PARADOCS : l'entremetteur de documents parallèles indépendant de la langue

Alexandre Patry[*,**] — **Philippe Langlais**[**]

[*] *KeaText*
845, Boulevard Décarie, bureau 202
Saint-Laurent, Canada H4L 3L7
alexandre.patry@keatext.com

[**] *Département d'Informatique et de Recherche Opérationnelle*
Université de Montréal
CP. 6128 Succ. Centre-Ville
Montréal, Canada H3C 3J7
{patryale,felipe}@iro.umontreal.ca

RÉSUMÉ. *Les corpus parallèles sont la pierre angulaire de plusieurs technologies de traduction automatique et des efforts conséquents sont régulièrement portés afin d'en réunir de nouveaux. L'expérience montre que la stratégie visant à réduire l'intervention manuelle dans cet exercice n'est jamais la même d'un corpus à l'autre. Ce constat nous a amené à développer* PARADOCS*, un entremetteur de documents parallèles qui utilise les entités numériques des documents afin de les apparier. Un classificateur est entraîné à décider des documents parallèles et un moteur de recherche d'information est utilisé afin de réduire l'espace de recherche des paires de documents parallèles. Nous montrons l'efficacité de* PARADOCS *sur de nombreuses tâches avec de nombreuses paires de langues.*

ABSTRACT. *Parallel corpora are the bread and butter of a number of machine translation technologies. Therefore, important efforts are regularly spent in acquiring new ones. This task often involves a rather cumbersome manual inspection and it is rather difficult to set up a strategy that fits all the needs. We thus developed* PARADOCS*, a system aiming at doing this automatically. Our solution exploits numerical entities in documents in order to pair them. A classifier trained to recognize parallel text coupled to an information retrieval engine controlling the search space of candidate pairs are the main components of our approach. We tested* PARADOCS *on a number of tasks involving numerous pairs of languages and report good results.*

MOTS-CLÉS : *corpus parallèles, recherche d'information, traduction automatique.*

KEYWORDS: *Parallel corpora, Information Retrieval, Machine Translation.*

1. Introduction

Un nombre grandissant de travaux dans le domaine de la traduction sont dédiés à l'exploitation de corpus comparables. L'idée que ces derniers sont disponibles en plus grande quantité que les corpus parallèles contribue à l'intérêt grandissant qu'on leur porte. Malgré cela, les corpus de *documents parallèles* (documents exprimant le même contenu dans le même ordre) continuent de jouer un rôle crucial dans les applications multilingues du traitement automatique des langues (Véronis, 2000). Aligné au niveau des phrases, une tâche pouvant être accomplie avec fiabilité (Langlais *et al.*, 1998), un corpus parallèle s'avère très utile aux concordanciers bilingues (Macklovitch *et al.*, 2000 ; Bourdaillet *et al.*, 2010). Il est la pierre angulaire de la plupart des systèmes commerciaux de mémoire de traduction. Aligné au niveau des mots, une tâche relativement bien maîtrisée (Brown *et al.*, 1993 ; Och et Ney, 2003), un corpus parallèle peut servir à plusieurs applications telles que la traduction automatique (Brown *et al.*, 1993) et l'extraction d'informations translinguistiques (Wessel *et al.*, 2003).

Le nombre de *corpus parallèles* (ensemble de documents parallèles) alignés au niveau des phrases – c'est à dire de *bitextes* – prêts à l'emploi augmente régulièrement. Historiquement, les débats parlementaires canadiens ont été parmi les premiers textes parallèles anglais/français mis à profit par la communauté de traduction statistique. LDC[1] distribue une version (payante) des hansards canadiens de 2,87 millions de paires de phrases (Ma, 1999). ISI[2] propose une version alignée des hansards du 36^{e} parlement canadien qui contient 1,3 million de paires de phrases et qui est libre de droit[3]. Au RALI, nous mettons à jour une version des hansards qui contient en date de rédaction de cet article 8,3 millions de paires de phrases. C'est ce bitexte qui est utilisé par la nouvelle version du concordancier bilingue TransSearch (Bourdaillet *et al.*, 2010).

Différentes versions parallèles des hansards sont disponibles pour plusieurs paires de langues ; notamment, les débats de la de Chambre des communes de Hong Kong (chinois/anglais). Il existe même des hansards inuktitut/anglais[4]. Des listes plus ou moins à jour de ressources parallèles sont également disponibles, dont la page d'Olivier Kraif[5] et celle du département de linguistique de l'université d'Uppsala[6].

À notre connaissance, le dernier effort important mené pour acquérir du Web un bitexte anglais/français de grande taille est celui du 10^9 *word parallel corpus* (Callison-Burch *et al.*, 2009). Les auteurs ont aspiré différents sites Web bilingues anglais/français (gouvernement canadien, Nations unies, organisations internationales, etc.) totalisant plus d'un teraoctet de données. De ce matériel, 2 millions de

1. Linguistic Data Consortium `http://www.ldc.upenn.edu/`
2. Information Science Institute, University of Southern California.
3. `http://www.isi.edu/natural-language/download/hansard/`
4. `http://www.inuktitutcomputing.ca/NunavutHansards`
5. `http://w3.u-grenoble3.fr/kraif/index.php?option=com_content&task=view&id=22&Itemid=38`
6. `http://xml.coverpages.org/etap-over.html`

paires de documents anglais/français ont été identifiées à l'aide d'heuristiques simples sur les URLs (comme par exemple remplacer *en* par *fr*). 177 millions de segments (de l'ordre de la phrase) ont ensuite été alignés à l'aide de la technique décrite dans (Moore, 2002). Différents filtres éliminaient ensuite des paires de phrases qui n'étaient pas parallèles. Les auteurs ont en effet relevé de nombreux cas où des documents n'ont pas été traduits alors que leur URL le suggère et mentionnent également que de nombreux documents sont dupliqués. Au final, 28 millions de paires de phrases constituent la première version de ce corpus[7].

Les ressources que nous venons de décrire sont bilingues. Il existe également des ressources multilingues. Les bitextes extraits des textes parlementaires européens – le corpus EUROPARL (Koehn, 2005) – disponibles dans 11 langues[8]. La cinquième version de ce corpus comporte environ 1,8 million de phrases par langue. Une autre ressource importante est constituée des textes des acquis communautaires, textes de l'Union européenne (UE) en majorité législatifs, organisés en un corpus appelé JRC-ACQUIS[9] (Steinberger *et al.*, 2006). Les textes de ce corpus sont disponibles dans 20 des langues officielles de l'UE plus le roumain. JRC-ACQUIS n'offre pas moins de 190 bitextes alignés au niveau des pargraphes, pour des paires parfois inhabituelles comme maltais/dannois. Chaque langue contient une moyenne de 9 millions de mots.

Maeda *et al.* (2008) décrivent les efforts récents mis en place à LDC[10] afin d'extraire des corpus parallèles arabe/anglais et chinois/anglais à partir de files de nouvelles disponibles sur le Web. Le cœur de leur approche réside dans l'utilisation du module d'appariement de documents du système BITS, un système d'extraction de documents parallèles depuis le Web développé à LDC il y a plus d'une décennie (Ma et Liberman, 1999). Ce module identifie dans deux ensembles de textes, l'un dans une langue source, l'autre dans une langue cible, les paires de documents parallèles en calculant un score de similarité pour chaque paire possible et en étiquetant comme parallèles les paires de documents dont le score de similarité dépasse un seuil donné. La similarité entre deux documents est, quant à elle, calculée à l'aide du ratio sur le nombre de mots traduits selon un lexique bilingue.

À toutes ces ressources extraites de sources institutionnelles, s'ajoutent des bitextes de nouvelles, comme le corpus anglais/norvégien ENPC[11] réalisé en 1994 à l'université d'Oslo ou encore le roman de George Orwell, *1984*, disponible en anglais et plusieurs langues de l'Europe de l'est (Erjavec, 2004). L'université d'Uppsala met également à la disposition de la communauté scientifique un bitexte suédois/turc constitué notamment des fictions *Beyaz Kale* d'Orhan Pamuk et *Sofies verden* de Jostein Gaardner (Megyesi *et al.*, 2006). De manière plus anecdotique, Resnik *et al.*

7. La version la plus récente, qui a subi d'autres étapes de nettoyage contient maintenant 22 millions de paires de phrases et est disponible à `http://www.statmt.org/wmt10`.
8. `http://www.statmt.org/europarl/`.
9. `http://langtech.jrc.it/JRC-Acquis.html`.
10. Linguistic Data Consortium (`http://www.ldc.upenn.edu`).
11. `http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/index.html`.

