Alignement de documents multilingues sans présupposé de parallélisme

Charlotte Lecluze

▶ **To cite this version:**

Charlotte Lecluze. Alignement de documents multilingues sans présupposé de parallélisme. Traitement du texte et du document. Université de Caen, 2011. Français. NNT : . tel-01075742

HAL Id: tel-01075742
https://hal.science/tel-01075742

Submitted on 20 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE CAEN BASSE-NORMANDIE

U.F.R. DE SCIENCES

ÉCOLE DOCTORALE
STRUCTURE, INFORMATION, MATIÈRE ET MATÉRIAUX

THÈSE

présentée par

CHARLOTTE LECLUZE

et soutenue

le 5 décembre 2011

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE CAEN

Spécialité : informatique et applications

Arrêté du 7 août 2006

ALIGNEMENT DE DOCUMENTS MULTILINGUES SANS PRÉSUPPOSÉ DE PARALLÉLISME

MEMBRES DU JURY

M. Philippe LANGLAIS, professeur, université de Montréal *(rapporteur)*
M. Eric GAUSSIER, professeur, université de Grenoble *(rapporteur)*
M. Patrick CONSTANT, président et fondateur de Pertimm
M^me^ Christine DURIEUX, professeur, université de Caen
M. Emmanuel GIGUET, chargé de recherche HDR, université de Caen *(co-directeur)*
M. Jacques VERGNE, professeur, université de Caen *(directeur)*

MERCIS

Merci à Jacques Vergne et Emmanuel Giguet d'avoir ouvert les portes du laboratoire à des étudiants venant d'un autre horizon. Merci pour votre encadrement tout au long de cette thèse, pour votre investissement et votre grande disponibilité à tous les deux, ainsi que pour vos remarques enrichissantes et surtout pour la confiance que vous m'avez accordée. Je sais que sans vous cette thèse n'aurait pu aboutir.

Merci à Pertimm de m'avoir accueillie pendant ces trois années, me permettant de m'enrichir au contact de son équipe, jeune, innovante.

Je remercie Éric Gaussier et Philippe Langlais d'avoir accepté de rapporter sur cette thèse, et Christine Durieux et Patrick Constant de faire partie du jury.

Merci à Loïs Rigouste et Romain Brixtel, je suis très heureuse d'avoir croisé vos routes. Acteurs et témoins « privilégiés » (si l'on peut dire ! !) de ma mutation. Il vous en aura fallu de la patience pour m'épauler dans cet apprentissage tardif de l'informatique ! Merci à vous deux pour toutes ces discussions fructueuses, toujours dans la bonne humeur !

Merci à Régis Clouard d'avoir immédiatement adhéré au projet, de s'être toujours montré disponible et intéressé. Et de m'avoir fait bénéficier de ses précieuses compétences en traitement d'images.

Merci à Gaël, Leslie et Estelle, qui ont bien voulu prendre de leur temps pour me relire, même quand ils n'en avaient pas !

Merci enfin à ma famille et à mes amis, qui pendant ces trois années de travail m'ont toujours entourée et encouragée. Et un merci tout spécial à Samuel, qui a vécu (et survécu à) ces années bien spéciales au jour le jour !

SOMMAIRE

INTRODUCTION

La traduction : Un enjeu de société

Le web est à l'origine d'une explosion de l'information. Chaque jour, le nombre de textes disponibles en différentes langues augmente et avec lui la nécessité de faire face à un flux d'informations résolument multilingue. Celle-ci est spécialement ressentie par les instances européennes et mondiales qui doivent non seulement préserver la diversité linguistique en soutenant l'apprentissage des langues étrangères, mais également garantir l'égalité des e-citoyens européens en assurant l'accès aux documents dans leur propre langue. Cependant cet objectif s'avère humainement difficile à atteindre puisque le processus de traduction fait que l'on traduit vers sa langue maternelle et qu'il n'existe pas suffisamment de traducteurs pour certains couples de langues.

C'est face à ce double constat de nécessité et d'incapacité qu'a mûri l'idée de convertir et valoriser les traductions réalisées par des traducteurs humains. Au début du XIX^{ème} siècle, Champollion face à la Pierre de Rosette prenait déjà conscience qu'un document traduit en plusieurs langues peut s'avérer une grande source de connaissances sur les langues en présence : lexicales, syntaxiques…

Les organisations ayant un rayonnement international proposent des informations en différentes versions linguistiques : documentation technique, texte réglementaire, document contractuel, information commerciale, communiqué de presse.

Des opérations de rétro-ingénierie sur ces documents peuvent apporter une aide tant en amont du processus de traduction qu'en aval. En amont, elles participent à la création d'outils d'aides à la traduction : ressources dictionnairiques, terminologiques, mémoires de traduction. En aval, elles peuvent s'avérer utiles pour contrôler a posteriori la traduction, voire le cas échéant pour orienter une révision de la traduction en mettant par exemple en lumière certaines divergences entre le texte source et le texte cible. Ces outils visent à augmenter la productivité de traducteurs humains. Cela est rendu possible par la croissance des capacités de calcul des ordinateurs. Ces traductions d'une même information font depuis plusieurs années l'objet de recherches en Traitement Automatique des Langues. L'informatique alliée à la linguistique de corpus offrent un nouveau regard sur ce matériau linguistique.

Les techniques qui permettent la mise en correspondance de zones sémantiquement équivalentes, sont des techniques dites d'alignement. Les correspondances sémantiques peuvent être faites à plusieurs niveaux : paragraphes, phrases, mots…

L'état de l'art pour automatiser cette mise en correspondance fait l'hypothèse simplificatrice du parallélisme au niveau sur-phrastique, hypothèse qui sous-tend que l'ordre du discours est globalement préservé.

Cependant celle-ci n'est pas toujours vérifiée et des verrous demeurent qui empêchent de valoriser pleinement cette mine d'informations, d'en extraire aussi massivement qu'envisagé des ressources pourtant utiles tant aux traducteurs qu'aux lexicologues. Il nous semble qu'il existe une marge de progression. Certains aspects des documents parallèles méritant d'être approfondis, notamment leur mise en forme et les cas d'inversions et de suppressions au niveau sur-phrastique.

Nos travaux portent sur la recherche d'une méthode d'alignement prenant en considération le travail de réécriture que constitue la traduction.

À l'image de notre cursus universitaire, ces travaux sont de deux types : observations linguistiques et réalisations informatiques. Notre démarche consiste à partir d'une observation multi-échelle des documents multilingues pour mettre en place une méthode générique d'extraction d'équivalences sémantiques entre ces traductions.

L'objectif de ces travaux est double : appariement et alignement, i.e. création de ressources et analyse de document.

La première partie de ce document pose les bases nécessaires à l'élaboration de notre méthode d'alignement, en mettant l'accent sur la question du parallélisme à travers différentes illustrations en contexte et une vue d'ensemble des méthodes d'alignement. La deuxième partie met ces observations à profit pour dégager une méthode sans présupposé de parallélisme. Enfin la troisième partie expose la mise en œuvre de cette méthode.

Première partie

DE L'ÉTUDE DE CORPUS DE DOCUMENTS PARALLÈLES À L'ÉTUDE DE COLLECTIONS DE MULTIDOCUMENTS

1

OBSERVATIONS LINGUISTIQUES ET TRADUCTOLOGIQUES SUR LES DOCUMENTS PARALLÈLES

« Peut-on se contenter de soutenir que, traduire, c'est dire la même chose en d'autres mots ? Pas si simple, […] la ligne de partage entre simple reproduction, traduction et libre adaptation est pour le moins fluctuante. […] la traduction, avant d'opérer *ab extra*, à la frontière extérieure des langues, travaille de l'intérieur de la moindre de nos paroles [1]. Voilà donc que [cette] problématique acquiert une portée insoupçonnée au départ : aussi vaste désormais que le langage lui-même. » (Ost, 2009, p.13)

Dans ce premier chapitre, nous parcourons pas à pas la distance qui sépare une langue d'une autre. Nous commençons par une description de l'opération traduisante. Puis, nous présentons plusieurs phénomènes linguistiques dont l'actualisation est propre à chaque langue, ce que nous illustrons à travers des exemples de traductions multilingues pris en contexte. Une telle observation traductologique témoigne rapidement de certaines nécessités si l'on souhaite mettre en œuvre un système de Traitement Automatique des Langues et plus particulièrement, comme c'est notre cas, un système d'alignement. Ce premier chapitre nous amènera naturellement au chapitre 2 consacré à un rappel de l'existant méthodologique en matière d'alignement.

SOMMAIRE

1. Chaque langue (…) peut se traduire elle-même. (Dakhli, 2009)

1.1 LA TRADUCTION : UNE OPÉRATION LINGUISTIQUE ET HUMAINE

La traduction est une opération complexe : logique, psychologique et linguistique ; au même titre que l'énonciation à l'origine du document source. Ce n'est pas un processus linéaire. Il s'agit au contraire d'un processus circulaire qui commence par une interprétation globale d'un texte en langue source, révisée ensuite par une analyse du texte source et l'élaboration de stratégies pour produire le texte cible. Cette suite de procédés contient elle-même un grand nombre de mouvements circulaires plus petits ou « boucles » qui ne cessent de revenir sur le texte source et sa situation, le texte cible et sa situation, les niveaux d'analyse individuels et sur l'analyse du texte source et la production du texte cible. Le traducteur doit ainsi constamment reconsidérer des éléments déjà analysés, chaque information obtenue au fil du processus d'analyse et de compréhension nécessite d'être confirmée et corrigée à travers le prisme des nouveaux éléments. (Nord, 2010)

Concrètement le passage d'un document d'une langue à une autre dépend entre autres :

- de la langue source ;
- de la langue cible (Chamsine, 2005) ;
- du destinataire : connaissances, cultures… (Abudayeh, 2010) ;
- du traducteur : compétences, connaissances du domaine (concepts et terminologie), mais également connaissances des cultures et des langues sources et cibles… ;

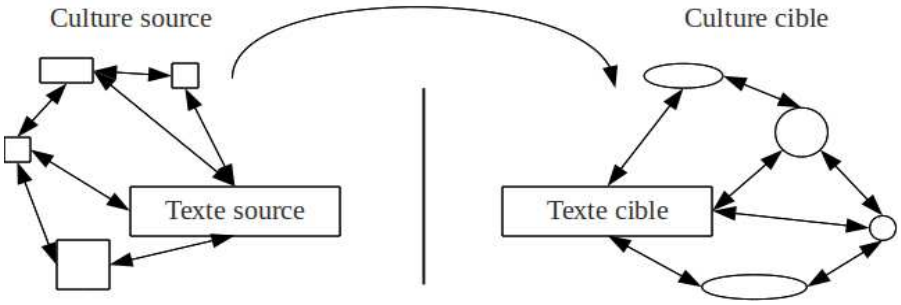


FIGURE 1 – L'intertextualité dans le processus de traduction (Nord, 2010). Les rectangles à gauche de la figure de même que les ovales à droite symbolisent des textes en relation avec les textes source et cible : articles, romans…

- du type de document : « On ne traduit pas de la même façon un bulletin météo, une dépêche diplomatique ou un texte littéraire. Parmi ces derniers, on n'assimilera pas la traduction d'un roman à celle d'une poésie, dont il convient de rendre avant tout la musicalité, ou celle d'une pièce de théâtre, dont il importe de

restituer l'efficacité scénique et le rythme des dialogues » (Ost, 2009, p.227)

– des outils à disposition : dictionnaires monolingues et bilingues, des documents auxiliaires : parallèles ou comparables (voir chapitre 2)…

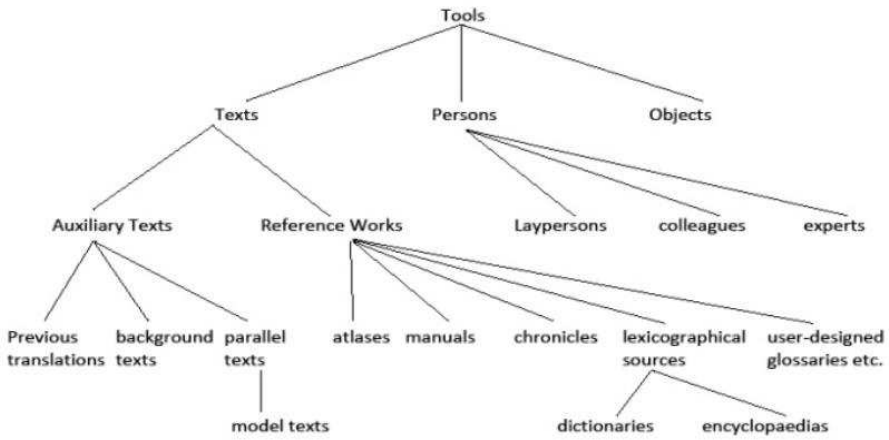


FIGURE 2 – Les outils du traducteur (Nord, 2002)

La problématique de la traduction est donc « aussi vaste que le langage lui-même », c'est donc naturellement que les traductions sont devenues des objets d'études à part entière et qu'une science proposant d'en faire l'étude est née dans les années 60.

1.2 LES TRADUCTIONS : DES OBJETS D'ÉTUDE

La *traductologie (translatology (Harris, 1988))* est le nom donné par Harris en 1973 à la méta-opération d'ordre analytique ayant pour objet les traductions (Harris, 1973). Ce néologisme est à l'époque venu remplacer les périphrases : « the Sciences of Translation » (Nida, 1964) et « the Linguistic Theory of Translation » (Catford, 1965). L'objet primordial de la traductologie est la traduction naturelle traitée d'une façon descriptive et explicative. Le travail du traductologue se situe donc en aval de la traduction, au niveau du texte traduit et pas de la traduction, il n'a initialement pas de vocation prescriptive. Les traductologues s'intéressent notamment à des phénomènes tels que : la surtraduction, la perte de sens, l'erreur, le foisonnement (Durieux, 1990; Juhel, 1999; Cochrane, 2007; Ballard, 1999) ; dont ils identifient les causes. Harris dit conserver le terme « traduction » pour l'opération traduisante, et celui de « texte traduit » pour son produit. Quelques années après dans (Harris, 1988), il propose le terme *bi-texte* pour désigner le couple texte source-texte cible, par extension on trouve également le terme *multi-texte* pour désigner un ensemble constitué d'un texte source et de ses traductions dans plusieurs langues. Zimina (2006), quant à elle, propose de nommer *volet* chacune

des versions d'un tel ensemble. Dans le laboratoire du GREYC, nous avons créé le néologisme *multidocument* qui inclut, en tant que grain supérieur au multitexte, les dimensions de mise en forme matérielle et de structure des documents.

1.3 DES TÉMOINS PRIVILÉGIÉS DE LA VARIÉTÉ DES LANGUES

Un rapide tour d'horizon de traductions en langues européennes témoigne déjà des principales différences morphologiques et syntaxiques que peuvent avoir des langues entre elles, et par là d'une part des difficultés à traduire, mais également d'autre part à mettre en place d'éventuelles opérations de rétro-ingénierie sur des documents traduits, telles que l'alignement.

1.3.1 *Au niveau morphologique*

Si à l'intérieur d'un document, par souci de cohésion interne, un terme est habituellement traduit au moyen du même équivalent, il existe des possibilités de variations morphologiques (Giguet, 2005), entraînant des décalages d'effectifs de ces équivalents. Ce problème est particulièrement présent et gênant dans les langues flexionnelles [2], et dans les langues agglutinantes [3] qu'elles englobent et qui déclinent le groupe nominal.

