, puis des cognats, pour apparier des phrases de combinaisons de langues déterminées. Une étape de complétion basée sur l'algorithme de Gale & Church (1991) est enfin appliquée, tant pour remplir le treillis de façon optimale que pour produire, si besoin est, des alignements complets deux à deux.

Dans une expérimentation sur un multi-texte de 11 langues, nous montrons qu'un tuilage des couples de langues suffit à obtenir des résultats satisfaisants, même s'il est possible d'optimiser ce tuilage en se basant sur les densités de cognats propres à chaque couple de langues. Nos résultats montrent par ailleurs que le multi-alignement produit de meilleurs résultats qu'une simple méthode bilingue telle que celle de Gale & Church (1991), et paraît adapté pour fournir des pré-alignements de qualité – avec seulement des appariements 1:1 – pour guider dans un second temps des méthodes d'alignement bilingues destinées à extraire des alignements complets.

Par la suite, nous avons montré que la prise en compte de plusieurs langues dans la même structure de données rend l'alignement, grâce au tuilage, très robuste face aux ruptures de parallélisme : que l'on supprime de nombreuses phrases disséminées çà et là, ou des blocs de grande taille, le pré-alignement stocké dans le treillis de multi-alignement reste d'une grande précision. Les comparaisons avec l'aligneur YASA montrent que celui-ci reste supérieur dans la tache d'alignement bilingue proprement dite, sauf dans les cas de ruptures importantes de parallélisme, où JAM se révèle plus robuste et plus rapide. Dans des cas de figure où de nombreuses langues sont disponibles, et où les ruptures de parallélisme sont fréquentes, le multi-alignement peut donc constituer une alternative crédible, en termes de robustesse et de fiabilité.

Notons enfin que le multi-alignement présente l'avantage de fournir en sortie une structure de données compacte renfermant un grand nombre de couples – avec une complexité en espace pour le stockage des résultats bien meilleure (en $O(n)$ pour n langues, contre $O(n^2)$ dans le cas bilingue) : un seul fichier au format TMX ou CesAlign peut renfermer les appariements de n langues - avec un rappel oscillant entre 80 % et 90 % vu qu'on n'y retient que les alignements 1:1.

Dans des travaux futurs, nous envisageons d'intégrer l'approche de YASA, qui combine densité de cognats et longueurs de phrases dans la phase de programmation dynamique, en sortie de JAM, pour améliorer l'extraction d'alignements bilingues autour des points d'ancrage. Par ailleurs, le corpus sur lequel nous avons mené cette étude est de taille modeste : il serait intéressant de l'élargir à un corpus plus volumineux et diversifié. Il serait utile, également, de proposer un cadre algorithmique pour étendre l'idée du multi-alignement à l'alignement sous-phrastique – en tirant réellement partie du principe de transitivité, contrairement aux méthodes multilingues proposées jusqu'ici (cf. Lardilleux, 2010). On pourra alors pleinement confirmer que pour l'alignement, tout comme pour la désambiguïsation multilingue, plusieurs langues valent mieux que deux.

Remerciements

Merci à Bettina Schrader, avec qui j'avais commencé à explorer cette piste de recherche il y a quelques années dans le cadre du projet Carmel.

Références

Brown P., Lai J., Mercer R. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, *ACL-91*, Morristown, NJ, 169-176.

Braune, Fabienne and Alexander Fraser. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of 23rd COLING*, 81–89.

CHEN, B., EL-BÈZE, M., HADDARA, M., KRAIF, O., MOREAU DE MONTCHEUIL, G. (2005). Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale, *Actes de TALN-RECITAL 2005*, 6-10 juin 2005, Dourdan, vol. 1, 415-420.

CHIAO Y.-C., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F., VÉRONIS J., ZAGHOUANI W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project, *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, Genova, May 2006.

DAGAN I., ITAI A., SHWALL U. (1991). Two Languages Are More Informative Than One. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, *ACL-91*, Morristown, NJ, 130-137.

DANIELSSON P., RIDINGS D. (1997). Practical presentation of a "vanilla" aligner. Presented at the *TELRI Workshop on Alignment and Exploitation of Texts*. Institute Jožef Stefan, Ljubljana.

GALE W., CHURCH K. (1991). "A Program for Aligning Sentences in Bilingual Corpora," *Association for Computational Linguistics,* 177-184.

HEWAVITHARANA S., VOGEL S. (2011). Extracting Parallel Phrases from Comparable Data, *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web.*

KAY M., RÖSCHEISEN M. (1993). Text-translation alignment. *Computational Linguistics* 19, 1 (March 1993), 121-142.

KRAIF O. (2001). Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation, *TAL* 42 :3, ATALA, Paris, 833-867.

LAMRAOUI F., LANGLAIS P. (2013). Yet Another Fast and Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment ?, *Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), 77–84.*

LARDILLEUX A. (2010). Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle. Thèse de doctorat, sous la direction d'Yves Lepage. Université de Caen, 2010.

LI, P., SUN M., XUE P. (2010). Fast-champollion: a fast and robust sentence alignment algorithm. In *Proceedings of 23rd COLING*, 710–718.

LEFEVER E., HOSTE V., DE COCK M. (2013) Five languages are better than one: an attempt to bypass the data acquisition bottleneck for WSD, CICLing 2013, Part I, LNCS 7816, Springer-Verlag: Berlin, 343-354.

MOORE, R.-C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *5th AMTA*, 135–144.

MUNTEANU D.S., FRASER A., MARCU D. (2004). Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora, P*roceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics* (HLT-NAACL).

SIMARD M. (1999). Text-Translation Alignment: Three Languages Are Better Than Two. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* 2-11.

SIMARD M., FOSTER G., ISABELLE P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, *TMI-92*, Montréal, CCRIT, 67-81.

TIEDEMANN, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'2012).

VARGA D., NÉMETH L., HALÁCSY P., KORNAI A., TRÓN V., NAGY V. (2005). Parallel corpora for medium density languages. *In Proceedings of 3rd RANLP,* 590–596.