(1999) ont organisé les écrits de la Bible dans 13 langues, offrant ainsi 169 bitextes pour des paires de langues parfois atypiques, comme Cebuano/Indonésien. Le corpus ESPC[12] suédois/anglais regroupe également un certain nombre d'écrits littéraires, scientifiques et légaux (Altenberg et Aijmer, 2000).

Enfin, des textes techniques libres de droit sont également organisés en bitextes. La collection de bitextes OPUS[13] est mise à jour régulièrement. Elle est disponible gratuitement (Tiedemann, 2009). Ishisaka *et al.* (2009) ont également récemment compilé un bitexte anglais/japonais d'environ 500 000 paires de phrases à partir de documentations de logiciels libres disponibles sur le Web.

Dresser la liste exhaustive des ressources existantes n'est pas l'objectif de cet article. Nous pensons cependant avoir montré à l'aide de cet inventaire représentatif que de nombreux bitextes sont déjà disponibles et que de nouveaux sont créés régulièrement. Nous adhérons de plus à l'idée avancée par Maeda *et al.* (2008) que les corpus parallèles sont de plus en plus nombreux et nous pensons qu'il est donc pertinent de développer des méthodes pour en faciliter l'acquisition. Nous constatons également que les efforts mis en avant pour réunir des bitextes ne sont pas anodins et qu'ils se heurtent tous au problème canonique auquel nous nous intéressons dans cette étude : identifier les relations de traduction entre un ensemble de documents en langue source et un ensemble de documents en langue cible.

Dans la suite de cet article, nous dressons en section 2 une revue des approches proposées dans la littérature pour l'appariement de documents parallèles. Nous décrivons l'approche mise en avant dans PARADOCS en section 3. Nous décrivons en section 4 les expériences réalisées sur quatre corpus faisant intervenir différentes paires de langues. Différentes instances de PARADOCS sont comparées à un système de référence et démontrent la supériorité de notre approche. Nous concluons cette étude en section 5 et dressons des perspectives qu'ouvre ce travail.

2. Revue de littérature

Si les *documents comparables* (documents traitant d'un même sujet dans deux langues différentes) ont suscité et continuent de susciter de nombreux travaux[14], il n'en reste pas moins que les corpus parallèles sont pour le moment encore incontournables pour plusieurs applications. Dans cette section, nous décrivons différentes approches qui ont été proposées afin d'identifier dans deux ensembles de textes (l'un en langue source, l'autre en langue cible), les paires de documents parallèles.

L'idée la plus populaire consiste à utiliser les noms ou les URL des documents comme indice permettant de les apparier. Cette idée a été étudiée séparément par Chen et Nie (2000) et Resnik et Smith (2003) dans le cadre des systèmes PTMiner

12. `http://www.sol.lu.se/engelska/corpus/corpus/espc.html`.
13. `http://urd.let.rug.nl`
14. Lire par exemple (Fung et Cheung, 2004) pour une discussion de plusieurs d'entre eux.

et STRAND respectivement, deux systèmes d'extraction de documents parallèles depuis le Web. Les auteurs de PTMiner apparient les URL qui varient par un préfixe ou un suffixe identifiant la langue des documents (comme *fr*, *french*, etc.). Ainsi, les paires d'URL des exemples a) et b) de la figure 1 seront-elles appariées par ce système. Dans STRAND, différentes variantes de cette approche sont proposées. Une d'entre elles autorise plus d'une substitution pour que les URL comme ceux de l'exemple c) de la figure 1 soient appariées correctement.

a) `http://example.com/mondoc.fr`
   `http://example.com/mondoc.en`
b) `http://example.com/french-doc`
   `http://example.com/english-doc`
c) `http://example.com/French/mondoc.fr`
   `http://example.com/English/mondoc.en`

Figure 1. *Exemples de paires d'URLs appariées*

Resnik et Smith (2003) soulignent que dans de grandes collections de documents, comparer les paires d'URL du produit cartésien de tous les documents à disposition est un processus trop coûteux. Ils décrivent un processus simple qui permet de réduire les paires d'URL candidates. Une partition de l'ensemble des URL est effectuée en calculant une forme canonique pour chaque URL[15] ; seules les paires d'URL ayant la même forme canonique sont considérées pour l'appariement.

Deux URL appariées n'indiquent pas toujours que les documents associés sont parallèles (Chen et Nie, 2000 ; Resnik et Smith, 2003 ; Shi *et al.*, 2006). Les systèmes se fondant sur ce type d'information ont donc recourt à la longueur des documents afin d'éliminer les paires invraisemblables[16]. La longueur d'un document est un indice pour le moins très indirect de son contenu. Nous présentons donc d'autres indices qui ont été utilisés pour déterminer si deux documents sont parallèles ou non.

Certaines pages Web contiennent des liens HTML vers leur traduction (ex. : `<a href="mon-doc.en"><img src="flag.en" alt="english"/></a>`). Cette information a par exemple été utilisée par Deléger *et al.* (2006) dans leurs travaux sur l'extraction de terminologies bilingues et fait également partie des informations utilisées par le système STRAND. Resnik et Smith (2003) vont plus loin dans l'exploitation du marquage HTML et proposent d'utiliser une forme linéaire de la structure HTML de chaque document candidat à l'appariement (ex. : `[START:HTML] [START:TITLE] [chunk:13] [END:TITLE] ...`) puis de comparer entre elles ces représentations à l'aide d'un calcul de distance d'édition. Ils utilisent ensuite un classificateur pour identifier les paires parallèles à partir des caractéristiques de l'alignement menant à la plus petite distance d'édition. Une approche similaire est également décrite dans

15. Cette forme est calculée en retirant de l'URL des chaînes caractéristiques comme *8859-6, a, ar, ara, arab, arabic* ou *cp1256*.
16. Deux documents parallèles A et B vérifient $|A| \simeq \alpha|B|$ pour une constante α ajustée à la paire de langues traitée.

(Shi *et al.*, 2006). Les auteurs ajoutent aux traits utilisés par le classificateur, un score d'alignement au niveau des phrases qui résulte d'un alignement de la structure HTML (DOM) des documents.

Nie et Cai (2001) décrivent une cascade de filtres faisant usage de l'alignement au niveau des phrases des documents candidats à l'appariement. L'un de ces filtres consiste à favoriser les paires de documents dont les phrases alignées contiennent des entrées dans un lexique bilingue. L'usage d'un lexique bilingue était au cœur du système BITS, à la différence que Ma et Liberman (1999) proposent de compter les mots traduits si on suppose que les paires de documents sont parfaitement parallèles, ce qui les affranchit d'avoir à calculer l'alignement phrastique.

S'inspirant des travaux de Nadeau et Foster (2004) sur l'extraction de textes parallèles dans des files d'actualités, Patry et Langlais (2005) proposent d'entraîner de manière supervisée un classificateur (AdaBoost) à reconnaître les documents parallèles à l'aide d'éléments lexicaux, comme les entités numériques, les entités nommées ou les signes de ponctuation souvent invariants dans les traductions. Leur approche est comparée à une approche de base (*baseline*) qui apparie les documents qui partagent un maximum de mots selon un lexique bilingue à grande couverture. Seules les paires de mots peu fréquentes sont considérées. Les auteurs montrent le comportement parfait des deux systèmes sur des textes d'EUROPARL. Ils ont également aspiré les pages du Web de PAHO[17] (*Pan American Health Organization*) en anglais/espagnol et ont entraîné un système de traduction statistique à partir du bitexte obtenu par leur approche. Les traductions produites par ce système obtenaient de meilleurs scores (BLEU) que les traductions produites par le système entraîné depuis le bitexte obtenu à l'aide de l'approche baseline.