Les langues suivantes déclinent le groupe nominal : allemand (4 cas), finnois (15), grec (4), hongrois (18), letton (6), polonais (7). Le finnois et le hongrois utilisent un nombre important de cas, et n'utilisent donc pas, comme le français par exemple, les adpositions. Le sens d'une préposition française est souvent traduit par un suffixe dans ces langues, que ce soit une désinence flexionnelle ou une postposition, qui se distinguent mal. Cette grande diversité de cas couvre des nuances très précises, comme en témoignent les quinze cas du finnois.

Le statut du mot

Ainsi, définir le concept de mot, ne serait-ce que pour les langues européennes, s'avère déjà complexe. Cela dépend en fait du point de vue adopté : lexical ou graphique. Ces deux points de vue ne sont pas toujours en correspondance.

2. Dans une langue flexionnelle, les radicaux sont pourvus d'affixes grammaticaux variables et exprimant plus ou moins à la fois, par exemple, le genre, le nombre et le cas, ou la personne, le temps, le mode, la voix… La plupart des langues européennes sont des langues considérées comme flexionnelles.

3. Dans une langue agglutinante, on juxtapose au radical une série de morphèmes distincts servant à exprimer les rapports grammaticaux. Dans ce type de langue, chacun des affixes (préfixes, infixes ou suffixes) est clairement analysable et identifie précisément une fonction grammaticale ou syntaxique.

Considérons pour illustrer ce fait, les traductions du syntagme nominal « les transports en commun » dans 4 langues européennes [4] présentant une disparité notable du grain mot : anglais (en), français (fr), hongrois (hu), finnois (fi) présentées dans le tableau 1 :

LANGUE	MOT POLYLEXICAL	NOMBRES DE MOTS GRAPHIQUES
fr	transport en commun	3 mots graphiques
en	public transport	2 mots graphiques
hu	a tömegközlekedés	2 mots graphiques
fi	joukkoliikenne	1 mot graphique

Tableau 1 – Illustration du décalage interlangue entre le niveau lexical et le niveau graphique du concept de mot, à partir de l'exemple de « transport en commun ».

Cette question est d'autant plus complexe que l'on a à traiter des *mots polylexicaux* (ou complexes) à savoir « toute unité composée de deux mots simples ou mots dérivés préexistants [...] les mots polylexicaux (ou complexes) peuvent être soudés (et alors, du point de vue informatique, ils peuvent être assimilés à des mots simples) [...] ou comporter un séparateur » [5]. La forme graphique d'une unité lexicale composée tient de propriétés intralangues. Elle dépend des particularités morphologiques de flexions et de dérivations de chaque langue.
Au regard de ces caractéristiques morphologiques, le mot graphique n'apparaît pas suffisamment universel pour établir des correspondances. Une autre granularité doit être recherchée pour répondre au besoin de comparativité d'un système multilingue d'alignement, qui plus est, sans présupposé.

Le foisonnement

Le foisonnement est le terme utilisé pour définir « en traduction, (...) la prolifération de mots en surnombre, (...) l'augmentation de volume du texte d'arrivée par rapport au texte de départ. » (Durieux, 1990). Celui-ci peut-être fortuit et résulter d'un défaut de méthode. Mais sans nier le rôle du traducteur dans le foisonnement et sans envisager non plus un simple transcodage, nous estimons que, quoi qu'il arrive, certaines langues sont intrinsèquement plus foisonnantes que d'autres et qu'il existe une sorte de « servitude linguistique » [6] à laquelle le traduc-

4. Nous utilisons à partir d'ici les codes de langue tels qu'ils sont définis par la norme ISO 639-1.

5. G. Gross (2004) cité par (Neveu, 2004)

6. Le terme « servitude linguistique » désigne les contraintes auxquelles le traducteur est contraint pour respecter la syntaxe de la langue (p. ex. ajout d'articles et de joncteurs, étoffement des prépositions, etc)(Cochrane, 2007).

teur doit se plier. Nous constatons cependant que les variations peuvent autant correspondre à une réduction qu'à une augmentation du volume de mots d'un document, lors de sa traduction d'une langue à une autre. Les coefficients sont, en moyenne, ceux présentés dans le tableau 2, ils nous ont été fournis par l'ARI [7].

LANGUE D'ORIGINE	FRANÇAIS
anglais	+20%
allemand	+30%
néerlandais	+20%
italien	-10%
espagnol	-10%
portugais	-10%
suédois	+30%
danois	+30%
norvégien	+30%
japonais	-67%

TABLEAU 2 – Coefficients de foisonnement fournis par l'ARI.

Le tableau 2 montre que le japonais est beaucoup moins foisonnant que le français. La théorie de l'information peut nous en apporter une explication. La quantité d'information associée à un symbole de probabilité p est $\log \frac{1}{p}$. Si l'on considère, grossièrement, que les caractères sont équiprobables [8], la quantité d'information associée à chaque caractère est donc $\log \frac{1}{1/n} = \log n$ pour un alphabet de taille n.

Dès lors, si l'on suppose, là encore en simplifiant beaucoup, qu'il y a 26 caractères possibles en français et 7000 en japonais, on obtient que la quantité d'information est identique entre un texte de 1000 occurrences en français et un texte de 400 caractères en japonais : $1000 \log 26 \approx 400 \log 7000$. En d'autres termes, plus intuitifs, puisqu'on a le choix entre un plus grand nombre de caractères, chaque caractère est beaucoup plus précis et permet d'exprimer plus de choses. Incidemment, cela explique aussi pourquoi l'unité sémantique constituée par le mot est souvent de deux caractères uniquement en chinois et toujours beaucoup plus en moyenne dans les langues européennes.

Avant de servir à l'illustration des différences entre les langues, les coefficients de foisonnement ont tout d'abord un intérêt reconnu en matière de tarification des traductions. Les organismes professionnels conseillent en effet aux traducteurs d'en tenir compte pour établir leur

7. ARI, Assistants Record International : traduction, rédaction, PAO, interprétation, conseil ; 11, Rue des Réglises, 75020 Paris.

8. Ce qui est, bien sûr, tout à fait faux en pratique, mais permet ici de simplifier le propos en conservant l'essentiel de l'argumentation.

devis. Pour cela, ces organismes diffusent des coefficients de foisonnement de référence, c'est-à-dire la différence envisagée de volume entre le texte original et le texte traduit. Dans le tableau 2, nous présentons les seules données officielles que nous avons pu nous procurer. Celles-ci témoignent bien des variations de volume qui naissent de l'opération traduisante, c'est-à-dire que le volume d'un même texte varie selon la langue. Ainsi, par exemple, lors d'une traduction de l'anglais vers le français, le nombre de mots français sera plus important que le nombre de mots anglais. En outre, plus le texte est technique, plus le coefficient risque d'être élevé.

Concrètement, la tarification est le plus souvent établie au nombre de mots. Il existe une normalisation des mots, lignes, pages et feuillets :

- une page ou un feuillet contient 250 mots ou 1500 signes/caractères ;
- une ligne contient 10 mots et un mot contient environ 6 signes ou caractères.

Il reste toutefois un certain nombre de langues ou pays dans lesquels l'unité est plutôt la page ou la ligne.

1.3.2 *Au niveau syntaxique*

La métataxe

Dans son ouvrage intitulé *Éléments de syntaxe structurale*, Lucien Tesnière consacre le livre E à la présentation de la « métataxe ». Il s'y intéresse notamment au changement structural qui peut intervenir entre une phrase à traduire et une phrase traduite, c'est-à-dire lors du mécanisme de traduction.

Le plan structural et le plan sémantique sont théoriquement indépendants l'un de l'autre. La métataxe n'est qu'une application de ce principe de l'indépendance du structural et du sémantique. Elle correspond à la différence de stemma (changement structural) qui existe entre la phrase à traduire et la phrase traduite (sans changement sémantique), c'est-à-dire qu'elle intervient chaque fois que la structure actancielle d'un verbe diffère d'une langue à une autre.

La métataxe peut avoir plusieurs degrés, elle peut être simple ou complète :

- simple appel à une catégorie grammaticale différente (tableau 3) : chaque langue établit ses propres correspondances entre catégories de la pensée et catégories grammaticales, c'est pourquoi la traduction d'une langue à une autre nécessite parfois l'appel à une catégorie grammaticale différente.

  À cela s'ajoute un principe de solidarité métataxique. Quand un mot est solidaire d'un autre, le passage métataxique du premier

ALLEMAND	>	FRANÇAIS
Idée de déplacement	(= changement de lieu)	
Adverbes résultatifs ou particules séparables	>	Verbe à l'impératif
Adverbe : **Fort !**		Verbe : **Va-t-en !**

TABLEAU 3 – Simple appel à une catégorie grammaticale différente.

à une autre catégorie grammaticale a automatiquement pour effet d'entraîner parallèlement une transformation métataxique équivalente du second qui lui est solidaire. Concrètement, si on change un substantif en verbe ou inversement, il y a lieu de changer parallèlement l'adjectif en verbe ou inversement. Ceci est valable aussi bien en monolingue qu'en multilingue.

– transformation complète de l'ordonnance structurale avec changement de nœud central (tableau 4) : on dit qu'il y a interversion des actants, quand à un actant d'une langue correspond sémantiquement un autre actant dans une autre langue. La traduction de l'une à l'autre n'est possible qu'en changeant la nature de l'actant. Même si le niveau sémantique prévaut sur le structural, un verbe dont on connaît le sens, mais dont on ignore la structure actancielle, est inutilisable, d'où l'importance de la structure actancielle dans le passage d'une langue à une autre. Sans rappeler toute l'étude de la métataxe que propose Tesnière, on peut simplement rappeler que cette interversion des actants peut être notamment simple, double, intervenir entre des actants et des circonstants, ou dans le passage de l'actif au passif…

LATIN	>	FRANÇAIS
Tela milites deficiunt		les armes font défaut aux soldats
Actant 2		Actant 3

TABLEAU 4 – Transformation complète de l'ordonnance structurale avec changement de nœud central.

Cette liste des différents types de métataxe que Tesnière nous offre se veut une sorte de mode d'emploi de la traduction, grâce auquel il est possible d'éviter les pièges. Ici, pour nous, qui nous situons en aval de la traduction, il nous sert à prendre conscience de tous les changements qui s'opèrent dans le passage d'une langue à une autre, tant au niveau de l'ordre des constituants que de leur nature (même si ce dernier aspect ne nous intéresse que peu ici).

Les différents schémas Sujet-Verbe-Objet (SVO)

Il existe des différences de syntaxe courantes y compris entre des langues de la même famille linguistique. Ainsi, parmi les langues indo-européennes notamment, l'on dénombre plusieurs schémas SVO plus ou moins contraints. L'ordre des constituants de la phrase n'est donc pas nécessairement invariant et peut également poser problème, quand en allemand ou en grec par exemple, l'ordre de la phrase peut être Sujet-Verbe-Objet (SVO) ou Objet-Verbe-Sujet (OVS), (ou encore SOV parfois en allemand). Il n'est alors pas toujours évident de définir cet ordre et donc, dans le cadre d'une méthode d'alignement, d'établir des alignements. Les ressources linguistiques, à ce propos, elles-mêmes se contredisent parfois.

En théorie, en ce qui concerne les langues européennes, on attribue aux langues les ordres suivants :

- **langues SVO :** allemand (de), anglais (en), bulgare (bg), danois (da), espagnol (es), estonien (et), finnois (fi), français (fr), grec (el), italien (it), letton (lt), litunien (lt), maltais (mt), néerlandais (nl), polonais (pl), portugais (pt), roumain (ro), slovaque (sk), slovène (sl), suédois (sv), tchèque (cs) ;
- **langue SOV :** hongrois (hu), néerlandais ;
- **langue VSO :** espagnol, néerlandais ;
- **langue OSV :** roumain ;
- **ordre libre :** finnois, hongrois, polonais, slovaque, slovène.

Généralement, les langues sans déclinaison, comme le français ou l'espagnol, ont un ordre plus strict que celles qui se déclinent, mais ce n'est pas une règle. Les langues finno-ougriennes, utilisant peu la coordination ou la subordination au profit de la juxtaposition, donnent beaucoup d'importance à l'ordre des mots.

Nous pouvons également ici évoquer l'ordre déterminant/déterminé. Si en français l'ordre est principalement déterminé⇒déterminant, en anglais la règle est davantage celle du déterminant→déterminé, quoique les deux se rencontrent (tableau 5).

déterminé ⇒ déterminant	déterminant ⇒ déterminé	déterminé ⇔ déterminant
fr, es, it, mt, pt, ro	bg, cs, da, et, fi, hu, lt, lv, nl, sl, sk, sv	de, en, el, pl

TABLEAU 5 – Ordre déterminant-déterminé des langues de l'Union Européenne

Ainsi, en n'observant ne serait-ce que le couple français-anglais, il nous est déjà permis de rencontrer les deux cas de figure la similitude (ordre SVO) et la différence (ordre déterminant-déterminé) d'ordre au

niveau sous-phrastique. Nous les illustrons en contexte dans la section 1.3.3.

1.3.3 *Similitude et différence d'ordre au niveau sous-phrastique*

L'ordre des mots d'une phrase n'est généralement pas considéré comme préservé dans le passage d'une langue à une autre (figures 3a et 3b).

Cependant, le niveau sous-phrastique peut lui aussi être globalement préservé dans le passage d'une langue à une autre (figure 4), et les unités qui le composent dans le même ordre.

1.4 LES TRADUCTIONS : DES ÉNONCIATIONS UNIQUES

Le travail du traducteur constitue un véritable travail d'écriture (réécriture). Nous illustrons dans les sous-sections qui suivent quelques phénomènes résultant cette fois de la liberté d'adaptation dont bénéficie le traducteur et entraînant un foisonnement davantage artificiel que celui lié aux servitudes linguistiques.

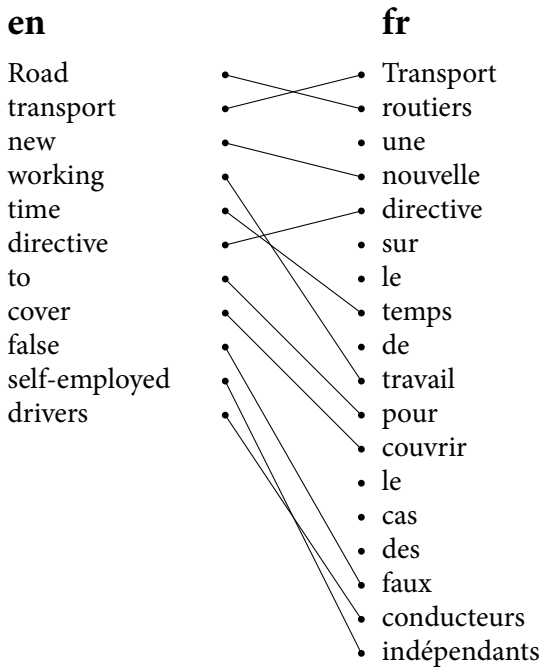
1.4.1 *L'implicite et l'explicite*

Des éléments sous-entendus, c'est-à-dire évoqués de manière implicite, dans certaines langues apparaissent de façon explicite dans d'autres langues. Ceci constitue évidemment un frein à l'alignement d'unités sémantiquement équivalentes, puisque certaines d'entre elles n'ont pas d'équivalent clairement explicité.