Enright et Kondrak (2007) décrivent une approche très élégante consistant à apparier les documents qui partagent le plus d'hapax d'au moins quatre lettres. Les auteurs évaluent leur approche sur les textes d'EUROPARL et rapportent des performances parfaites. Ils concluent que les textes des débats parlementaires européens ne sont pas propices à l'étude du problème d'appariement.

Dans la présente étude, nous revisitons le système PARADOCS décrit par Patry et Langlais (2005). Dans ce dernier, le produit cartésien des deux ensembles (source et cible) de textes est étudié, une solution qui n'est viable que pour de petites collections de textes. Nous proposons à la place d'utiliser un système de recherche d'information indépendant de la langue afin d'identifier les paires de documents parallèles. Nous étendons les traits utilisés lors de la classification d'une paire de documents. Enfin, nous comparons notre système à l'approche de Enright et Kondrak (2007) sur différentes tâches, certaines contrôlées, d'autres pas, en faisant varier différents paramètres pouvant influencer les performances.

17. `http://www.paho.org`

3. Approche

L'appariement d'un texte source à un quelconque des textes cibles présents dans la collection est réalisé dans la version actuelle de PARADOCS en trois étapes : la recherche d'un sous-ensemble de documents cibles candidats à l'appariement (section 3.1), l'identification des paires parallèles (section 3.2) et le filtrage éventuel des paires identifiées (section 3.3). Nous étudions chacune de ces trois étapes pour un ensemble de configurations que nous avons jugées pertinentes.

3.1. *Recherche de paires candidates*

Dans la version de PARADOCS décrite dans (Patry et Langlais, 2005), toutes les paires de documents possibles étaient considérées. Cette approche s'est avérée très lourde, même pour des collections de taille relativement modeste de l'ordre de quelques centaines de documents.

Comme l'ont proposés Munteanu *et al.* (2004) pour la détection de phrases parallèles dans un corpus comparable, nous avons restreint les documents cibles à considérer à l'aide d'un système de recherche d'information. Nous utilisons à cette fin Lucene[18], une librairie écrite en Java qui offre des outils performants d'indexation et de recherche dédiés aux textes[19].

Nous indexons tous les documents cibles et créons une requête pour chaque document source à apparier. Les paires candidates sont ensuite formées en jumelant chaque document source aux vingt premiers documents qui lui sont retournés par Lucene. Le défi repose donc dans la sélection des termes à utiliser pour l'indexation et la recherche.

Puisque nous revendiquons une approche indépendante de la langue et simple à mettre en œuvre, nous avons utilisé deux familles de termes : les nombres (`nombre`) et les hapax (`hapax`). Nous définissons les nombres comme une séquence d'au moins un chiffre. L'avantage de cette représentation est qu'elle est stable dans plusieurs paires de langues et qu'elle peut s'appliquer même à un texte où les frontières des mots n'ont pas encore été identifiées.

Tout comme Enright et Kondrak (2007), nous définissons les hapax comme les mots de plus de quatre lettres n'apparaissant qu'une seule fois dans le document. Ces mots sont de très bons indices de parallélisme lorsqu'ils s'écrivent de la même façon dans la langue source et la langue cible.

Utiliser un moteur de recherche pour restreindre les paires candidates est une arme à double tranchant. En effet, il peut arriver que le moteur ne retourne pas le bon document cible ou qu'il ne retourne rien du tout. Nous croyons que c'est cependant un

18. `http://lucene.apache.org`
19. *Lucene* utilise une représentation vectorielle des documents et des requêtes qu'il compare à l'aide d'un score de la famille de $tf{\cdot}idf$.

faible prix à payer pour identifier rapidement des paires de documents parallèles dans une grande quantité de documents.

3.2. *Identification des paires parallèles*

Lucene nous laisse avec un ensemble de paires de documents parmi lesquelles nous devons identifier celles qui sont parallèles. Nous commençons par représenter chaque document à l'aide de traits indépendants de la langue. La similarité entre les traits de chaque paire est ensuite mesurée et utilisée par un classificateur pour décider si cette paire est parallèle ou non.

Tout comme pour la recherche des paires candidates, nous souhaitons que l'identification soit la plus agnostique possible vis-à-vis des langues. C'est pourquoi nous avons représenté les documents par leurs nombres, par leurs hapax, par leurs nombres et leurs hapax ou par leurs nombres et leurs ponctuations[20].

Nous comparons ensuite les documents d'une paire en évaluant la similarité de leurs représentations. Pour ce faire, nous utilisons les trois métriques suivantes :

– la distance d'édition normalisée par la taille de la plus grande séquence :

$$\mathrm{ed}(\sigma(d_s), \sigma(d_t)) \,/\, \max(|\sigma(d_s)|, |\sigma(d_t)|)$$

où $\sigma(d)$ extrait les nombres, les hapax ou les ponctuations d'un document d. Nous normalisons la distance d'édition pour qu'elles soient toujours entre zéro et un ;

– la distance d'édition normalisée nous informe sur la fraction des traits comparés, mais ne dit rien sur le nombre total de ces traits. C'est pourquoi nous ajoutons le nombre total de traits dans les deux documents ($|\sigma(d_s)| + |\sigma(d_t)|$) à la distance d'édition normalisée ;

– intuitivement, il faut privilégier les paires de documents ayant la plus petite distance d'édition. Nous ajoutons donc un indicateur binaire marquant ces paires :

$$\delta(d_s, d_t) = \begin{cases} 1 & \text{si } \mathrm{ed}\left(\sigma(d_s), \sigma(d_t)\right) \leq \mathrm{ed}\left(\sigma(d_s), \sigma(d_{t'})\right) \ \forall\, d_{t'} \\ 0 & \text{sinon} \end{cases}$$

Ces mesures sont calculées pour chacune des représentations. Ainsi, l'expérience utilisant les nombres et les ponctuations assignera six mesures à chaque paire : trois pour les nombres et trois pour les ponctuations.

Contrairement à Patry et Langlais (2005) nous n'utilisons pas les traits obtenus par alignement au niveau des phrases. Les auteurs ont montré que ces traits n'étaient pas très discriminants. De plus, le calcul de l'alignement de phrases pour chaque paire candidate est une opération coûteuse (*a priori* quadratique par rapport au nombre de phrases dans les documents).

20. Nous nous sommes restreints aux ponctuations suivantes qui sont fréquemment traduites telles quelles parmi les paires de langues auxquelles nous nous sommes intéressées : `.!?():`

Nous avons également évalué plusieurs classificateurs : un arbre de décision (j48), un modèle bayésien naïf (bayes), un modèle AdaBoost (adaboost) ainsi qu'un modèle de régression logistique (régression). Tous ces modèles ont été entraînés à l'aide du logiciel libre Weka (Hall *et al.*, 2009)[21].

3.3. *Post-traitements*

Les classificateurs supposent que les paires sont indépendantes les unes des autres. Cette hypothèse est manifestement fausse, car chaque document ne devrait pas apparaître dans plus d'une paire.

Plusieurs solutions peuvent être appliquées pour filtrer les paires de documents appariées. L'une d'elles consiste à ne rien faire, c'est la solution appelée `dup` dans la suite. L'autre solution que nous avons testée élimine les paires partageant des documents ; nous l'appelons `nodup`. Une autre solution simple que nous n'avons pas testée ne consisterait qu'à garder la paire ayant le meilleur score.

4. Expériences

Nous décrivons dans cette section les quatre expériences que nous avons réalisées afin d'évaluer PARADOCS selon différents axes. La première série d'expériences menée sur le corpus EUROPARL (section 4.1) nous a servi à calibrer le système et à en tester le comportement sur de nombreuses paires de langues. Nous avons ensuite étudié deux paires de langues ne partageant pas le même alphabet : arabe/anglais (section 4.2) et inuktitut/anglais (section 4.3). Ces trois expériences sont *contrôlées*, dans la mesure où nous connaissons déjà les paires de documents parallèles, ce qui nous permet de mesurer des taux de précision et de rappel. En section 4.4, nous décrivons une expérience réalisée sur des pages en anglais et en français de WIKIPEDIA pour lesquelles nous ne connaissons pas les pages parallèles. Nous évaluons alors manuellement un échantillon de la sortie de PARADOCS.

4.1. EUROPARL

4.1.1. *Corpus*

Nous avons téléchargé la version cinq du corpus EUROPARL[22]. De l'ordre de 6 000 documents sont disponibles pour 11 langues (dont l'anglais), ce qui constitue 10 bitextes dont l'une des langues est l'anglais. Le nombre moyen de phrases des documents anglais est d'environ 273. Certains documents présentent des problèmes (pas

21. `http://www.cs.waikato.ac.nz/ml/weka/`.
22. `http://www.statmt.org/europarl/`

de phrases, arrêt du fichier en plein milieu, etc.) ; nous n'avons réalisé aucun traitement particulier pour y remédier.

De manière à mesurer l'influence de la taille des documents manipulés, nous avons artificiellement préparé différentes versions des ensembles source et cible de documents. Nous avons pour cela retenu de chaque document un nombre variable n de phrases parallèles parmi 10, 20, 30, 50, 70, 100 et 1 000. Nous avons ensuite lancé les différentes variantes de PARADOCS sur chacun de ces jeux de test.

4.1.2. *Protocole*

Sept paramètres contrôlent les différentes variantes de PARADOCS testées : la longueur des documents (7 valeurs différentes), la paire de langues (10 valeurs), les termes utilisés pour la recherche (2 valeurs), les termes utilisés pour la représentation des documents (4 valeurs), le score utilisé pour comparer les documents (1 valeur), le classificateur (4 valeurs) et la présence ou l'absence de post-traitement (2 valeurs). Nous avons donc réalisé un total de 4 480 expériences de recherche de paires de documents parmi 6 000 documents sources et 6 000 documents cibles. Il ne nous est donc pas possible de rapporter l'ensemble des résultats ; aussi nous contentons-nous de donner les tendances observées.

L'évaluation des variantes se fait en comparaison à la référence à l'aide des mesures de *rappel* (ratio du nombre de paires de documents identifiées au nombre total de paires dans la référence), de *précision* (pourcentage des paires identifiées qui sont dans la référence) et de *f-mesure* (moyenne harmonique de la précision et du rappel). Puisque PARADOCS requiert l'entraînement d'un classificateur, nous avons procédé par validation croisée et les résultats que nous rapportons ici sont les moyennes sur cinq strates ; quatre cinquièmes du corpus étant à chaque fois utilisés pour entraîner le classificateur, le cinquième restant servant à l'évaluer.

4.1.3. *Recherche des paires candidates*

Comme nous l'avons déjà mentionné, PARADOCS utilise un moteur de recherche pour restreindre les paires de documents qui seront identifiées. Cela réduit considérablement le temps nécessaire à l'évaluation des documents parallèles par rapport à Patry et Langlais (2005). Cependant, deux cas de figure peuvent conduire à un échec de *Lucene* : celui où il ne retourne aucun document et celui où il ne retourne que des mauvais documents.

En figure 2, nous montrons le pourcentage de documents sources pour lesquels *Lucene* ne retourne aucun document pour la paire de langues néerlandais/anglais (des courbes très similaires sont observées pour les autres paires de langues). On observe que la taille des documents influe directement sur ce pourcentage. Environ 11 % des documents de 100 phrases et plus n'obtiennent aucune réponse de *Lucene*. Pour les documents plus petits, jusqu'à 33 % des documents sont sans réponse. De l'ordre de 6 % des documents sources pour lesquels *Lucene* retourne des documents cibles ne contiennent pas le document cible apparié.

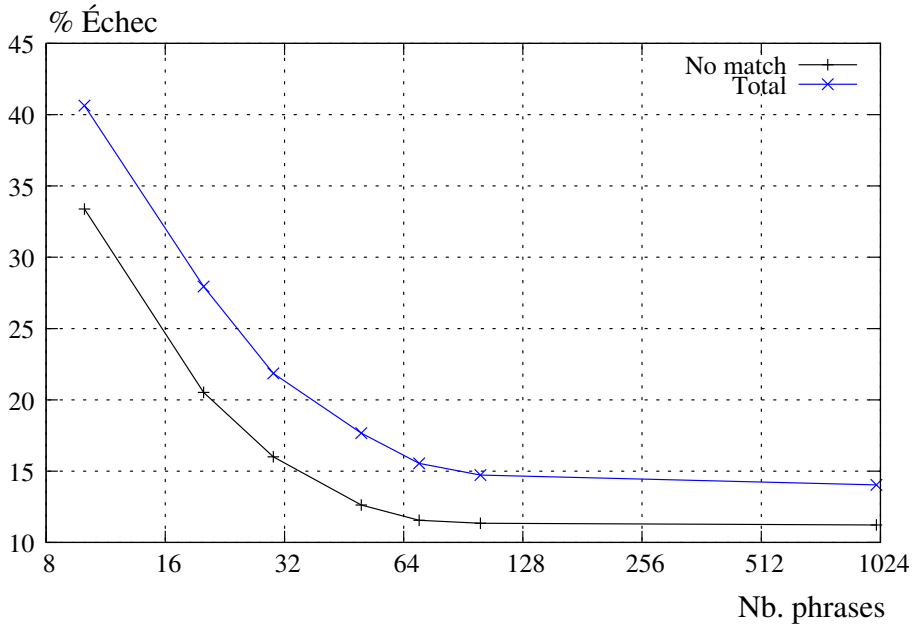

Figure 2. *Pourcentage de documents sources pour lesquels Lucene n'a retourné aucun document (no match) et aucun document correct (total) pour la paire de langues néerlandais/anglais, en fonction de la longueur des documents comptée en phrases*

Ce prix n'est cependant pas cher payé. Nous imaginons facilement un scénario où PARADOCS identifierait rapidement la majorité des documents parallèles et qu'un deuxième système plus lourd et plus complexe s'occuperait des autres. Ce deuxième système pourrait même utiliser des modèles statistiques entraînés sur la sortie de PARADOCS. Dans la suite, les évaluations de la tâche d'identification ne portent que sur les documents pour lesquels *Lucene* retourne le bon document dans sa liste de candidats.

4.1.4. *Influence de la langue et de la taille des documents*

Dès lors que les documents sont assez grands (100 phrases ou plus), il n'existe pas de différence majeure de performance en fonction de la paire de langues étudiée. Ceci peut-être observé sur la figure 3 où pour les documents de 100 phrases, la plus petite f-mesure est de $0,934$ pour la paire néerlandais/anglais et la plus grande de $0,949$ pour la paire français/anglais. Des différences plus marquées sont en revanche observées pour les petits documents, où là encore, la paire néerlandais/anglais est la moins facile à apparier. Nous notons que la paire français/anglais est de loin la plus facile à traiter et que la f-mesure pour cette paire de langues est déjà très élevée pour

les documents d'au plus 10 phrases. Mise à part cette paire de langues[23], on observe donc que PARADOCS n'est pas dépendant de la langue, ce qui n'est pas surprenant compte tenu des indices utilisés par ce système.

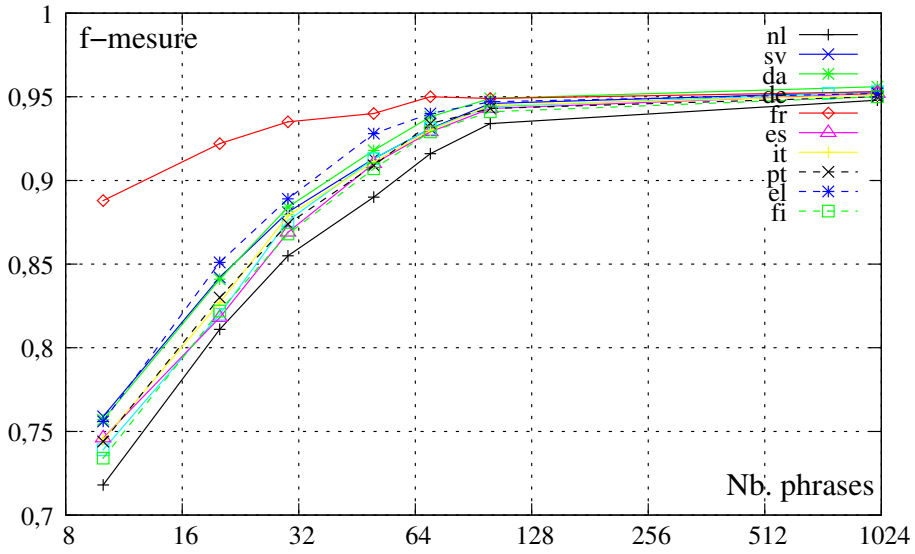

Figure 3. *f-mesure du module d'identification des paires parallèles de la meilleure variante de* PARADOCS *en fonction de la taille maximale des documents comptée en phrases pour les 10 bitextes d'*EUROPARL

On observe également que sur EUROPARL, la meilleure performance en terme de f-measure est de 0,95. Les configurations gagnantes de PARADOCS possèdent toutes une précision proche de 1, attestant du bon comportement des classificateurs gagnants.