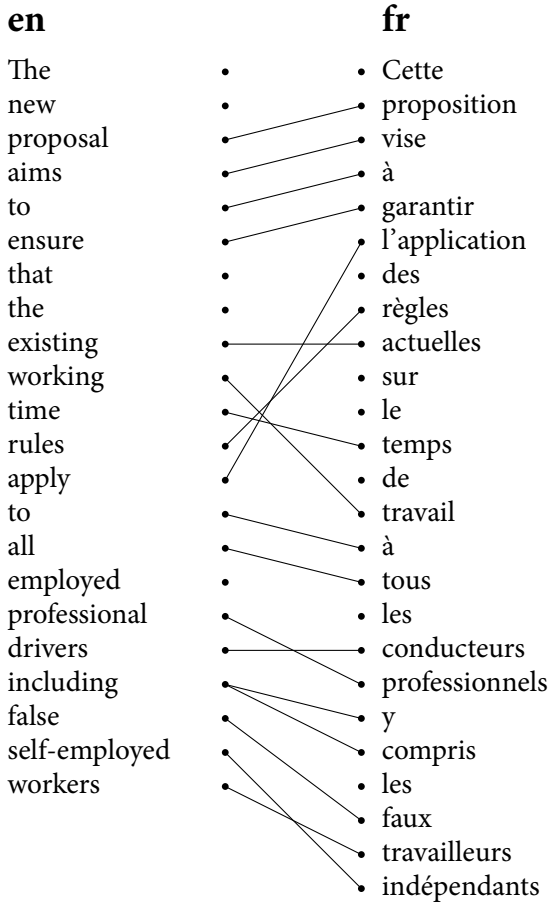
Exemple : document IP/05/975, ligne 8
FR : Jacques Barrot, Vice-Président de la Commission européenne, responsable des transports, a déclaré : (…)
ES : Jacques Barrot, Vicepresidente de la Comisión Europea y responsable de la política de transportes, se ha expresado **en los siguientes términos** : (…)

Exemple : document IP/05/975, ligne 9
FR : Les collectivités pourront soit fournir leurs transports locaux en régie, soit les confier en toute transparence à un opérateur spécialisé.
FI : Paikallisviranomaiset voivat joko vastata itse paikallisliikenteen palvelujen tarjonnasta tai uskoa niiden tarjonnan avoimelta pohjalta jollekin erikoistuneelle **liikenteen**harjoittajalle.

Le fait qu'il s'agisse d'un opérateur de transport est en finnois clairement explicité « **liikenteen**harjoittajalle », à la différence de ce qui est proposé dans la version française « un opérateur spécialisé ». Le fran-



(a) Titre de communiqué de presse.



(b) Résumé de communiqué de presse.

FIGURE 3 – Différence de l'ordre des mots au niveau sous-phrastique entre les extraits anglais et les extraits français.

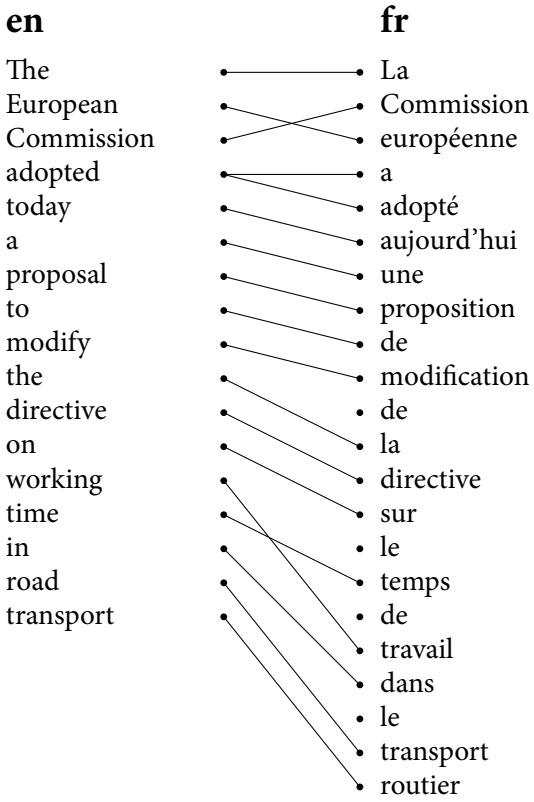


FIGURE 4 – Similitude de l'ordre des mots entre une série de phrases en anglais et leur traduction en français.

çais fait l'économie de ce complément puisqu'en début de phrase, le sujet des transports est clairement cité. Nous pouvons rapprocher cette particularité du finnois, de sa qualité de langue très redondante, peu anaphorique et utilisant peu les synonymes. Plus généralement, on peut dire que tout ce qui est contenu dans les phrases de départ, implicitement ou explicitement, l'est dans les phrases d'arrivée, implicitement ou explicitement.

1.4.2 *La synonymie*

« La synonymie est une relation sémantique fondée sur une similarité de signifiés entre des signifiants distincts. […] Cette similarité de signifiés est souvent présentée comme pouvant être totale ou partielle. En fait, le lien étroit dans le signe linguistique entre le signifié et le signifiant rend la synonymie totale, qui est indifférente au contexte, pratiquement inobservable en discours. Car si deux lexèmes peuvent manifester une référence similaire, l'effet sémantique produit par chacun d'eux ne révèlera pas la même situation énonciative. » (Neveu, 2004)

Cette définition de la synonymie et l'illustration qui en est faite au travers du tableau 6 témoignent bien du frein à l'alignement sémantique monolingue et multilingue que ce phénomène constitue.

LANGUE	FR	FI	EN	EL	ES
	Donner (l.4)	tarjotaan	Giving	Να δοθούν	Proporcionar a
	donner (l.6)	tarjoaa	provides	παρέχει	ofrece
	offrir (l.7)	tarjoamiseksi	offering	να παρέχεται	garantizar
	fournir (l.9)	tarjonnasta	running	να προσφέρουν	ofrecer
	fournit (l.13)	tarjoaa	provides	παρέχει	proporciona
	apportera (l.14)	taataan	will offer	θα προσφέρει	aportará
	une offre (l.14)	tarjonaan		την παροχή	una oferta de
	de fournir (l.17)	tarjonnasta	to provide	να παρέχουν	de proporcionar
	apporte (l.26)	merkitsee	provide	παρέχει	supone
Nombre de signifiants différents	4	3	4	3	5
Nombres d'occurrences du signifié « donner »	9	9	8	9	9

TABLEAU 6 – Illustration du phénomène de synonymie dans le multidocument IP/05/975 en français (fr), finnois (fi), anglais (en), grec (el) et espagnol (es).

1.4.3 *L'anaphore*

Ce terme désigne : « une relation référentielle qui s'exerce à l'intérieur du discours entre deux expressions linguistiques, dont l'une, dite anaphorique (ou forme de rappel), reçoit son interprétation de l'autre, dite source de l'anaphore (ou antécédent) qui lui est antéposée. » (Neveu, 2004)

Au travers du tableau 7, nous pouvons constater que l'usage de l'anaphore n'est pas uniforme (voir également Sachtouri, 2006). Il dépend d'une part, du jeu des synonymes qui s'opère dans chaque langue, et d'autre part, de l'usage que chacune d'entre elles fait des pronoms. À la ligne 24 du document français, le pronom « il » anaphorise son antécédent, le syntagme nominal « une proposition révisée d'un règlement » (l.5). En finnois l'antécédent de la ligne 5 est anaphorisé par le nom commun « asetuksessa », équivalent sémantique en contexte de « règlement » en français. De même, en hongrois et en grec, la reprise anaphorique ne se fait pas de manière pronominale, car ces langues en font souvent l'économie. Le pronom « il » n'a pas graphiquement d'équivalent sémantique, puisqu'il est contenu respectivement dans les verbes « Καθιερώνει » en grec et « Establece » en espagnol.

1.4.4 *Similitude et différence d'ordre au niveau sur-phrastique*

La conservation de l'ordre au niveau sur-phrastique d'un volet d'un multidocument à l'autre ne peut être présupposée. Dans la figure 5, l'ordre est effectivement globalement préservé entre les volets anglais (en) et allemand (de), tandis qu'il est inversé entre ces deux derniers volets et le volet français (fr) comme nous l'observons au travers de la figure 6.

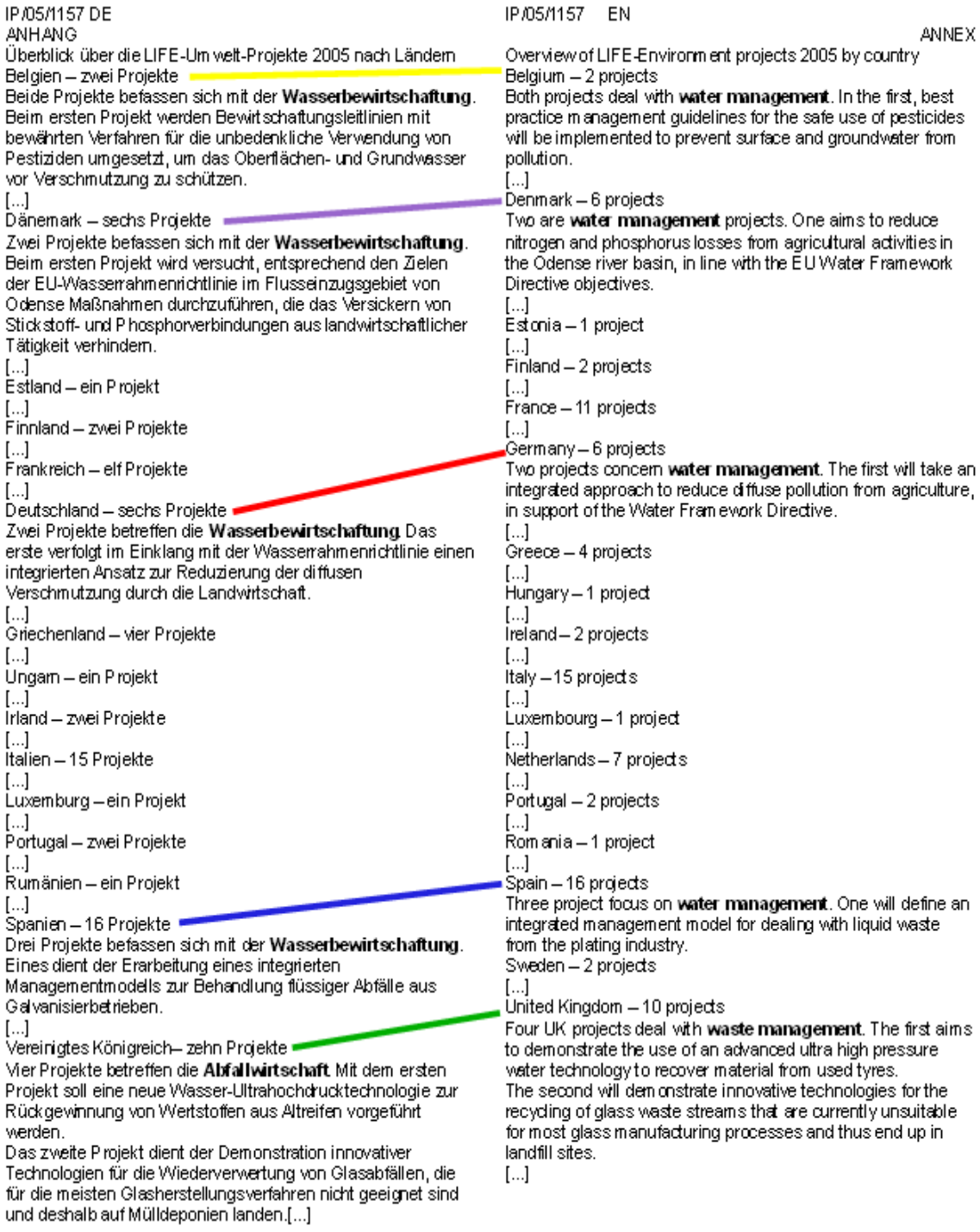
La présence d'une série de paragraphes débutant par le nom du pays concerné par les mesures évoquées et triés par ordre alphabétique de ces noms rend l'ordre largement différent d'un volet à l'autre. On observe un croisement des liens sémantiques. Dans cet exemple, l'inversion concerne des paragraphes, mais il pourrait tout aussi bien s'agir de documents entiers, de résumés…

1.5 CONTRAINTES ÉDITORIALES

La traduction en tant qu'opération est soumise à de nombreuses contraintes éditoriales d'ordre politique, économique, juridique, matériel et linguistique, comme nous l'avons vu précédemment. Le cycle de la traduction à la Commission européenne, tel qu'il est présenté dans le schéma à la page 24 du guide intitulé « Outils d'aide à la traduction et cycle de travail », datant de 2009 et diffusé par la DGT, témoigne lui aussi

LANGUE	FR	FI	HU	EL
Chaîne anaphorique principale ou antécédente	une proposition révisée d'un règlement (l.5)	tarkistetun ehdotuksen asetukseksi	módosított rendelettervezetet	αναθεωρημένη πρόταση κανονισμού
Chaînes anaphoriques secondaires ou de rappel	Ce texte rénové (l.6)	Ehdotus	Ez a felülvizsgált szöveg	Το αναθεωρημένο αυτό κείμενο
	Ce texte rénové (l.10)	Tämän tarkistetun ehdotuksen	Ez a módosított szöveg	Το ανανεωμένο αυτό κείμενο
	Le règlement actuel actuel (l.11)	Nykyinen asetus	A jelenleg érvényben lévőrendelet	Ο ισχύων κανονισμός
	La proposition révisée de règlement (l.13)	Tarkistettu asetusehdotus	A felülvizsgált rendelettervezet	Η αναθεωρημένη πρόταση κανονισμού
	Le règlement (l.23)	Asetuksessa	A rendelet	Ο κανονισμός
	Il (l.24)	Asetuksessa		

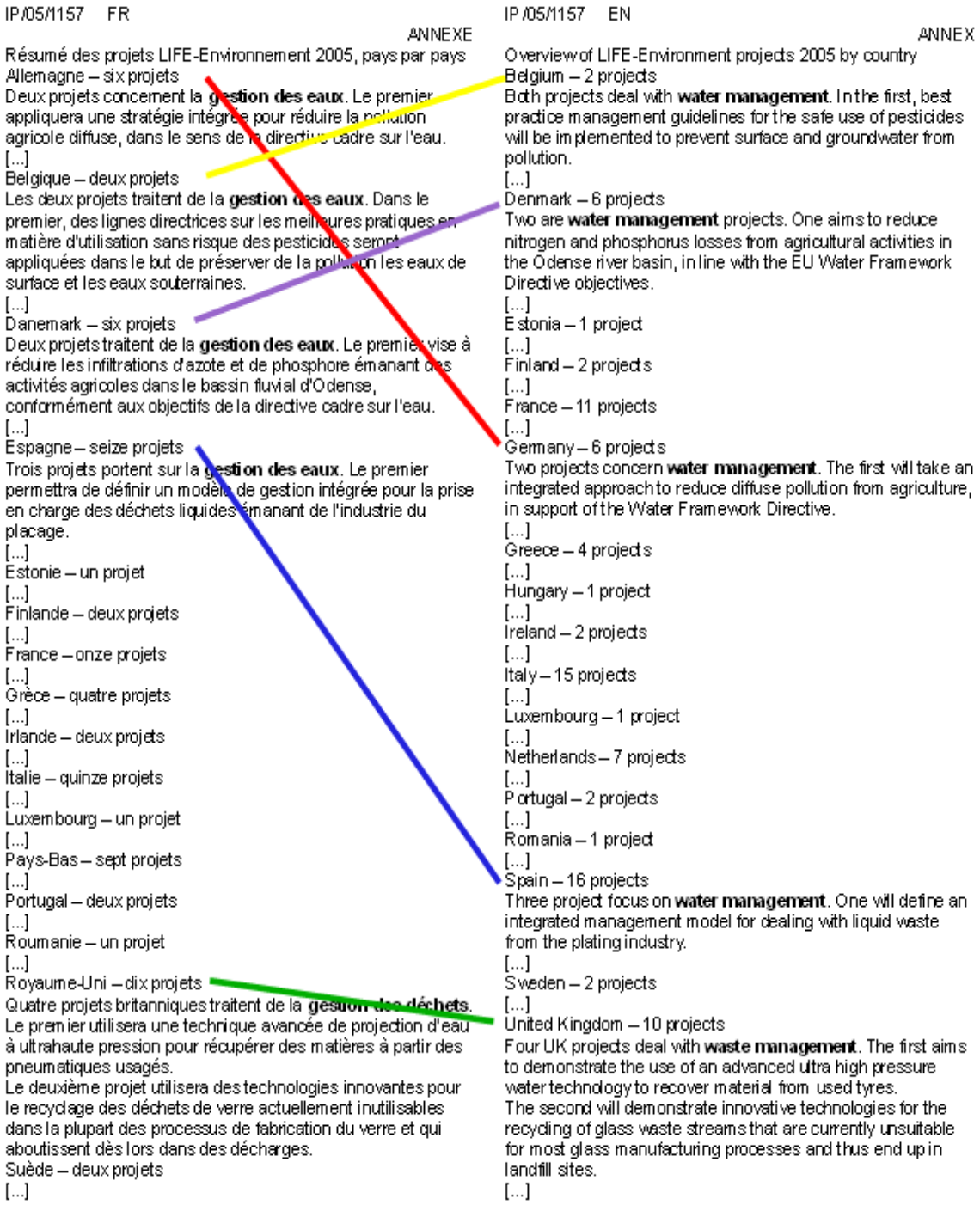
Tableau 7 – Illustration du phénomène d'anaphore dans le multidocument IP/05/975 en français (fr), finnois (fi), hongrois (hu) et grec (el).