4.1.5. *Variantes de* PARADOCS

Afin de déterminer l'importance des différents paramètres influençant PARADOCS, nous avons retenu les variantes obtenant la meilleure f-mesure pour les 70 expériences lancées en faisant varier la paire de langues (parmi 10 paires différentes) et la taille des documents (parmi 7 tailles différentes). Le décompte des configurations gagnantes est rapporté dans le tableau 1. Typiquement, environ trois configurations par expérience sont *ex aequo* et obtiennent la f-mesure la plus grande, ce qui explique que le total du nombre de configurations gagnantes (première colonne) soit supérieur à 70.

On observe que la recherche sur les entités numériques (`nombre`) amène plus souvent aux meilleurs résultats que celle sur les hapax (`hapax`). Nous revenons sur

23. Nous n'avons pas encore identifié de raison expliquant la plus grande facilité à la traiter.

ce point dans la section suivante. Le post-traitement semble également souhaitable (`nodup`).

Nb	Recherche	Identification	Classificateur	Filtre
1	`nombre`	`nombre+hapax`	j48	`dup`
1	`nombre`	`nombre`	j48	`dup`
1	`nombre`	`nombre+ponct`	j48	`dup`
7	`hapax`	`hapax`	j48	`nodup`
11	`nombre`	`nombre+hapax`	bayes	`nodup`
11	`nombre`	`nombre`	bayes	`nodup`
11	`nombre`	`nombre+ponct`	bayes	`nodup`
24	`nombre`	`nombre+hapax`	j48	`nodup`
24	`nombre`	`nombre`	j48	`nodup`
24	`nombre`	`nombre+ponct`	j48	`nodup`
31	`nombre`	`nombre+hapax`	régression	`nodup`
31	`nombre`	`nombre`	régression	`nodup`
31	`nombre`	`nombre+ponct`	régression	`nodup`

Tableau 1. *Caractéristiques des configurations gagnantes des 70 expériences réalisées en faisant varier la paire de langues et la taille des documents. La première colonne comptabilise le nombre de fois où la configuration concernée a obtenu la meilleure f-mesure (deux configurations peuvent obtenir le même score)*

En ce qui concerne l'identification des paires parallèles, il apparaît clairement que toutes les configurations gagnantes font usage des entités numériques. L'ajout des ponctuations (`nombre+ponct`) ou des hapax (`nombre+hapax`) ne semble pas améliorer les résultats. Aucune configuration faisant un usage exclusif des hapax n'a obtenu de meilleurs résultats qu'une configuration faisant usage des nombres. Finalement, on observe que la régression logistique est régulièrement parmi les meilleurs classificateurs. Cependant, comme les arbres de décision étaient toujours dans le peloton de tête, nous les avons préférés à la régression logistique où ce n'était pas toujours le cas.

4.1.6. *Entités numériques* versus *hapax*

Les observations précédentes indiquent que l'utilisation des entités numériques est préférable à celle des hapax, lorsque embarquée dans un classificateur. Afin de vérifier si cette observation ne serait pas simplement liée à un artefact de l'étape de classification de PARADOCS, nous avons implémenté la version de Enright et Kondrak

(2007) telle que décrite dans leur article. Le code shell[24] de cette variante est présenté à la figure 4[25].

```
foreach in ($1 $2)
  cat $in | tr '[:space:]' '\n' | sort | uniq -u
          | grep '\w\w\w\w' >! $in.hpx
end
sort $1.hpx $2.hpx | uniq -d | wc -l
```

Figure 4. *Programme c-shell qui affiche le score de l'appariement des deux fichiers dont les noms sont passés à la ligne de commande*

La figure 5 montre les gains absolus en f-mesure de la meilleure variante de PARADOCS par rapport à la f-mesure de l'approche de Enright et Kondrak (2007). Nous observons la supériorité de PARADOCS pour toutes les paires de langues et pour toutes les tailles de documents. Les gains les plus modestes sont observés pour le français, la langue pour laquelle nos résultats sont les plus élevés. Pour les documents les plus grands, le gain moyen en f-mesure, toutes paires de langues confondues, est de 13,6. Nous devons souligner que contrairement à Enright et Kondrak (2007), notre approche requiert l'entraînement d'un classificateur, ce qui la rend potentiellement moins intéressante dans certaines applications. C'est pourquoi nous testons en section 4.4 l'usage d'un classificateur entraîné sur une autre tâche que celle à laquelle PARADOCS est appliqué.

4.2. MOTASEM

Les expériences précédentes touchaient à des textes législatifs. Afin d'étudier le comportement de PARADOCS sur d'autres types de textes et impliquant une langue non européenne, nous avons utilisé un corpus mis au point par Alrahabi et Desclés (2007). Ce corpus est constitué de 115 paires de textes arabe/français du *Monde diplomatique* des années 2002 et 2003. Les paires parallèles ont été identifiées manuellement, un travail qui au dire des auteurs s'est avéré long et fastidieux. Les documents en français font en moyenne 21 ko, les documents en arabe 25 ko.

Les performances de PARADOCS faisant usage des arbres de décision sont présentées dans le tableau 2. On observe une f-mesure proche de celles observées sur la tâche EUROPARL. Les variantes faisant usage des hapax sont significativement moins

24. Les auteurs remercient Ken Church de les avoir inspirés avec *Unix for poets* (`http://people.sslmit.unibo.it/~baroni/compling04/UnixforPoets.pdf`).

25. En pratique cette approche est trop lente et nous avons plutôt utilisé les structures de hachage de la STL en C++. Des prétraitements (élimination de diacritiques, prise en compte de la ponctuation lors du découpage en mots) ont également été mis en place, conformément à l'étude de Enright et Kondrak (2007).

performantes (baisse de f-mesure d'environ 0,2). Des performances légèrement supérieures (0,949) sont obtenues en utilisant un classificateur bayésien naïf. Les variantes de notre système qui font un usage exclusif des hapax obtiennent au mieux un f-mesure de 0,868. Une grande partie de ces hapax correspond d'ailleurs à des données numériques. Notez que pour ce corpus, la recherche des documents de *Lucene* n'échoue que pour dix documents.

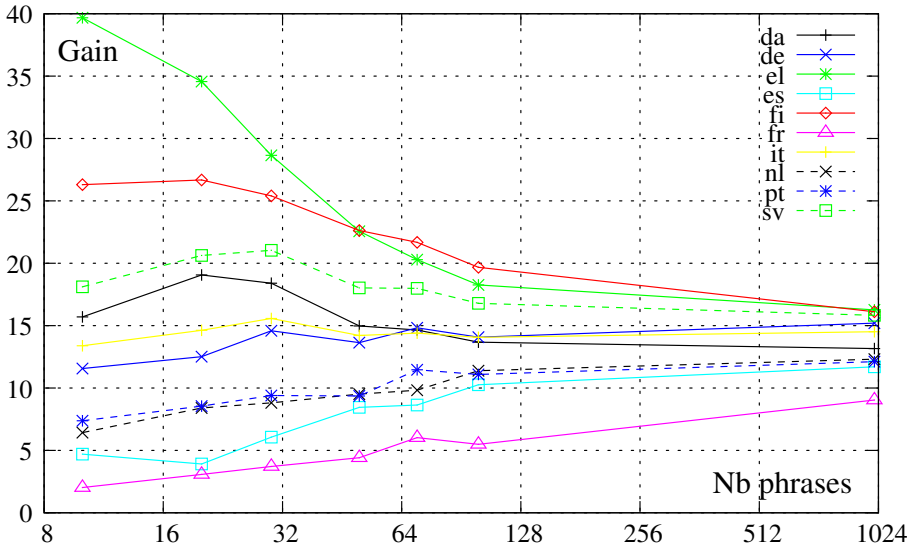

Figure 5. *Gains absolus en f-mesure de la meilleure variante de* PARADOCS *sur l'approche décrite par Enright et Kondrak (2007) en fonction de la taille des documents et les paires de langues étudiées*

Recherche	Identification	Filtre	Préc.	Rappel	F-mesure
`hapax`	`hapax`	`dup`	0,802	0,689	0,741
`hapax`	`hapax`	`nodup`	0,944	0,642	0,764
`nombre`	`nombre+hapax`	`nodup`	0,924	0,933	0,928
`nombre`	`nombre`	`nodup`	0,924	0,933	0,928
`nombre`	`nombre+ponct`	`nodup`	0,924	0,933	0,928
`nombre`	`nombre+hapax`	`dup`	0,901	0,962	0,93
`nombre`	`nombre`	`dup`	0,901	0,962	0,93
`nombre`	`nombre+ponct`	`dup`	0,901	0,962	0,93

Tableau 2. *Performances de variantes de* PARADOCS *faisant usage d'un arbre de décision et prenant en compte l'ordre des entités sur la tâche* MOTASEM

a)	The Legislative Assembly convened at 3-30 pm.	a)	maligaliurvik matuiqtaulauqtuq 3 :30mi unnusakkut
b)	Mr. Quirke (Clerk-Designate) :	b)	mista kuak (titiraqti - tikkuaqtausimajuq) :
c)	THURSDAY, APRIL 1, 1999	c)	sitamiq, ipuru 1, 1999

Figure 6. *Trois paires de phrases extraites d'un bitexte anglais/inuktitut de débats parlementaires du Nunavut*

4.3. NUNAVUT

Nous avons également mis à profit les textes des débats parlementaires du Nunavut (Canada) afin d'étudier le comportement de PARADOCS sur la paire inuktitut/anglais. Nous avons pour cela utilisé le corpus parallèle réalisé par Martin *et al.* (2003) et mis à la disposition des participants à l'atelier *Building and Using Parallel Texts - Data Driven Machine Translation and Beyond*[26]. Ce corpus est constitué de 155 paires de documents contenant en moyenne 2 000 phrases.

Comme dans l'expérience précédente, ces deux langues ne partagent pas le même alphabet, mais elles utilisent toutes les deux les chiffres arabes (du moins dans nos corpus). La figure 6 présente des exemples qui nous laissent croire que les données numériques sont intéressantes pour notre système même pour cette paire de langues plus atypique. Ceci est confirmé par les meilleures variantes de PARADOCS qui identifient sans erreur les 155 paires de documents parallèles. Seules les configurations faisant usage des entités numériques atteignent les performances optimales. Nous n'avons pas testé notre implémentation de Enright et Kondrak (2007) sur cette tâche puisque l'intersection des hapax dans les deux langues est constituée en grande majorité des données chiffrées. La meilleure configuration de PARADOCS qui fait un usage exclusif des hapax obtient tout de même une f-mesure de 0,487.

4.4. WIKIPEDIA

Il existe de nombreuses pages sur WIKIPEDIA[27] qui sont liées d'une langue à l'autre. Cela ne signifie pas que deux pages liées sont en relation de traduction, mais simplement qu'elles décrivent dans deux langues différentes la même entrée vedette. Ainsi, les pages associées aux entrées « Corpus » (page française) et « *Text Corpus* » (en anglais) sont liées mais ne sont pas parallèles. En revanche, la page française fait mention à la section sur les corpus parallèles que les deux pages liées « le Déclin de l'empire romain d'Occident » (en français) et « *Decline of the Roman Empire* » (en anglais) étaient parallèles en date du 26 octobre 2006[28].

26. `http://www.cse.unt.edu/~rada/wpt/`
27. `http://www.wikipedia.org`
28. C'est encore le cas en date de rédaction de cet article.

Cook is a railway station and crossing loop on the standard gauge Trans-Australian Railway from Adelaide to Perth, with no inhabited places around.**(1)**
The town was created in 1917 when the railway was built and is named after former Prime Minister Joseph Cook.**(2)** The town depended on the Tea and Sugar Train for the delivery of supplies, and is on the longest stretch of straight railway in the world, at 479 km which stretches from Ooldea to beyond Loongana.**(3)** When the town was active, water was pumped from an underground Artesian aquifer but now, all water is carried in by train.**(4)** Attempts have been made to introduce trees and other vegetation, but these have not been successful.**(5)**
Today, it is said to have a resident population of four, and is essentially a ghost town.**(6)** The town was effectively closed in 1997 when the railways were privatised and the new owners did not need a support town there, although the diesel re fuelling facilities remain, and there is overnight accommodation for train drivers.**(7)** Cook is the only scheduled stop on the Nullarbor Plain for the Indian Pacific passenger train across Australia and has little other than curiosity value for the passengers.**(8)** The bush hospital is closed, and the shop is only opened while the Indian Pacific is in town.**(9)** It has a few houses and fuel tanks for the locomotives.**(10)** The crossing loop can cross trains up to 1800m long.**(11)**
An interview with a resident from ulivewhere.com revealed they use the WA and SA time zones,for the WA and SA people, so they don't have to have to adjust their watches, except there is a time difference of around two or three hours.**(12)** When the railway was sold, 8 transportable houses were taken away and made into holiday houses on the coast.**(13)** There are apparently no children in the town.**(14)**

Cook est une gare ferroviaire et surtout un point de croisement sur la voie de chemin de fer à écartement standard, la Trans-Australian Railway qui permet à l'Indian Pacific de relier Adelaide à Perth, sans aucun habitant aux alentours.**(1)** Il y a quelques maisons abandonnées et quelques réservoirs de gas-oil pour les trains.**(10)** Ce point de croisement permet le croisement de trains jusqu'à 1800 mètres de long.**(11)**
La ville fut créée en 1917 lors de la construction de la voie de chemin de fer.**(2)** Elle doit son nom au premier ministre australien de l'époque Joseph Cook.**(2)** La ville fut abandonnée en 1997 quand les chemins de fer furent privatisés et que les nouveaux propriétaires estimèrent qu'ils n'avaient pas besoin de personnel à cet endroit, gardant simplement des réserves de gas-oil en cas de nécessité, des logements pour permettre au personnel du train de se reposer la nuit et un magasin qui ouvre lorsque le train est en gare.**(7)**
Cook est le seul arrêt prévu pour l'Indian Pacific dans la plaine de Nullarbor et n'a pas d'autre intérêt pour le voyageur qu'un arrêt en plein désert dans une ville fantôme.**(8)** Quand la ville était habitée, elle était alimentée en eau par pompage dans un puits artésien mais à l'heure actuelle l'eau est transportée par le train.**(4)** Des essais de plantation de végétaux ont été faits dans la région mais ils n'ont pas été couronnés de succès.**(5)** Cook est sur la plus longue ligne droite de chemin de fer au monde avec une ligne droite de 479 km.**(3)**

Figure 7. *Textes de* WIKIPEDIA *concernant la vedette anglaise «* Cook, South Australia *» et la vedette française associée « Cook (Australie méridionale) ». L'alignement des phrases françaises aux phrases anglaises est indiqué par le jeu d'indices en gras*

4.4.1. *Corpus*

Durant l'été 2009, nous avons aspiré de WIKIPEDIA l'ensemble des pages françaises et anglaises liées entre elles. Un pré-traitement sommaire de ces pages a été effectué afin d'éliminer une partie du marquage spécifique à WIKIPEDIA. Nous avons obtenu un ensemble de 537 067 pages dans chaque langue. Les pages récupérées contiennent en moyenne 445 mots en français et 711 en anglais.