IP/05/1157 DE
ANHANG
Überblick über die LIFE-Umwelt-Projekte 2005 nach Ländern
Belgien – zwei Projekte
Beide Projekte befassen sich mit der **Wasserbewirtschaftung**. Beim ersten Projekt werden Bewirtschaftungsleitlinien mit bewährten Verfahren für die unbedenkliche Verwendung von Pestiziden umgesetzt, um das Oberflächen- und Grundwasser vor Verschmutzung zu schützen.
[...]
Dänemark – sechs Projekte
Zwei Projekte befassen sich mit der **Wasserbewirtschaftung**. Beim ersten Projekt wird versucht, entsprechend den Zielen der EU-Wasserrahmenrichtlinie im Flusseinzugsgebiet von Odense Maßnahmen durchzuführen, die das Versickern von Stickstoff- und Phosphorverbindungen aus landwirtschaftlicher Tätigkeit verhindern.
[...]
Estland – ein Projekt
[...]
Finnland – zwei Projekte
[...]
Frankreich – elf Projekte
[...]
Deutschland – sechs Projekte
Zwei Projekte betreffen die **Wasserbewirtschaftung**. Das erste verfolgt im Einklang mit der Wasserrahmenrichtlinie einen integrierten Ansatz zur Reduzierung der diffusen Verschmutzung durch die Landwirtschaft.
[...]
Griechenland – vier Projekte
[...]
Ungarn – ein Projekt
[...]
Irland – zwei Projekte
[...]
Italien – 15 Projekte
[...]
Luxemburg – ein Projekt
[...]
Portugal – zwei Projekte
[...]
Rumänien – ein Projekt
[...]
Spanien – 16 Projekte
Drei Projekte befassen sich mit der **Wasserbewirtschaftung**. Eines dient der Erarbeitung eines integrierten Managementmodells zur Behandlung flüssiger Abfälle aus Galvanisierbetrieben.
[...]
Vereinigtes Königreich– zehn Projekte
Vier Projekte betreffen die **Abfallwirtschaft**. Mit dem ersten Projekt soll eine neue Wasser-Ultrahochdrucktechnologie zur Rückgewinnung von Wertstoffen aus Altreifen vorgeführt werden.
Das zweite Projekt dient der Demonstration innovativer Technologien für die Wiederverwertung von Glasabfällen, die für die meisten Glasherstellungsverfahren nicht geeignet sind und deshalb auf Mülldeponien landen.[...]

IP/05/1157 EN
ANNEX
Overview of LIFE-Environment projects 2005 by country
Belgium – 2 projects
Both projects deal with **water management**. In the first, best practice management guidelines for the safe use of pesticides will be implemented to prevent surface and groundwater from pollution.
[...]
Denmark – 6 projects
Two are **water management** projects. One aims to reduce nitrogen and phosphorus losses from agricultural activities in the Odense river basin, in line with the EU Water Framework Directive objectives.
[...]
Estonia – 1 project
[...]
Finland – 2 projects
[...]
France – 11 projects
[...]
Germany – 6 projects
Two projects concern **water management**. The first will take an integrated approach to reduce diffuse pollution from agriculture, in support of the Water Framework Directive.
[...]
Greece – 4 projects
[...]
Hungary – 1 project
[...]
Ireland – 2 projects
[...]
Italy – 15 projects
[...]
Luxembourg – 1 project
[...]
Netherlands – 7 projects
[...]
Portugal – 2 projects
[...]
Romania – 1 project
[...]
Spain – 16 projects
Three project focus on **water management**. One will define an integrated management model for dealing with liquid waste from the plating industry.
Sweden – 2 projects
[...]
United Kingdom – 10 projects
Four UK projects deal with **waste management**. The first aims to demonstrate the use of an advanced ultra high pressure water technology to recover material from used tyres.
The second will demonstrate innovative technologies for the recycling of glass waste streams that are currently unsuitable for most glass manufacturing processes and thus end up in landfill sites.
[...]

FIGURE 5 – Similitude d'ordre au niveau sur-phrastique entre les annexes des documents anglais et allemand du multidocument IP/05/1157. Les […] ont été introduits par nos soins, ils symbolisent des paragraphes entiers de plusieurs lignes (de 3 à plusieurs dizaines).



IP/05/1157 FR

ANNEXE

Résumé des projets LIFE-Environnement 2005, pays par pays

Allemagne – six projets

Deux projets concernent la **gestion des eaux**. Le premier appliquera une stratégie intégrée pour réduire la pollution agricole diffuse, dans le sens de la directive cadre sur l'eau.

[...]

Belgique – deux projets

Les deux projets traitent de la **gestion des eaux**. Dans le premier, des lignes directrices sur les meilleures pratiques en matière d'utilisation sans risque des pesticides seront appliquées dans le but de préserver de la pollution les eaux de surface et les eaux souterraines.

[...]

Danemark – six projets

Deux projets traitent de la **gestion des eaux**. Le premier vise à réduire les infiltrations d'azote et de phosphore émanant des activités agricoles dans le bassin fluvial d'Odense, conformément aux objectifs de la directive cadre sur l'eau.

[...]

Espagne – seize projets

Trois projets portent sur la **gestion des eaux**. Le premier permettra de définir un modèle de gestion intégrée pour la prise en charge des déchets liquides émanant de l'industrie du placage.

[...]

Estonie – un projet

[...]

Finlande – deux projets

[...]

France – onze projets

[...]

Grèce – quatre projets

[...]

Irlande – deux projets

[...]

Italie – quinze projets

[...]

Luxembourg – un projet

[...]

Pays-Bas – sept projets

[...]

Portugal – deux projets

[...]

Roumanie – un projet

[...]

Royaume-Uni – dix projets

Quatre projets britanniques traitent de la **gestion des déchets**. Le premier utilisera une technique avancée de projection d'eau à ultrahaute pression pour récupérer des matières à partir des pneumatiques usagés.

Le deuxième projet utilisera des technologies innovantes pour le recyclage des déchets de verre actuellement inutilisables dans la plupart des processus de fabrication du verre et qui aboutissent dès lors dans des décharges.

Suède – deux projets

[...]

IP/05/1157 EN

ANNEX

Overview of LIFE-Environment projects 2005 by country

Belgium – 2 projects

Both projects deal with **water management**. In the first, best practice management guidelines for the safe use of pesticides will be implemented to prevent surface and groundwater from pollution.

[...]

Denmark – 6 projects

Two are **water management** projects. One aims to reduce nitrogen and phosphorus losses from agricultural activities in the Odense river basin, in line with the EU Water Framework Directive objectives.

[...]

Estonia – 1 project

[...]

Finland – 2 projects

[...]

France – 11 projects

[...]

Germany – 6 projects

Two projects concern **water management**. The first will take an integrated approach to reduce diffuse pollution from agriculture, in support of the Water Framework Directive.

[...]

Greece – 4 projects

[...]

Hungary – 1 project

[...]

Ireland – 2 projects

[...]

Italy – 15 projects

[...]

Luxembourg – 1 project

[...]

Netherlands – 7 projects

[...]

Portugal – 2 projects

[...]

Romania – 1 project

[...]

Spain – 16 projects

Three project focus on **water management**. One will define an integrated management model for dealing with liquid waste from the plating industry.

[...]

Sweden – 2 projects

[...]

United Kingdom – 10 projects

Four UK projects deal with **waste management**. The first aims to demonstrate the use of an advanced ultra high pressure water technology to recover material from used tyres.

The second will demonstrate innovative technologies for the recycling of glass waste streams that are currently unsuitable for most glass manufacturing processes and thus end up in landfill sites.

[...]

FIGURE 6 – Ordre différent au niveau sur-phrastique entre les annexes des documents anglais et français du multidocument IP/05/1157. Les […] ont été introduits par nos soins, ils symbolisent des paragraphes entiers de plusieurs lignes (de 3 à plusieurs dizaines).

de certaines de ces contraintes, notamment les contraintes matérielles, auxquelles les traducteurs doivent faire face (figure 7).

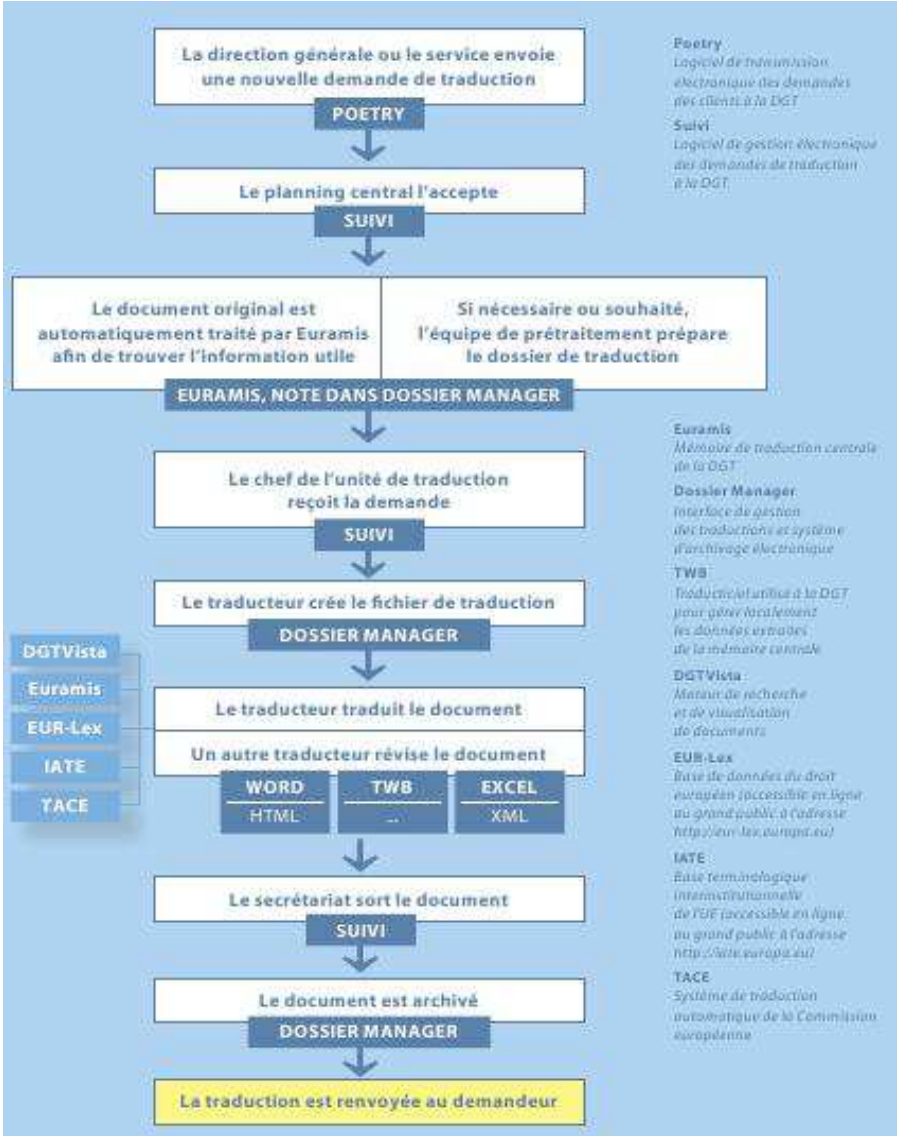


FIGURE 7 – Cycle de la traduction à la Commission européenne.

Les besoins en traduction amènent des contraintes, notamment de rapidité. Les services de traduction doivent répondre au mieux aux demandes de traductions. Néanmoins, les conditions ne sont pas toujours réunies, selon le couple de langues notamment, le système de Traduction Automatique de la Commission Européenne, TACE, ne couvre que 18 couples de langues (et ces couples ne recouvrent que 7 langues), et les traducteurs sur certains couples de langues ne sont pas légion. Ces contraintes structurelles donnent lieu à des choix, tel que celui présenté dans la figure 8 dans lequel la zone de texte commençant par « Next

steps » en anglais n'a été traduite dans aucune des autres langues dans lesquelles ce document est disponible (20 volets en tout).

1.6 CONSTAT : L'ALIGNEMENT AUTOMATIQUE, UN ENJEU DE TAILLE

Nous avons fait ici l'illustration de la complexité de la traduction, tant comme opération que comme produit. Parmi la variété des phénomènes linguistiques, ce chapitre a porté sur ceux concernant les niveaux morphologiques, syntaxiques et textuels. Les conclusions de ce chapitre sont que :

- au niveau sous-phrastique, l'ordre des constituants dépend principalement des langues en présence. Bien que l'ordre des mots n'y soit pas toujours préservé, il arrive néanmoins que dans certains cas il le soit.
- au niveau sur-phrastique, l'ordre du discours dépend principalement de choix du traducteur et bien que l'ordre du texte en langue cible soit généralement le même que celui du texte en langue source, certaines contraintes peuvent engendrer des inversions, des suppressions ou des reformulations.

Nous retenons donc comme observation principale de ce premier chapitre qu'au niveau sur-phrastique comme au niveau sous-phrastique, on ne peut présupposer ni de similitudes, ni de différences d'ordre.

Dans le chapitre 2, nous faisons le tour des différentes approches proposées à ce jour pour traiter ce matériau complexe que sont les traductions : corpus, concepts, indices, approches du point de vue grain analysé/grain aligné. Nous aurons un intérêt particulier pour la façon dont elles prennent en charge les différences et les similitudes d'ordre au niveau sur- et sous-phrastique. Après avoir tiré les constats qui s'imposent nous présenterons dans le chapitre 3 les grandes lignes de notre approche et le corpus sur lequel nous l'avons mise en place, corpus que nous avons voulu d'emblée représentatif de la diversité des langues et des documents.

en	fr
IP/05/473 Brussels, 24 April 2005	**IP/05/473** Bruxelles, le 24 avril 2005
European Commission launches investigations into sharp surge in Chinese textiles imports	**La Commission européenne ouvre des enquêtes sur la brusque hausse des importations de textiles chinois**
Trade Commissioner Peter Mandelson today announced that he has decided to ask the European Commission to authorise him to launch investigations into nine categories of Chinese textile exports to the EU. [...]	***M. Peter Mandelson, commissaire responsable du commerce, a annoncé ce jour qu'il avait décidé de demander à la Commission européenne l'autorisation de lancer des enquêtes concernant les exportations chinoises de neuf catégories de produits textiles à destination de l'Union européenne. [...]***
Peter Mandelson said: "Member States have finally made available the import statistics for the first quarter of 2005. [...]	Peter Mandelson a déclaré: «Nous venons de recevoir les statistiques d'importation des États membres pour le premier trimestre 2005. [...]
The product categories to be covered by the investigation are: T-shirts, pullovers, blouses, stockings and socks, men's trousers, women's overcoats, brassieres, flax or ramie yarn and woven fabrics flax. [...]	Les catégories de produits couvertes par l'enquête sont: les T-shirts, les pull-overs, les chemisiers, les bas et les chaussettes, les pantalons pour hommes, les manteaux pour femmes, les soutiens-gorge, les fils de lin ou de ramie et les tissus de lin. [...]
The product categories concerned cover 7 of the 12 product categories identified by the European textile manufacturers association Euratex in a letter to the Commission on 9 March 2005. [...]	Les catégories en cause couvrent sept des douze catégories recensées par Euratex, l'association européenne des fabricants de produits textiles, dans la lettre qu'elle a adressée à la Commission le 9 mars 2005. [...]
The Textile Specific Safeguard Clause in China's WTO Accession Protocol (2001) [...]	La clause spécifique de sauvegarde relative aux produits textiles du protocole d'adhésion de la Chine à l'OMC (2001) [...]
Next Steps	
These investigations will last for a maximum of 60 days, of which the first 21 will be used to take submissions from parties. [...]	
The Commission reserves the right, should massive and imminent damage to European textile producers [...]	
At the end of the investigation, if the Commission determines that serious market disruption has occurred it can [...]	
As set out by the conditions of the Textiles Specific Safeguard Clause, these formal consultations shall last ninety days. [...]	
At no stage of the process is there any automatic advance to the next stage.	
Any possible safeguard measures would take the form of a quantitative import restriction and could be put in place until December 31 of the current year, or for twelve months if the request for formal consultations comes in the last three months of the calendar year.	

FIGURE 8 – Illustration d'un cas de suppression entre le volet anglais du communiqué de presse IP/05/473 et sa traduction en français. Les […] ont été introduits par nos soins, ils symbolisent la fin du paragraphe qui les précède.

2

EXISTANT MÉTHODOLOGIQUE

Ce chapitre est consacré à un tour d'horizon des principales approches de l'état de l'art en matière d'alignement de corpus parallèles. Nous y présentons les multiples définitions du parallélisme, avant de montrer qu'il existe de nombreuses techniques d'alignement, différentes notamment du point de vue des unités de base mises en jeu : phrases parallèles, paragraphes parallèles ou documents parallèles. Nous présentons les corpus, les concepts et les indices qu'elles exploitent. Nous verrons ensuite sur plusieurs d'entre elles l'usage qui en est fait et quel grain elles analysent pour aligner tel ou tel autre grain.

Les constats que nous tirons de ces principales techniques, notamment du point de vue du parallélisme, nous amènent à présenter au chapitre 3 à la fois les grandes lignes de notre approche et notre corpus tant du point de vue des langues que du type de documents.

SOMMAIRE

2.1 CORPUS PARALLÈLES ET DÉFINITIONS DU PARALLÉLISME

2.1.1 *Définitions du parallélisme*

Le terme *parallèle* revêt un sens différent selon les communautés et les dimensions des textes qu'elles étudient.

Le parallélisme stylistique en versification

Jakobson (1963) dans son article intitulé « linguistique et poétique » introduit le terme *parallélisme* pour désigner un phénomène stylistique consistant à souligner la correspondance entre deux parties de l'énoncé (similitude, opposition, complémentarité). Le parallélisme fait appel à différents « procédés permettant de contraster dans la structure d'une image deux ou plusieurs termes qui peuvent être contraires, ou homonymes, ou synonymes, ou presque homonymes ou presque synonymes » (Becquey, 2003b). L'observation montre une grande variété de types d'associations pour lesquels il faut examiner le nombre de termes en parallèle (couplets, triplets, quadruplets… inventaires), leur taille (parallélisme de 1 à x termes), leur distance (de la connexité à l'éloignement), leur composition (chiasmes, échos, canon…) (Becquey, 2003a). Cette définition du parallélisme sert ici à définir un phénomène monolingue du domaine de l'oralité, agissant notamment à travers la syntaxe des énoncés, les lexèmes, les sons de la langue…

kubin int'an utalam ki'ichkelem injajal yúum	ma parole va à mon vrai beau et mystérieux seigneur
kumani tyosa **kpixan**	qui se déplace pour nos **âmes**
kumani tyosa **klu'uma**	qui se déplace pour nos **corps**

Tableau 8 – Illustration du parallélisme en versification sur un couplet en yucatèque d'une prière d'offrande agricole (Becquey, 2003a)

Le tableau 8 illustre un cas de « microparallélisme », rendu par la répétition partielle de vers à vers qui établit des cadres syntagmatiques au sein desquels on trouve également un contraste paradigmatique (symbolisé en gras dans le tableau).

Cette définition renvoie aux deux modes fondamentaux d'arrangement utilisés dans le comportement verbal, la *sélection* et la *combinaison* :

- la sélection : « la sélection entre des termes alternatifs implique la possibilité de substituer l'un des termes à l'autre, équivalent du premier sous un aspect et différent sous un autre. En fait, sélection et substitution sont les deux faces d'une même opération. » Jakobson (1963)

– la combinaison : « tout signe est composé de signes constituants et/ou apparaît en combinaison avec d'autres signes. Cela signifie que toute unité linguistique sert en même temps de contexte à des unités plus simples et/ou trouve son propre contexte dans une unité linguistique plus complexe. D'où il suit que tout assemblage effectif d'unités linguistiques les relie dans une unité supérieure : combinaison et contexture sont les deux faces d'une même opération. » Jakobson (1963)

Ces deux modes d'arrangement s'actualisent d'une façon propre à chaque langue et dépendent de chacun des six facteurs de la communication présentés par Jakobson : un *émetteur* transmet un *message* à un *récepteur* par le biais d'un *canal* (visuel, auditif...) en utilisant un *code* (pictural, linguistique...), le tout dans un *contexte* donné [1].

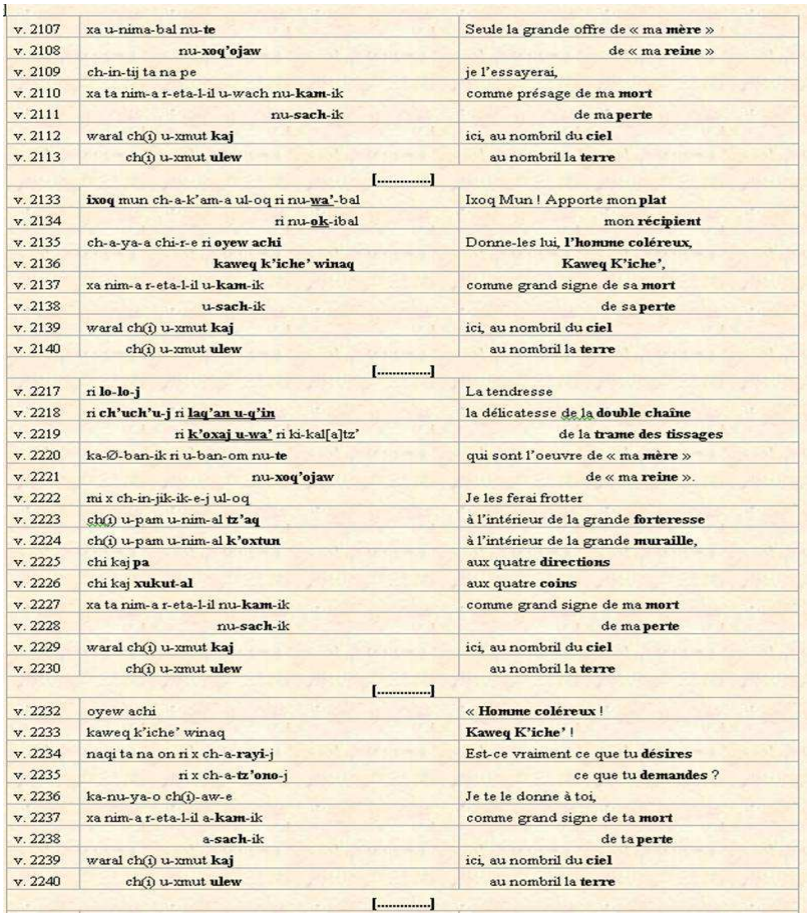
Ainsi, le parallélisme peut être moins littéral que ce que nous avons illustré au travers du tableau 8, il peut aboutir à des niveaux de « macro-parallélisme » intra- (figure 9) voir inter-textuel.

Le parallélisme textuel

Selon Heather et Rossiter (1990), on peut distinguer quatre types de parallélisme textuel en fonction de l'organisation sémantique et structurelle de l'ensemble des données à l'intérieur des documents : explicite, fonctionnel, latent et implicite.

– **Parallélisme explicite** : les deux textes partagent les mêmes identificateurs d'unités textuelles sous forme de clés facilement accessibles par l'ordinateur.
  Exemple : les différentes éditions de la Bible ;
– **Parallélisme fonctionnel** : les deux textes ont, essentiellement, la même structure mais possèdent des identificateurs différents. Une correspondance fonctionnelle peut être établie.
  Exemple : deux versions successives d'un document juridique comportant des différences dans le système de numérotation de sections, paragraphes, phrases, etc. (partial mapping), ainsi que des différences dans le contenu ;
– **Parallélisme latent** : il s'agit de textes qui sont proches dans leurs contenus. Cependant, cette proximité n'est pas manifeste au niveau structurel. Pour mettre en évidence les liens sémantiques qui réunissent l'ensemble de ces textes, il faut entreprendre une réorganisation sémantique ou insérer des identificateurs supplé-

1. Chacun des six facteurs de la communication assure une des six fonctions de base de la communication verbale, respectivement : émotive, poétique, conative (« parce que vous le valez bien ! »), métalinguistique (« cadeaux » prend un « x » au pluriel), phatique (comme le « allô » dit au téléphone), référentielle. « […] si nous distinguons ainsi six aspects fondamentaux dans le langage, il serait difficile de trouver des messages qui ne rempliraient seulement une seule fonction. La diversité des messages réside non dans le monopole de l'une ou l'autre fonction, mais dans les différences de hiérarchies entre celles-ci. » Jakobson (1963)



v. 2107	xa u-nima-bal nu-**te**	Seule la grande offre de « ma **mère** »
v. 2108	nu-**xoq'ojaw**	de « ma **reine** »
v. 2109	ch-in-tij ta na pe	je l'essayerai,
v. 2110	xa ta nim-a r-eta-l-il u-wach nu-**kam**-ik	comme présage de ma **mort**
v. 2111	nu-**sach**-ik	de ma **perte**
v. 2112	waral ch(i) u-xmut **kaj**	ici, au nombril du **ciel**
v. 2113	ch(i) u-xmut **ulew**	au nombril la **terre**
	[.....]	
v. 2133	**ixoq** mun ch-a-k'am-a ul-oq ri nu-**wa'**-bal	Ixoq Mun ! Apporte mon **plat**
v. 2134	ri nu-**ok**-ibal	mon **récipient**
v. 2135	ch-a-ya-a chi-r-e ri **oyew achi**	Donne-les lui, **l'homme coléreux**,
v. 2136	**kaweq k'iche' winaq**	**Kaweq K'iche'**,
v. 2137	xa nim-a r-eta-l-il u-**kam**-ik	comme grand signe de sa **mort**
v. 2138	u-**sach**-ik	de sa **perte**
v. 2139	waral ch(i) u-xmut **kaj**	ici, au nombril du **ciel**
v. 2140	ch(i) u-xmut **ulew**	au nombril la **terre**
	[.....]	
v. 2217	ri **lo-lo-j**	La tendresse
v. 2218	ri **ch'uch'u-j** ri **laq'an u-q'in**	la délicatesse de la **double chaîne**
v. 2219	ri **k'oxaj u-wa'** ri ki-kal[a]tz'	de la **trame des tissages**
v. 2220	ka-Ø-ban-ik ri u-ban-om nu-**te**	qui sont l'oeuvre de « ma **mère** »
v. 2221	nu-**xoq'ojaw**	de « ma **reine** ».
v. 2222	mi x ch-in-jik-ik-e-j ul-oq	Je les ferai frotter
v. 2223	ch(i) u-pam u-nim-al **tz'aq**	à l'intérieur de la grande **forteresse**
v. 2224	ch(i) u-pam u-nim-al **k'oxtun**	à l'intérieur de la grande **muraille**,
v. 2225	chi kaj **pa**	aux quatre **directions**
v. 2226	chi kaj **xukut-al**	aux quatre **coins**
v. 2227	xa ta nim-a r-eta-l-il nu-**kam**-ik	comme grand signe de ma **mort**
v. 2228	nu-**sach**-ik	de ma **perte**
v. 2229	waral ch(i) u-xmut **kaj**	ici, au nombril du **ciel**
v. 2230	ch(i) u-xmut **ulew**	au nombril la **terre**
	[.....]	
v. 2232	oyew achi	« **Homme coléreux** !
v. 2233	kaweq k'iche' winaq	**Kaweq K'iche'** !
v. 2234	naqi ta na on ri x ch-a-**rayi**-j	Est-ce vraiment ce que tu **désires**
v. 2235	ri x ch-a-**tz'ono**-j	ce que tu **demandes** ?
v. 2236	ka-nu-ya-o ch(i)-aw-e	Je te le donne à toi,
v. 2237	xa nim-a r-eta-l-il a-**kam**-ik	comme grand signe de ta **mort**
v. 2238	a-**sach**-ik	de ta **perte**
v. 2239	waral ch(i) u-xmut **kaj**	ici, au nombril du **ciel**
v. 2240	ch(i) u-xmut **ulew**	au nombril la **terre**
	[.....]	

FIGURE 9 – Illustration du macroparallélisme intratextuel. (Becquey, 2003b)

mentaires.
Exemple : plusieurs textes traitant des mêmes thèmes. On parle aussi dans ce cas de corpus comparables ;

– **Parallélisme implicite** : les deux textes sont présentés sous un format qui ne permet pas d'établir des correspondances directes. Néanmoins, il y a suffisamment d'information pour mettre en correspondance les différentes parties de ces textes.
Exemple : deux versions d'un même traité dans deux langues différentes.