4.4.2. *Protocole*

Nous avons appliqué PARADOCS à ces deux collections de documents. Nous avons pour cela utilisé un arbre de décision avec post-traitement `nodup` entraîné sur un corpus aspiré de la portion anglais/français du site officiel des jeux Olympiques[29]. Pour accélérer le traitement, nous avons limité le nombre de documents retournés par *Lucene* à 5 plutôt que 20. Cette expérience nous permet de vérifier l'utilité de nos classificateurs lorsqu'ils sont appliqués sur des corpus de nature différente à celle des corpus sur lesquels ils ont été entraînés.

Des 537 067 documents sources, 106 896 n'ont reçu aucune réponse de *Lucene*. Un total de 126 438 paires de documents ont été sélectionnées par le classificateur. Le filtre (`nodup`) a éliminé un peu moins de la moitié de ces paires et un total de 69 834 paires de pages WIKIPEDIA ont finalement été appariées par PARADOCS. De ces documents, nous n'avons gardé que ceux étant liés par la langue par WIKIPEDIA pour terminer avec 56 801 paires de documents identifiées comme parallèles.

Vérifier manuellement la véracité de PARADOCS sur l'ensemble de ces paires constituerait un travail colossal, aussi avons-nous entrepris une vérification manuelle d'un échantillon aléatoire de 50 paires de documents identifiées par PARADOCS[30]. Nous avons lu les textes de chaque paire et avons décidé s'ils étaient parallèles, en majorité parallèles, partiellement parallèles ou non parallèles.

Nous tenons à souligner la difficulté d'une telle tâche et le caractère subjectif qu'elle revêt. Nous invitons pour cela le lecteur à consulter la figure 7 qui montre deux fiches de WIKIPEDIA qui décrivent la gare de Cook, l'une en anglais « *Cook, South Australia* »[31], l'autre en français « Cook (Australie méridionale) »[32]. Dans cet exemple, on observe que les 10 phrases du texte français sont en fait des traductions des phrases anglaises. Il ne s'agit cependant pas d'une traduction complète puisque 5 des 14 phrases anglaises ne sont pas traduites[33]. Il est frappant d'observer que l'alignement des phrases est loin d'être monotone et poserait en fait des problèmes à bon nombre de techniques d'alignements phrastiques classiques. La deuxième phrase française est par exemple la traduction de la dixième phrases anglaise. On observe également que les phrases françaises 4 et 5 sont la traduction de la deuxième phrase

29. `http://www.olympic.org`
30. Ce travail délicat a nécessité plusieurs heures d'inspection.
31. `http://en.wikipedia.org/wiki/Cook,_South_Australia`
32. `http://fr.wikipedia.org/wiki/Cook_(Australie_méridionale)`
33. Les phrases 6, 9, 12, 13 et 14.

anglaise. Enfin, l'alignement de la dernière phrase française avec la troisième phrase anglaise est partiel. Cette paire de documents a été notée comme majoritairement parallèle lors de notre évaluation manuelle.

Nous avons souvent observé lors de notre analyse des sorties de PARADOCS que des passages entiers sont produits par traduction sans pour autant qu'un texte au complet soit la traduction de l'autre. En fait, ce travail nous a permis de réaliser que la définition d'un corpus parallèle que nous avons donnée au début de cet article (deux documents véhiculant le même contenu dans le même ordre) est très approximative et que si elle convient aux bitextes habituellement manipulés par la communauté de traduction fondée sur l'exemple (débats parlementaires, textes législatifs), elle ne rend compte que très partiellement des nombreuses situations où l'un des textes est produit en traduisant dans un ordre éventuellement différent des phrases d'un texte source. À l'instar des travaux menés sur les corpus comparables, il serait donc pertinent de définir le degré avec lequel deux documents sont parallèles.

4.4.3. *Résultats*

Cet exercice d'annotation nous a également appris qu'il existe de nombreuses pages véritablement parallèles. WIKIPEDIA s'avère donc une source de bitextes abordant des thèmes variés qui à notre connaissance n'a pas encore été exploitée ni même organisée. Par exemple, les fiches de la vedette française « Mécanique matricielle »[34] et de la vedette anglaise « *Matrix mechanics* »[35] qui font respectivement 57 ko et 53 ko sont en relation de traduction.

Les résultats de notre annotation sont consignés dans le tableau 3. Nous observons que 40 % des paires identifiées par PARADOCS comme parallèles sont en fait des pages parallèles et que 58 % de ces pages sont en fait en majorité parallèles. La raison pour laquelle des pages non parallèles sont identifiées à tort par PARADOCS comme parallèles est liée au fait que les pages sont en fait comparables dans le sens où elles parlent d'une même entrée vedette. Les paires identifiées partagent souvent des entités numériques, parfois dans le même ordre. Par exemple, même si deux biographies (l'une en anglais, l'autre en français) ne sont pas produites par traduction, elles n'en partagent pas moins des dates et des lieux communs qui respectent le plus souvent l'ordre chronologique.

Afin d'apprécier ces résultats, nous avons aussi annoté un échantillon aléatoire de 50 paires d'articles liés par la langue dans WIKIPEDIA. De ces paires, plus de la moitié n'étaient pas parallèles du tout et seulement 26 % étaient parallèles ou majoritairement parallèles.

Des quatre paires de documents parallèles que nous avons annotés de ce corpus, la moitié ont été identifiées par PARADOCS. Nous avons observé deux problèmes qui ont empêché PARADOCS d'identifier les deux autres paires. Nos scripts d'extraction n'ont

34. `http://fr.wikipedia.org/wiki/Mécanique_matricielle`
35. `http://en.wikipedia.org/wiki/Matrix_mechanics`

pas nettoyé certaines informations de formatage et certains tableaux denses en entités numériques. Les nombres dans ces parties qui auraient dû être enlevées ont confondu PARADOCS. Nous avons également observé que *Lucene* considère les nombres communs entre le document et la requête, mais qu'il ne pénalise pas les nombres présents seulement dans le document. Certains documents contenant plusieurs nombres sont donc retournés pour les requêtes formées à partir de documents n'en contenant que quelques-uns.

	WIKIPEDIA	PARADOCS
Parallèle	4	20
En majorité parallèle	9	9
Partiellement parallèle	11	6
Non parallèle	26	15

Tableau 3. *Analyse manuelle d'un échantillon de 50 paires de documents liés par la langue dans* WIKIPEDIA *(première colonne) et de 50 paires de documents identifiées comme parallèles par* PARADOCS *(deuxième colonne)*

5. Conclusion

Nous avons présenté le système PARADOCS, un système capable d'identifier dans une collection de textes de deux langues les documents parallèles. Nous avons montré que le système était indépendant de la paire de langues étudiée et que ses performances étaient meilleures pour les documents les plus grands. Nous avons montré la supériorité de notre système sur l'approche décrite par Enright et Kondrak (2007) et pensons donc avoir démontré l'utilité des entités numériques pour la détection de documents parallèles ; du moins dans les corpus auxquels nous nous sommes intéressés.

Pour l'appariement de transcriptions de débats parlementaires, notre système peut s'ancrer sur les dates et les prix. Les dates sont aussi très présentes dans les corpus journalistiques et dans WIKIPEDIA. Les nombres seraient probablement moins utiles sur des corpus littéraires, mais les systèmes comme PARADOCS sont habituellement utilisés avec des données du Web, où ces textes sont plus rares.

Le système que nous avons présenté est sans aucun doute perfectible. Nous avons notamment mentionné la simplicité de notre étape de post-traitement et des classificateurs que nous avons utilisés ainsi que la nécessité de traiter les cas où la présélection des documents candidats à l'appariement échoue. D'autres traits peuvent facilement être ajoutés aux classificateurs comme le nombre de mots ou de caractères dans les textes à apparier. Nous pensons cependant qu'en l'état, PARADOCS est un composant fiable qui peut être utilisé avec un bon rendement dans les chaînes de collectes de bitextes.