Dans les formations de traduction, Hartmann (1980) et Spillner (1981) ont défini les textes parallèles comme étant des documents authentiques, i.e. non traduits, des textes choisis dans le répertoire du texte-cible de la culture, car ils représentent le genre auquel le texte cible devrait appartenir (Nord, 2010). Cette utilisation renvoit à la notion d'intertextualité qui reconnaît dans tout texte la présence d'autres textes, par le biais par exemple de la citation, de l'allusion, du plagiat, de la référence et du lien hypertexte, c'est-à-dire de façon plus ou moins explicite pour le lecteur. Les documents auxiliaires utilisés en traduction

couvrent les trois premiers types de parallélisme vu précédemment : explicite, fonctionnel et latent.

L'École Coseriu de la linguistique contrastive favorise la dernière acceptation du terme parallèle, l'implicite. Elle a utilisé les originaux et leurs traductions comme « textes parallèles » pour l'analyse des sources et l'utilisation de la langue cible, faisant valoir qu'eu égard aux fonctions de communication énoncées par Jakobson, il n'existe pas de textes aussi « parallèles ». La notion de corpus parallèles utilisés dans les études de traduction sur corpus se réfère généralement également à un corpus de textes traduits tandis qu'un corpus de textes non traduits est appelé « corpus comparable » [2].

Dans le domaine du TAL, comme en linguistique contrastive, l'on considère que des corpus parallèles sont constitués d'ensembles de documents composés d'originaux et de leurs traductions. Mais l'idée de parallélisme en TAL va plus loin et opère également dans les dimensions horizontale et verticale des textes. On suppose globalement que la combinaison et la sélection des unités sont réalisées de la même façon d'une langue à l'autre à l'intérieur des documents. Nous revenons sur cette définition du parallélisme en TAL dans la partie consacrée à l'hypothèse de parallélisme ou hypothèse de synchronicité pour limiter les ambiguités (Voir 2.2).

Nous situant dans le domaine de la traduction sur corpus, nous utilisons l'expression corpus parallèles pour désigner un ensemble constitué de textes parallèles, i.e. de documents sources et de plusieurs de leurs traductions. Néanmoins, nous adhérons à l'idée que les fonctions communicatives des textes et de leurs traductions ne sont pas toujours les mêmes et par conséquent que la structure des documents en relation de traduction n'est pas toujours la même, les arrangements de sélection et de combinaison étant propre à chaque langue. Pour ces raisons, à compter du chapitre 3 qui présente notre approche, nous favoriserons l'expression *collection de multidocuments*, dépourvue d'ambiguité et de présupposé quant au parallélisme des documents que nous traitons.

Nous présentons dans la section 2.1.2 les principaux corpus parallèles à disposition, avant de présenter l'hypothèse de parallélisme sous-jacente à la quasi-totalité des méthodes.

2. Néanmoins, l'expression « textes parallèles » continue de prospérer dans les formations de traductions pour désigner des documents non traduits. Deux raisons à cela : d'une part, elle a été utilisée dans la formation des traducteurs (au moins en allemand) bien avant que les études de traduction sur corpus aient émergé et d'autre part, les universitaires spécialisés en traduction ont toujours pris le parti de ne pas considérer les traductions comme une source fiable pour l'étude de l'utilisation du langage, car il n'existe aucune preuve empirique que les fonctions communicatives des textes et leurs traductions sont toujours les mêmes.

2.1.2 *Corpus parallèles*

Le terme *textes parallèles* désigne un ensemble de textes en relation de traduction mutuelle. En fonction des applications visées, ces corpus parallèles correspondent à des corpus de phrases parallèles ou à des corpus de textes parallèles que l'on dira *alignés*[3], si des sous-parties des différents volets sont explicitement mises en relation d'équivalence traductionnelle en phrases ou en paragraphes. Certains d'entres eux ont été partiellement alignés dans le cadre de campagne d'évaluation :

- le Hansard est le premier et le plus connu des corpus parallèles, collecté par l'IBM T.J. Watson Research Center et Bell Communications Research dans les années 80. C'est une sorte d'étalon pour l'évaluation et la mise au point des systèmes. Il s'agit de débats du parlement canadien disponibles en français et anglais. Des parties de ce corpus ont été utilisées notamment par Gale et Church (1993) ou encore Brown *et al.* (1991), avant de servir dans le cadre des deux campagnes d'évaluation ARCADE 1 (Véronis et Langlais, 1999; Véronis, 2000) et ARCADE 2 (Chiao *et al.*, 2006), mais également dans le cadre du projet Portage (Sadat *et al.*, 2006). Malheureusement ce corpus se trouve limité à un seul genre et un seul couple de langues, ce qui ne le rend pas très représentatif ni pour le couple français-anglais ni a fortiori pour les autres couples de langues ;
- Le JRC-ACQUIS Communautaire est disponible en 20 langues [4]. Il comporte environ 800 textes incluant l'ensemble des textes et des traités qui constituent le socle législatif de l'UE. Ce corpus parallèle multilingue a été collecté par l'équipe des technologies du language du centre commun de recherche de le Commission Européenne (JRC) ;
- le European Corpus Initiative de l'International Telecommunications Union CCITT handbook (13,5 M de mots) et l'International Labour Organisation (5M) voient le jour entre 1992-93, ils comportent le français, l'anglais et l'espagnol. Puis entre 1994-95, le projet MULTEXT-MLCC constitue un corpus de questions écrites de parlementaires sur plusieurs sujets (10M de mots) et de débats du parlement européen (environ 60M), disponibles en 9 langues européennes. Ide et Véronis (1994) ont aligné environ 1M de ces mots au niveau des phrases. Erjavec *et al.* (1995) à travers le projet MULTEXT-EAST ont constitué, quant à eux, un corpus de langues de pays européens de l'Est, partiellement alignés en phrases ;

3. Corpus alignés = textes et annotations, métainformation d'équivalences entre des niveau de granularité : paragraphes ou phrases.

4. http://wt.jrc.it/lt/Acquis/

– d'autres projets, tel le Projet JEIDA (Isahara et Hiruno, 2000), ont visé la constitution de corpus parallèles pour les langues asiatiques.

Dans les textes juridico-administratifs, l'alignement de phrases est très souvent de type (1 : 1), d'où l'élargissement à d'autres types de textes dans le cadre de la campagne ARCADE 1 : articles scientifiques, manuels techniques, littérature :

– Science : 5 articles, totalisant 50 000 mots par langue ;
– Tech : 1 manuel de documentation technique, 39328 mots anglais, et 46828 mots français ;
– Verne : le roman De la terre à la lune, 40161 mots anglais et 53181 mots français.

Cependant, la plupart des méthodes trouvent leur limite dans la nécessité qu'elles ont de prendre en entrée de leur système des corpus préalablement alignés en phrases. La disponibilité et la variété de tels corpus sont telles que l'objectif de fournir, grâce aux techniques d'alignement, des ressources électroniques en quantité au traducteur ou au terminologue, s'en trouve compromis. Néanmoins des systèmes d'identification automatique de corpus parallèles voient également le jour. C'est le cas de celui proposé par Patry et Langlais (2005) ou encore par Enright et Kondrak (2007) qui utilisent pour l'un quelques connaissances lexicales et pour l'autre des similitudes de répartition.

2.2 MÉTHODES D'ALIGNEMENT ET HYPOTHÈSE DE PARALLÉLISME

2.2.1 *Définition de l'alignement*

L'alignement ou l'appariement recouvre deux aspects : il s'agit de repérer les mots et expressions du texte source et du texte cible, puis de les mettre en correspondance.

Nous considérons pour notre part, comme le propose Kraif (2001), une distinction entre aligner et apparier, entre alignement et appariement. Dans le cas d'un alignement, nous dirons qu'à une occurrence d'une unité correspond une occurence d'une autre unité dans une autre langue, il s'agit d'une correspondance observable en contexte, tandis qu'un appariement est une correspondance sémantique fortement généralisée telle qu'on en trouve dans un dictionnaire.

Concrètement, aligner des mots, ou des unités sémantiquement équivalentes, est donc l'opération consistant à identifier des relations bilingues ou multilingues entre des mots ou des unités, dans des corpus parallèles (i.e traductions), bilingues ou multilingues, autrement dit des bi-textes ou des multidocuments. Cette démarche s'inscrit dans le but de les réutiliser dans le traitement des langues naturelles, comme la lexicographie bilingue (Klavans et Tzoukcrmann, 1990; Langlois, 1996),

la Traduction Automatique (TA), la Traduction Assistée par Ordinateur (TAO), via notamment des Mémoires de Traduction (Planas, 2000) ou des concordanciers bilingues (Huet *et al.*, 2009), ou encore la création de bases de données terminologiques multilingues (Wu, 1994; Lin *et al.*, 2008) et la détection de plagiat (Brixtel *et al.*, 2009).

Langlais (1997) définit un système d'alignement multilingue « idéal », comme : « un processus qui prend en entrée un corpus multilingue ; c'est-à-dire un ensemble de textes traitant d'un même sujet dans des langues différentes (et qui) produit une sortie constituée d'appariements [5] mettant en correspondance les régions (ou segments) qui sont en relation de traduction dans l'ensemble des textes du corpus. Une région est une unité textuelle pouvant relever de différents niveaux comme le chapitre, la division, le paragraphe, la phrase, la proposition, le terme, le mot, ou encore le caractère. »

Nous adhérons à cette définition multilingue et multiéchelle d'un système d'alignement idéal. Cependant nous devons noter que cette définition très générique ne correspond pas à celle utilisée par les différentes approches de l'état de l'art, tant du point de vue des corpus utilisés, il s'agit le plus souvent de phrases, que des unités qu'elle souhaite aligner. Les méthodes existantes tiennent pour vraie une hypothèse de parallélisme ou de synchronicité trop contraignante y compris dans le cadre de corpus parallèles, de documents traductions. Elles présupposent en effet que *tout est là et/ou tout est dans le même ordre*.

2.2.2 *Hypothèse de parallélisme (de synchronicité)*

L'hypothèse de parallélisme est largement exploitée par les systèmes d'alignement qu'ils soient sous- ou sur-phrastiques. Les fonctions d'alignement pour maximiser leur résultat présupposent un parallélisme fort. On suppose que l'ordre des unités textuelles à aligner est le même, ou presque, à tous les volets du corpus et ce d'autant plus que l'on a à faire à des unités supra-phrastiques. Derrière cette hypothèse, on trouve deux présupposés exposés par Langé et Gaussier (1995), celui de *quasi-synchronisation* et celui de *quasi-bijection* définis comme suit :

- quasi-bijection : toute phrase source a en général un correspondant dans le texte cible, et réciproquement. Dans ce sens, Debili et Sammouda (1992) utilisent la notion de *proximité de taille* ;
- quasi-synchronisation ou quasi-monotonie : la séquence des phrases sources doit suivre, à quelques variations locales près, la séquence des phrases cibles correspondantes. Dans ce sens, Debili et Sammouda (1992) utilisent la notion de *proximité de rang*. Cette hypothèse de la conservation de la séquentialité des idées dans le processus de traduction présuppose deux choses : la

5. Appariement est ici pris au sens d'alignement, la distinction de Kraïf n'ayant été introduite qu'en 2001.

première, plus on descend dans l'échelle, plus il y a de désordre et la deuxième la phrase est la plus petite unité dont l'ordre sera presque toujours maintenu.

Si ces présupposés sont vérifiés et que l'hypothèse de parallélisme est pleinement validée, l'alignement peut être illustré comme sur la figure 10).

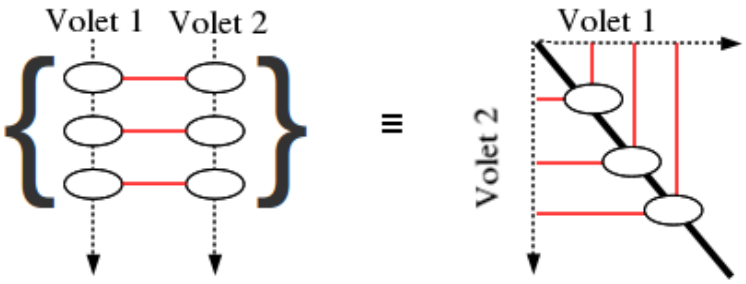


FIGURE 10 – Illustration du parallélisme à l'intérieur d'un bi-texte, composé de deux volets, respectivement en langue 1 et en langue 2.

Le tour d'horizon des méthodes existantes, que nous faisons par la suite, montre que l'alignement de phrases, comme l'alignement sous-phrastique peut être considéré comme résolu lorsque les traductions remplissent effectivement ces hypothèses. Cependant ces contraintes de quasi-synchronisation et de quasi-bijection des unités textuelles ne sont pas toujours vérifiées comme nous l'avons observé et illustré dans le chapitre 1. La traduction n'est pas un simple transcodage, la linéarité du discours n'est pas toujours conservée. Il existe des différences d'ordre tant au niveau sur- que sous-phrastique, et des suppressions massives peuvent intervenir.

Ainsi, bien que l'alignement automatique de traductions soit considéré comme un domaine verrouillé, un problème quasi résolu, et que les recherches s'orientent maintenant vers les corpus comparables, il convient de nuancer ce propos en distinguant notamment les différents types de corpus utilisés. Si l'on peut en effet dire que l'alignement sur- et sous-phrastique sur corpus de phrases parallèles ou de documents parallèles et synchrones est résolu, il n'en est cependant pas de même pour l'alignement sur- et sous-phrastique de textes parallèles asynchrones ou comme on peut les trouver nommés : *complexe*, *bruités (noisy)*, *croisés*, *avec déplacement*. Mais comme le souligne, Church (1993), « Real texts are noisy ». Cette affirmation met à part la traduction « traditionnelle » de roman par exemple, et vise davantage les traductions « tout venant » comme celles publiées sur internet qui pour des raisons de mise en page ou de gains subissent régulièrement des suppressions et/ou des inversions. Il s'agit là de documents quasi-parallèles à mi-chemin entre les

documents parallèles et les documents comparables.

Nous faisons dans la section suivante le tour d'horizon des principales méthodes d'alignement existantes. Nous y exposons pour chaque grain aligné, principalement phrases et mots, les indices et les ressources auxquelles celles-ci font appel, ainsi que l'utilisation qu'elles en font.

2.3 MÉTHODES D'ALIGNEMENT : LA CIRCULARITÉ

Historiquement, les recherches ont d'abord porté sur des méthodes d'alignement de phrases. Mais la quasi-résolution de ce problème, et surtout le constat que l'alignement de phrases est intimement lié à celui des mots (Debili et Sammouda, 1992), et plus généralement aux unités sous-phrastiques, quelles qu'elles soient, ont fait émerger rapidement des méthodes proposant d'aligner aux grains inférieurs à celui de la phrase : mots (Gale et Church, 1991), chunks (Zhou *et al.*, 2004), propositions (Nakamura-Delloye, 2007), …

Debili et Sammouda (1992) décrivent en effet un phénomène de circularité. Les méthodes d'alignement de phrases peuvent utiliser comme point d'ancrage un alignement même partiel de mots. À l'inverse, l'alignement de phrases peut être un point de départ à l'alignement de mots. Dans ce dernier cas, on ne peut se satisfaire d'alignements grossiers. Deux écoles s'affrontent donc : l'une prenant le problème par le haut, par l'alignement de phrases, l'autre choisissant de partir du bas, par l'alignement de mots. Les deux méthodes partagent l'objectif de s'incrire dans un « cercle vertueux ». Mais chacune comporte deux étapes successives et les résultats de la deuxième étape sont toujours dépendants des résultats obtenus par la première.

Les méthodes d'alignement automatique proposées vont du tout statistique (Gale et Church, 1993), à des méthodes hybrides (Langlais, 1997; Moore, 2002) alliant tant des indices de longueurs en mots (Brown *et al.*, 1991) ou en caractères (Gale et Church, 1993) que des indices de fréquences, de distributions (Kay et Röscheisen, 1993; Fung et Church, 1994) ou des indices lexicaux (Church, 1993; Chen, 1993; Simard *et al.*, 1992; Kraif, 1999).

2.3.1 *Méthodes d'alignement de phrases*

Les travaux d'alignement ont d'abord porté sur l'alignement de phrases. L'alignement de phrases consiste à identifier des correspondances entre une phrase dans une langue et d'autres phrases dans d'autres langues. Cette opération précède l'ambition plus grande d'aligner des mots. Elle fait également parfois suite à un alignement de paragraphes

(voire de divisions lorsque le marquage du corpus l'autorise, système LORIA), effectué : manuellement (Gale et Church, 1993), semi-automatiquement, ou automatiquement (Gerdes, 2008). Comme lui, l'alignement de phrases a pour objectif de réduire la combinatoire en vue d'un alignement de mots. Gale et Church (1993) suggèrent même qu'il serait peut-être préférable d'ajouter des étapes d'alignement aux niveaux propositions.

De façon opératoire et non linguistique, la phrase est définie comme un niveau de découpage, délimité par la ponctuation et les majuscules. Une phrase correspond à un segment de texte s'étendant le plus souvent d'un **. Majuscule** à un autre **. Majuscule**. Pour un certain nombre de langues, ce traitement ne réclame pas de ressource. Néanmoins, dans certaines langues, la phrase ne répond pas à ce type de description ou cette description ramène autre chose que des phrases, on fait alors appel à des ressources légères.

Les similitudes de longueur

La méthode d'alignement de phrases sur corpus bilingue de Gale et Church (1993) est statistique et ne se base pas sur le contenu lexical. Ce modèle se base sur l'observation que « des régions de texte plus longues ont tendance à avoir des traductions plus longues, et les régions les plus courtes, des traductions plus courtes », il suggère également que ce rapport est constant. Autrement dit il existe une forte corrélation entre la longueur en caractères d'un paragraphe et la longueur en caractères de sa traduction. Ceci suggère que la longueur en caractères peut être un indice à la fois simple et fort pour l'alignement de phrases. Cette méthode aligne tout à 4% prêt et si l'on sélectionne 80% des alignements ayant le meilleur score, le taux d'erreur passe de 4% à 0,7%. Ce modèle constitue un raffinement de la méthode de Brown *et al.* (1991) basée sur le rapport de longueur en mots entre les phrases. Ces deux systèmes ont prouvé que la longueur en mots et surtout en caractères peut être un indice efficace pour l'alignement de phrases. Ils sont encore largement exploités.

Gale et Church (1993) exploitent un second indice de surface pouvant contribuer à l'alignement de phrase : la fréquence d'apparition de sa configuration. L'analyse d'un corpus déjà aligné en phrases permet en effet de dégager un nombre limité de schémas de correspondances phrastiques (tableau 9), étant entendu que la fréquence des schémas d'appariement dépend grandement du type de textes traités, comme le soulève Langlais (1997).

Les invariants graphiques

D'autres méthodes ont par la suite essayé de conjuguer ces principes en ajoutant et en faisant primer des indices lexicaux, comme par exemple

NOMBRE DE PHRASES			TYPES DE PARALLÉLISME
en L1		**en L2**	
1	⇒	1	Bi-univocité
2	⇒	1	Fusion
1	⇒	2	Scission
2	⇒	2	Bi-univocité multiple
1	⇒	0	Suppression
0	⇒	1	Insertion

TABLEAU 9 – Correspondances phrastiques entre une langue 1 et une langue 2 d'après le modèle de Gale et Church (1993).

la présence de mots comportant des similitudes de surface (Church, 1993; Chen, 1993; Simard *et al.*, 1992; Kraif, 1999). On en distingue deux types :

- les cognats : deux mots d'étymologie commune présentant une similitude de surface que Brown *et al.* (1991) considèrent comme des *ancres faibles* ;
- les transfuges : chaînes de caractères invariantes entre 2 traductions : nombre, noms propres ou emprunts, ponctuation, que Brown *et al.* (1991) considèrent comme des *ancres fortes*.

La recherche de ces invariants repose sur ce que Kraif (1999) appelle l'hypothèse de *cognacité* et qu'il formule de la façon suivante : « la densité de cognats observée entre deux phrases est probablement plus élevée si elles sont traductions l'une de l'autre que si elles sont prises au hasard ». Les méthodes basées sur les cognats s'appuient sur la longueur de la suite maximale de n caractères contigus communs. Certains systèmes (Simard *et al.*, 1992; Church, 1993) en prenant n=4 ont obtenu des résultats significatifs qui, selon Kraif (1999), peuvent être améliorés par un raffinement de cette approximation.

Pour minimiser les ambiguités dues à la notion de ressemblance, il propose donc une définition opératoire des cognats. Ainsi, deux mots (M) sont cognats si et seulement si :

- il existe deux phrases (P1, P2) dont l'une est traduction de l'autre, et dans lesquelles ils sont traductions l'un de l'autre ;
- M1 et M2 présentent un lien étymologique (emprunt, origine commune) perceptible dans leur signifiant, ce à quoi il ajoute les transfuges.

Cependant le premier critère de traductibilité implique des difficultés. D'une part, un mot peut être traduit par un phrasème (« because » ⇔ « à cause »). Kraïf retient alors le couple portant l'étymon commun : « because » ⇔ « cause ». D'autre part il est parfois difficile de déterminer si un mot peut en traduire un autre : la traduction mot-à-mot est un cas limite, éloigné de la pratique effective de la traduction. Kraif (1999)

prend lui le parti restrictif de ne garder que les cognats effectifs du corpus, ceux qui sont effectivement traduits l'un par l'autre et qui de fait peuvent servir à l'alignement de celui-ci.

Dans chacune de ces méthodes, les invariants graphiques, transfuges et cognats, permettent la réduction de l'espace de recherche, la constitution d'un certain nombre de ce que Kraif (1999) appelle des « îlots de confiance » entre les points à aligner. Cette étape de réduction de l'espace à parcourir précède la phase d'alignement à l'intérieur de ces îlots de confiance.

On peut également mentionner ici le système LIA, proche du système Jacal mais moins restrictif, qui fait appel à une étape de pré-traitement basée sur les cognats. Le système propose un alignement en phrase par programmation dynamique pour délimiter un espace de recherche pertinent, en utilisant une fonction de score faisant intervenir de manière pondérée les informations suivantes : longueur des phrases, cognats, dictionnaire de transfert (extrait automatiquement), fréquence des schémas de traduction (`1:1`,`1:2`...).

Néanmoins, si ces similitudes sont fréquentes entre les langues indo-européennes, elles s'avèrent plus rares et insuffisantes entre les langues de différentes familles (indo-européennes et asiatiques par exemple).

Les similitudes de distribution

Kay et Röscheisen (1993) s'inspirent de la technique d'ancrage lexical. Pour cela, ils utilisent d'une part des dictionnaires bilingues et d'autre part ils procèdent à un repérage des cognats grâce au coefficient de Dice. Leur modèle est basé non seulement sur la correspondance phrase/phrase mais aussi mot/mot. Selon Kay et Röscheisen, pour que les phrases d'une langue soient alignées, il faut que les mots de ces phrases soient plus ou moins en correspondance. Même si l'alignement de ces mots est imparfait, c'est un bon point de départ à l'alignement de phrases. Il faut donc comme point de départ trouver des phrases qui fassent office de point d'ancrage aux autres : les meilleures candidates sont les premières et dernières phrases, les plus susceptibles d'être effectivement alignées. La distribution des mots de cet ensemble de deux phrases est pris comme point de départ, on fait l'hypothèse que si ces distributions sont similaires au-delà d'un certain seuil pour un couple de mots donné, ces mots ont de bonnes chances d'être en relation de traduction. Ces mots font office de point d'ancrage, dès que l'on trouve des couples similaires, on aligne, chaque nouveau groupe de mots alignés est un nouveau point d'ancrage jusqu'à la solution optimale.

De la même manière, le système IRMC propose un alignement en phrases s'appuyant sur des liens entre les mots composant ces phrases. Il fait intervenir un dictionnaire de transfert ainsi qu'une mesure de proximité entre mots (Debili et Sammouda, 1992). L'alignement en phrase est

alors réalisé par un algorithme qui recherche la solution qui optimise différents critères comme la conservation de l'ordre des mots dans le processus de traduction ou encore la synchronisation des textes à aligner.

Dans la lignée de ces travaux, Chen (1993) s'appuie sur un lexique construit à la volée, avec lequel il obtient un taux d'erreur de 0,4% sur des données du Hansard.

À ce niveau, on constate que des heuristiques simples basées sur la longueur des phrases en mots (Brown *et al.*, 1991) ou en caractères (Gale et Church, 1993), utilisant éventuellement des points d'ancrage (Brown *et al.*, 1991) ou un lexique construit à la volée (Chen, 1993) ont permis d'atteindre des taux de réussite avoisinant les 100%.

Langlais (1997); Langlais et El-Bèze (1997); Melamed (2000) montrent l'importance de la combinaison de ces différentes sources d'informations.

L'alignement de phrases étant considéré comme résolu, les recherches se sont rapidement tournées vers de l'alignement d'unités sous-phrastiques. Cependant l'alignement de phrases a des limites importantes comme en témoigne la campagne d'évaluation ARCADE 1 (Véronis et Langlais, 1999) révélant les meilleurs résultats sur le corpus JOC : corpus marqué en paragraphes et divisions, pas d'interprétation dans la traduction, schéma le plus généralement (1 : 1) ; et les pires résultats sur le corpus VERNE, il « recueille (…) les plus mauvais résultats. (…) c'est sur ce corpus que les systèmes présentent des performances les plus disparates (de 22% à 90% de précision au niveau des caractères). Ces mauvais résultats s'expliquent par la nature littéraire du corpus, qui contient beaucoup moins d'alignements (1 : 1) que les autres (75% seulement). De plus la version anglaise est abrégée et présente des omissions par rapport à la version française ce qui conduit à des « décrochements » des systèmes. »

Les méthodes sous-phrastiques reposant largement sur l'hypothèse que ce prétraitement est correctement réalisé, subissent des dégradations de résultats lorsque ce n'est pas le cas (cf ARCADE 2 (Chiao *et al.*, 2006) et autres évaluations).

2.3.2 *Méthodes d'alignement sous-phrastique*

Les méthodes d'alignement sous-phrastique prennent, pour la plupart, en entrée, un corpus de phrases préalablement alignées. Leurs résultats sont donc largement dépendant de la qualité de cet alignement de phrases. Nous avons vu précédemment que si l'alignement de phrase pouvait se contenter d'une correspondance mot/mot relativement grossière, il n'en est pas de même pour l'alignement en unités sous-phrastiques. L'alignement d'unités inférieures à la phrase peut

être vu comme un raffinement de la technique d'alignement phrase/-phrase dont le but est d'arriver à une granularité plus petite. La tâche est très complexe car il n'est pas possible d'envisager un alignement fin au niveau lexical sans se pencher sur les nombreuses difficultés que cela engendre :

- les textes sont fortement constitués d'occurrences en rapport complexe : mots composés, locutions, phraséologies, et aucun alignement ou extraction ne peut sérieusement être fait sans prendre en considération ces phénomènes, à la fois recherchés en terminologie et nécessaires pour le travail sur certaines langues comme le suédois ou l'allemand pour n'en citer que deux.
- les textes sont fortement constitués de mots grammaticaux (50% des occurrences d'un texte) dont la traduction est encore moins biunivoque que celle des mots lexicaux.

Deux types d'approches ont émergé certaines purement linguistiques et d'autres hybrides basées sur la combinaison des méthodes statistiques avec les premières et généralement basées sur la reconnaissance de patrons et modèles à l'aide d'expressions régulières ou de grammaires locales. Mais l'introduction de connaissances linguistiques spécifiques à chaque langue est coûteuse et rend les systèmes dépendants des langues.

Deux approches ont été explorées : l'approche estimative et l'approche associative

- l'approche estimative ou par modèles statistiques introduite par (Brown *et al.*, 1990) est inspirée de la traduction automatique statistique, où le calcul d'alignement de mots est la base du calcul des modèles de traduction. Elle commence par déterminer les meilleurs alignements en contexte avant d'en dériver éventuellement des tables de traductions. (Och et Ney, 2003)
- l'approche associative ou par modèles heuristiques introduite par (Gale et Church, 1991). Cette approche descendante utilise la mesure de similarité de chaîne, des heuristiques d'ordre des mots, ou des mesures de co-occurrences telles que le score d'information mutuelle (Fung et Church, 1994)(une paire de mots co-occurre-t-elle plus souvent que par hasard ?), le pourcentage de plus longue sous-séquence commune (Melamed, 1995), le coefficient de Dice (Smadja *et al.*, 1996), des mesures de log-vraissemblance (Tufiş et Barbu, 2002) ou encore le cosinus (Giguet et Luquet, 2006). Les méthodes relevant de cette approche commencent par extraire des traductions avant de créer des alignements.

Ainsi, beaucoup d'études se sont attachées à l'extraction de dictionnaires de mots simples, le plus souvent par des méthodes statistiques (Dagan *et al.*, 1993; Dagan et Church, 1994; Wu et Xia, 1994; Resnik et Melamed, 1997). Très rapidement, les travaux se sont toutefois orientés vers l'extraction d'unités plus longues que le mot graphique : collocations, terminologie et phraséologie (Daille *et al.*, 1994; Gaussier, 1998; Zimina-

Poirot, 2004; Giguet et Apidianaki, 2005; Lardilleux, 2010). Mais peu de travaux s'attachent à l'alignement d'unités plus courtes, il convient de mentionner ici la tentative de Cromières (2006) de réaliser un alignement sous-phrastique par calcul de coefficients de corrélation entre des N-grammes de caractères de taille non prédéfinie. Il conseille particulièrement l'utilisation du grain caractère sur les langues asiatiques, où le mot n'est pas facile à isoler. Pour les langues occidentales, Cromières a également appliqué son algorithme au grain caractère sur un petit corpus de bi-phrases tirées du corpus Europarl, à cause de limites de mémoire.

L'alignement sous-phrastique se heurte immédiatement à la délimitation des unités, notamment lorsque le mot n'est pas physiquement marqué, ou bien lorsque la langue est agglutinante. En outre, on ne peut présumer une quelconque préservation de l'ordre des unités dans la phrase. Pour pallier cette difficulté, le recours à un dictionnaire bilingue est souvent l'option choisie, mais cette technique exclut d'emblée l'analyse des langues faiblement dotées en matière de ressources linguistiques, pose le problème de la qualité de ces dictionnaires et rend l'analyse d'une nouvelle langue coûteuse. Nous noterons également que l'alignement au niveau sous-phrastique suit généralement un alignement phrastique et qu'il est donc largement dépendant de la qualité de celui-ci.

2.4 ALTERNATIVES POUR APPRÉHENDER LA CIRCULARITÉ

Les méthodes présentées dans cette section visent là encore un alignement sous-phrastique mais l'amorcent de façon plus progressive et moins contrainte. L'objectif est de pouvoir traiter aussi bien des documents synchrones qu'asynchrones.