Si l'appariement de documents est un problème de base commun à toute entreprise d'acquisition de bitextes, nous pensons néanmoins qu'un certain nombre de problèmes restent ouverts. En particulier, il convient de revoir les stratégies de fouille de données parallèles sur le Web. Si des approches comme PTMiner (Chen et Nie, 2000) ou STRAND Resnik et Smith (2003) offrent déjà des solutions viables au problème, nous pensons que trop peu d'efforts ont été dédiés à l'acquisition de bitextes de spécialité. La fouille de données scientifiques, par exemple, comme les articles scientifiques pour lesquels il existe souvent des résumés dans plusieurs langues, ou encore les articles que des chercheurs publient dans plusieurs langues (souvent en anglais et dans leur langue maternelle) sont des sources d'informations pertinentes qui requièrent le déploiement de stratégies de recherche d'information ciblées. Nous avons également montré dans cette étude que WIKIPEDIA était une source potentielle de bitextes de spécialité pour autant que l'on sache les retrouver.

Une autre piste de recherche que nous souhaitons poursuivre consiste à unifier les efforts menés en acquisition de corpus comparables et parallèles. Selon nous, il existe un continuum caractérisant les ressources bilingues. À l'une des extrémités de ce continuum, il y a les documents bilingues qui ne sont pas reliés thématiquement. À l'autre extrémité, ceux qui sont en relation de traduction (corpus parallèles). Entre les deux, il y a les corpus plus ou moins comparables.

À notre connaissance, la seule étude tentant de mesurer le degré de comparabilité de documents bilingues est celle de Fung et Cheung (2004). Nous souhaitons adapter PARADOCS à cette problématique. Nous avons également observé sur les pages de WIKIPEDIA que certains documents dans deux langues sont produits par traduction partielle. Identifier des passages parallèles dans des ressources bilingues est une problématique intéressante. Munteanu *et al.* (2004) décrivent un système qui détecte les paires de phrases parallèles dans des corpus comparables. Nous souhaitons vérifier dans quelle mesure l'identification de segments parallèles de l'ordre du paragraphe permettrait d'améliorer un tel système.

Remerciements

Cette recherche a été partiellement financée par le Conseil de Recherche en Science Naturelle et Génie du Canada. Les données utilisées dans l'expérience WIKIPEDIA ont été réunies par Matthew Leon Grinshpun.

6. Bibliographie

Alrahabi M., Desclés J.-P., « Annotation automatique des citations en arabe et en français, vers une carte sémantique des modalités énonciatives », *Congrès de l'ACFAS*, Trois-Rivières, Québec, Canada, 2007.

Altenberg B., Aijmer K., « The English-Swedish Parallel Corpus : A resource for contrastive research and translation studies », *Conference on English Language Research on Computerized Corpora (ICAME 20)*, p. 15-33, 2000.

Bourdaillet J., Huet S., Langlais P., Lapalme G., « TransSearch : from a bilingual concordancer to a translation finder », *Machine Translation*, 2010.

Brown P. F., Pietra S. A. D., Pietra V. J. D., Mercer R. L., « The Mathematics of Statistical Machine Translation : Parameter Estimation », *Computational Linguistics*, vol. 19, n° 2, p. 263-311, 1993.

Callison-Burch C., Koehn P., Monz C., Schroeder J., « Findings of the 2009 Workshop on Statistical Machine Translation », *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Athens, Greece, p. 1-28, March, 2009.

Chen J., Nie J.-Y., « Parallel Web text mining for cross-language IR », *RIAO*, Paris, France, p. 62-67, 2000.

Deléger L., Merkel M., Zweigenbaum P., « Contribution to Terminology Internationalization by Wiord Alignment in Parallel Corpora », *American Medical Informatics Association (AMIA)*, Washington, USA, p. 185-189, 2006.

Enright J., Kondrak G., « A Fast Method for Parallel Document Identification », *NAACL HLT 2007, Companion Volume*, Rochester, NY, p. 29-32, 2007.

Erjavec T., « MULTEXT-East Version 3 : Multilingual Morphosyntactic Specifications, Lexicons and Corpora », *LREC*, Lisbon, Portugal, 2004.

Fung P., Cheung P., « Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus », *Proceedings of the 20th international conference on Computational Linguistics (COLING'04)*, Association for Computational Linguistics, 2004.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., « The WEKA Data Mining Software : An Update », *SIGKDD Explorations*, 2009.

Ishisaka T., Utiyama M., Sumita E., Yamamoto K., « Development of a Japanese-English Software Manual Paralell Corpus », *MT Summit XII*, Ottawa, Canada, 2009.

Koehn P., « Europarl : A Multilingual Corpus for Evaluation of Machine Translation », *10th Machine Translation Summit*, Phuket, Thailand, sep, 2005.

Langlais P., Simard M., Véronis J., « Methods and practical issues in evaluating alignment techniques », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, Montréal, Quebec, Canada, August, 1998.

Ma X., « Parallel Text Collections at Linguistic Data Consorsium », *Machine Translation Summit VII*, Singapore, sep, 1999.

Ma X., Liberman M., « BITS : A Method for Bilingual Text Search over the Web », *Machine Translation Summit VII*, Singapore, sep, 1999.

Macklovitch E., Simard M., Langlais P., « TransSearch : A Free Translation Memory on the World Wide Web », *Second International Conference On Language Resources and Evaluation (LREC)*, vol. 3, Athens Greece, p. 1201-1208, jun, 2000.

Maeda K., Ma X., Strassel S., « Creating Sentence-Aligned Parallel Text Corpora from a Large Archive of Potential Parallel Text using BITS and Champollion », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may, 2008.

Martin J., Johnson H., Farley B., Maclachlan A., « Aligning and Using an English-Inuktitut Parallel Corpus », *HLT-NAACL Workshop : Building and Using Parallel Texts - Data Driven Machine Transkation and Beyond*, Edmonton, Canada, p. 115-118, May, 2003.

Megyesi B. B., Johansson E. C., Hein A. S., « Building a Swedish-Turkish Parallel Corpus », *LREC*, Genoa, Italy, 2006.

Moore R. C., « Fast and Accurate Sentence Alignment of Bilingual Corpora », *Proceedings of the fifth Conference of Association for Machine Translation in the Americas (AMTA)*, Tiburon, California, p. 135-144, oct, 2002.

Munteanu D. S., Fraser A., Marcu D., « Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora », *in* D. M. Susan Dumais, S. Roukos (eds), *HLT-NAACL 2004 : Main Proceedings*, Association for Computational Linguistics, Boston, Massachusetts, USA, p. 265-272, May, 2004.

Nadeau D., Foster G., « Real-Time Identification of Parallel Texts from Bilingual News feed », *CLINE 2004*, Computational Linguistics in the North East, 2004.

Nie J.-Y., Cai J., « Filtering Noisy Parallel Corpora of Web Pages », *IEEE Symposium on Natural Language Processing and Knowledge Engineering*, Tucson (AZ), USA, p. 453-458, 2001.

Och F. J., Ney H., « A Systematic Comparison of Various Statistical Alignment Models », *Computational Linguistics*, vol. 29, n° 1, p. 19-51, March, 2003.

Patry A., Langlais P., « Paradocs : un système d'identification automatique de documents parallèles », *12e Conference sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France, p. 223-232, jun, 2005.

Resnik P., Olsen M. B., Diab M., « The Bible as a Parallel Corpus : Annotating the 'Book of 2000 Tongues' », *Computers and the Humanities*, vol. 33, n° 1-2, p. 129-153, 1999.

Resnik P., Smith N. A., « The Web as a Parallel Corpus », *Computational Linguistics*, vol. 29, p. 349-380, 2003. Special Issue on the Web as a Corpus.

Shi L., Niu C., Zhou M., Gao J., « A DOM tree alignment model for mining parallel data from the web », *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, Sydney, Australia, p. 489-496, 2006.

Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufis D., Varga D., « The JRC-Acquis : A multilingual aligned parallel corpus with 20+ languages », *5th International Conference on Language Resources and Evaluation*, Genoa, Italy, p. 2142-2147, May, 2006.

Tiedemann J., « News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces », *in* N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds), *Recent Advances in Natural Language Processing*, John Benjamins, Amsterdam/Philadelphia, p. 237-248, 2009.

Véronis J. (ed.), *Parallel Text Processing, Alignment and Use of Translation Corpora*, Kluwer Academic, 2000.

Wessel K., Jian-Yun N., Michel S., « Embedding web-based statistical translation models in cross-language information retrieval », *Computational Linguistics*, vol. 29, n° 3, p. 381-419, 2003.