2.4.1 *L'alignement de phrases : une interrogation documentaire*

Fluhr *et al.* (2000) proposent une approche originale affranchie des hypothèses contraignantes précitées, dans laquelle les textes ne sont plus traités séquentiellement mais comme des bases de données qui sont alors considérées comme un système de recherche d'informations : le problème de l'alignement de phrases est ainsi ramené à celui d'une interrogation documentaire multilingue, dont le but est de ramener la phrase la plus similaire dans le texte à partir de la « requête » que constitue la phrase source.

2.4.2 *Méthodes d'alignement sous-phrastique affranchies d'un alignement de phrases*

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

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

Plusieurs auteurs ont utilisé des matrices de points (dotplots, techniques empruntées à l'analyse ADN, et d'abord reprise pour explorer du code source (Church et Helfman, 1993)) les appariements ainsi révélés transformant le problème de l'alignement en un problème de traitement d'image Church (1993); Chang et Chen (1997); Langlais (1997); Melamed (1999) ou exploitant des hypothèses similaires pour la détection de plagiat (Brixtel *et al.*, 2010).

2.4.3 *Utilisation des structures hiérarchiques des documents*

Brixtel (2011) met, quant à lui, en évidence le fait que les marques de structure et de mise en forme des documents peuvent servir à délimiter des zones de recherche pour les alignements phrastiques et sous phrastiques. Ses expériences ont été réalisées sur des documents extraits du site Europa, présentés sous la norme XHTML. De nombreuses traces non textuelles comme des liens hypertextes, des tableaux, les séparations horizontales ou l'application de gras ou d'italique, peuvent y être repérés via la Mise en Forme Matérielle (MFM). Brixtel soutient que la MFM peut être considérée comme un vecteur de sens préservé dans le processus de traduction, cela le conduit à exploiter ces marques en tant qu'invariant entre les documents de différentes langues pour identifier leur structure. « L'idée est d'exploiter une hiérarchie des constituants la plus fine possible pour s'assurer de la construction d'un espace de recherche à un niveau de résolution le plus bas possible en passant par des paliers fiables ». Ainsi, cette segmentation-alignement au grain alinéa, plus élevé que la phrase, permet de restreindre les espaces de recherche d'équivalences sémantiques entre les documents d'un multidocument et d'identifier des suppressions. Ces macro-alignements posent les bases de la détection d'appariements sous-phrastiques à laquelle il procède par la suite.

2.5 CONSTATS : MÉTHODES D'ALIGNEMENT EXISTANTES ET APPLICATIONS

Le problème de l'alignement est par définition celui de la localisation et de la délimitation précise des segments à mettre en correspondance entre les langues.

Si les différentes méthodes d'alignement au grain paragraphe ou phrase ont fait leurs preuves sur certains types de documents, il est néanmoins à noter qu'elles reposent sur des hypothèses simplificatrices à propos du parallélisme de la structure des documents :

- l'ordre des phrases dans les deux textes est identique ou très proche ;
- les textes contiennent peu de suppressions ou d'adjonctions ;
- les alignements (1:1) (de longueur équivalente) sont très largement prépondérants et les rares alignements $m : n$ sont limités à de petites valeurs de m et n (typiquement 2).

Elles ne sont par conséquent que très peu tolérantes aux variations dispositionnelles du contenu. Les résultats des analyseurs basés sur ces hypothèses se dégradent lorsqu'elles ne sont pas vérifiées dans le corpus. La qualité des alignements est globalement fonction du corpus, satisfaisants sur des textes juridiques et techniques (textes « simples », où les schémas (1:1) mot et phrase sont les plus courants), médiocres sur des textes scientifiques, témoignant davantage d'un travail de traduction-

réécriture que d'un travail de traduction-transcodage et elle se dégrade encore à mesure que l'on tend vers des textes littéraires.

Des alternatives ont été proposées pour appréhender différemment la circularité et dépasser le problème de l'alignement de phrases, mais celles-ci ne réglent pas les questions fondamentales liées à la présence d'inversions, de suppressions ou de reformulations massives.

En outre, il faut signaler qu'à de rares exceptions près (Simard, 1999; Lardilleux, 2009), ces méthodes sont bilingues et que peu d'entre elles sont endogènes, c'est-à-dire ne requièrentt aucune ressource dictionnairique (Giguet, 2005; Giguet et Luquet, 2006; Brixtel, 2011).

Ainsi, la question qui demeure est de savoir comment parvenir à aligner massivement de façon peu supervisée et donc peu coûteuse, des documents traduits, y compris de façon asynchrone, présentant des cas d'inversions, mais aussi de suppressions/omissions. Une des pistes que nous privilégions est celle d'un travail sur les caractères amorcé par Cromières, voie prometteuse pour un alignement indépendant des langues.

Nous présentons dans le chapitre 3 les principales caractéristiques de notre méthode d'alignement de documents multilingues sans présupposé de parallélisme. Cette présentation générale ouvre la voie à la présentation plus détaillée qui se tient dans la deuxième partie de notre rapport.

3

POUR UNE MÉTHODE SANS PRÉSUPPOSÉ DE PARALLÉLISME SOUS- OU SUR-PHRASTIQUE

Nous nous sommes intéressée aux limites rencontrées par les méthodes reposant sur l'hypothèse du parallélisme sur-phrastique. Au regard de celles-ci, l'enjeu de notre méthode est de mettre en place une méthode affranchie des contraintes liées tant à la disponibilité de corpus préparés ou sélectionnés pour leur parallélisme sur-phrastique avéré, qu'à celles de ressources dictionnairiques ou à la proximité des langues.

Nous présentons dans ce chapitre les principales caractéristiques de notre méthode ainsi que le corpus de langues et de documents que nous avons délibérément choisi pour sa variété morphologique afin de pouvoir directement éprouver notre méthode sur des données représentatives.

SOMMAIRE

3.1 CARACTÉRISTIQUES GÉNÉRALES DE NOTRE APPROCHE

Le principal objectif de notre méthode est de prendre en charge les cas de suppressions/omissions d'une partie d'un des deux volets d'un bi-document (cf. figure 8), aussi bien que les cas d'inversions (cf. figure 6). Pour cela, nous choisissons de nous affranchir d'un alignement préalable au grain phrase (Church, 1993) et d'au contraire traiter les documents dans leur intégralité et avec leur MFM (Brixtel, 2011; Resnik et Smith, 2003).

Notre deuxième objectif est de mettre au point une méthode adaptée à toutes les langues : indépendante de l'ordre des constituants de la phrase et de la disparité du grain mot. Pour cela, nous traitons toutes les langues avec des chaînes de caractères comme le propose (Cromières, 2006) pour les langues asiatiques.

Enfin, dans l'esprit des travaux de l'équipe DLU du laboratoire GREYC, nous souhaitons élaborer une méthode endogène qui exploite le corpus pour analyser le corpus autrement dit qui n'utilise que les connaissances intrinsèquement contenues dans les traductions. Notre objectif est de pallier ainsi le manque voire l'absence de ressource dictionnairique disponible pour l'analyse de certaines langues, ainsi que le coût de l'ajout éventuel d'une langue dans le corpus.

3.2 CORPUS DE LANGUES MORPHOLOGIQUEMENT DIFFÉRENTES

Nous introduisons volontairement dès le début des langues très différentes du point de vue du foisonnement, de l'alphabet, de la morphologie… Ces différences nous aideront à valider et renforcer l'intérêt de certains concepts à la base de notre méthode appliquée à une collection de documents, comme l'alignement de N-grammes de caractères ou le concept de multizones, ainsi que le caractère indépendant des langues que revêt la méthode dans son ensemble.

Tous les schémas SVO et déterminé-déterminant sont représentés, au travers de deux couples de langues proches et plusieurs couples de langues différentes selon plusieurs aspects : plus ou moins agglutinant, plus ou moins flexionnel.

3.2.1 *Langues indo-européennes*

Langues romanes

Dans ce groupe linguistique, composé de l'espagnol, du français, de l'italien, du portugais et du roumain, nous avons conservé le français et l'espagnol :

– le français, car c'est notre langue maternelle, mais également du fait de son importance dans la traduction. Le français est

souvent, pour autant que nous le sachions car cette information n'est jamais mentionnée, la langue du document source de nos multidocuments issus de la Commission Européenne (voir la section 3.3).
– l'espagnol, car c'est l'une des deux langues, avec le grec, les moins synthétiques des langues de l'Union Européenne à l'exception du roumain, du bulgare et du gaélique. En outre, nous avions des connaissances de cette langue, préalables à cette étude.

Langues germaniques

Dans ce groupe subdivisé en 2 sous-groupes appelés : langues germaniques occidentales et langues scandinaves, se situent d'une part l'allemand, l'anglais, et le néerlandais et d'autre part, le danois et le suédois. Nous avons choisi l'anglais, l'allemand et le danois :
– l'anglais, car au même titre que le français, il correspond souvent à la langue du document source de nos multidocuments, et également pour nos connaissances de cette langue ;
– l'allemand pour sa syntaxe particulière ;
– le danois pour sa proximité avec l'allemand, amenant à deux le nombre de couples proches avec celui composé par le français et l'espagnol.

Langue hellénique

Le grec est seul dans ce groupe. C'est avec l'espagnol, une des deux langues les moins synthétiques. C'est également une des langues de l'Union Européenne qui s'écrit avec un alphabet différent.

3.2.2 *Langues ouraliennes*

Langues finno-ougriennes

Ce groupe linguistique est lui aussi subdivisé en 2 sous-groupes, langues fenniques et langue ougrienne, composés pour l'un du finnois et de l'estonien, et pour le second du hongrois.

Notre choix s'est porté sur le finnois pour son caractère très synthétique.

Nous faisons le choix de ne pas nous intéresser plus en profondeur aux langues slaves occidentales et méridionales (le polonais, le slovaque et le tchèque et de l'autre, le slovène et le bulgare), ni aux langues baltes, groupe linguistique composé du letton et du lituanien.
Un tel corpus de langues nous amène notamment à nous interroger sur le statut du mot dans chacune de ces langues. Et à proposer une délimitation adaptée des unités à aligner.

3.3 CORPUS DE DOCUMENTS EN RELATION DE TRADUCTION

Nos expérimentations ont été menées sur un corpus est constitué de communiqués de presse de l'Union Européenne. Il s'agit de communiqués de presse au format HTML et encodé en utf-8, émanant de la Commission Européenne et disponibles sur le site Europa, le portail de l'Union européenne [1], source importante de documents traduits jusque dans 23 langues [2]. Les documents que nous observons sont considérés a priori comme traductions pour la simple raison qu'ils sont présents sur le même site et portent le même nom. Nous choisissons ce corpus car nous avons déjà pu observer qu'il contient des inversions sur- et sous-phrastiques, ainsi que des suppressions plus ou moins massives (cf. figure 8) que nous cherchons à découvrir automatiquement par notre méthode. Nous ne réalisons pas de prétraitement sur ce corpus et le traitons directement avec son source en HTML.

De ce corpus de communiqués, nous avons extrait les documents disponibles dans les sept langues que nous avons annoncé vouloir traiter dans la section 3.2. Chaque document source et ses traductions ont été placés dans un dossier numéroté constituant ainsi un multidocument. De cette façon, nous avons isolé 385 multidocuments. Nous ferons une synthèse des résultats obtenus sur 194 de ces multidocuments ventilés sur 6 collections différentes dans le chapitre 7. Les raisons sous-jacentes à la constitution de collections sont détaillées dans le chapitre 4 et la nature des collections utilisées pour l'évaluation est présenté au chapitre 7.

Dans ce chapitre, nous avons présenté les grandes lignes de notre approche sans présupposé de parallélisme entre les volets d'un multidocument ainsi que le corpus que nous souhaitons analyser. Ce corpus se veut réel, empreint de diversités linguistiques et de la marque du travail de réécriture que constitue la traduction.

1. http://europa.eu
2. Nous le mettons à la disposition de la communauté : http://code.google.com/p/europa-corpus/

Deuxième partie

MÉTHODE D'ALIGNEMENT SANS PRÉSUPPOSÉ DE PARALLÉLISME

4

NOS CONCEPTS

Nous présentons ici les concepts utilisés pour définir en contexte si les documents que nous cherchons à aligner sont effectivement traductions, si oui dans quelles mesures et pour révéler les unités qui sont effectivement en correspondance. Notre approche est résolument orientée analyse textuelle en cela qu'elle s'applique à des multidocuments dans leur intégralité. Nous utilisons les N-grammes de caractères, les collections de multidocuments et la Mise en Forme Matérielle (MFM) pour leur capacité à révéler de la répétition. Enfin, dans un but opératoire, nous introduisons le concept de *multizone*.

SOMMAIRE

4.1 LE MULTIDOCUMENT

Les systèmes d'alignement sous-phrastique prennent généralement en entrée un corpus de documents parallèles préalablement alignés en phrases ou un ensemble de phrases parallèles. Notre méthode orientée analyse textuelle prend en entrée des multidocuments. Comme nous l'avons mentionné dans le chapitre 1, le néologisme *multidocument* a été crée au laboratoire du GREYC. Il inclut, en tant que grain supérieur au multitexte, les dimensions de mise en forme matérielle et de structure de documents. Si le document est l'unité la plus apte à rendre compte des résultats de l'acte de langage, le multidocument est le plus intéressant pour étudier l'opération de réécriture qu'est la traduction et les phénomènes auxquels elle donne lieu : choix des mots mais également ce que nous souhaitons étudier dans nos travaux : l'inversion, la suppression… À la différence de la phrase, le document présente une autonomie permettant de travailler sur des répartitions autres que des répétitions à l'identique.

4.2 LA COLLECTION DE MULTIDOCUMENTS

La collection nous sert de cadre pour étudier les distributions des éléments contenus dans chacun des multidocuments : lexique et structure. Elle nous permet d'augmenter les informations sur le contenu de chacun des multidocuments de la collection et notamment de :

- trouver d'autres occurrences d'unités hapax dans un document à analyser : dans un document pris isolément, l'on dénombre un grand nombre de mot hapax, et ce d'autant plus que la langue est morphologiquement riche. De par le volume qu'ils représentent, ces hapax de document sont difficiles à aligner a fortiori si l'on décide de ne pas présupposer le parallélisme (la synchronicité) entre deux volets d'un multidocument, c'est-à-dire de ne pas considérer leur position à l'intérieur des volets.
- révéler simplement à partir de leurs distributions intra- et inter-langue et sans traitement spécial des éléments de structures présents dans les différents volets des multidocuments. En multilingue, une chaîne de caractères largement ventilée sur les différents volets et les différentes langues a de fortes chances de correspondre à un élément de structure.

Ces informations seront autant d'indices supplémentaires pour mettre en évidence des différences et des similitudes entre les volets des multidocuments et les unités qui les composent.

Ces collections ont comme caractéristiques principales de :

- regrouper plusieurs multidocuments ;
- être équilibrées du point de vue des langues, autant de documents pour chaque langue afin de limiter les décalages de fréquences, déjà forcément présents d'une langue à une autre ;

– être éventuellement thématiquement homogènes afin de maximiser l'apparition de répétitions intermultidocument.

4.3 LE DOCUMENT ET SA MISE EN FORME

Dans la lignée des travaux de Brixtel (2011), nous considérons que la mise en forme est porteuse de sens et doit de ce fait être utilisée pour l'alignement de multidocuments. Cependant, à la différence de celui-ci qui recherche et interprète les indices de forme, nous choisissons de prendre en compte la structure et le contenu par la même méthode, sans leur accorder un traitement particulier. Prendre les documents avec le source permet une fois encore de faire ressortir des éléments répétés, pour le coup pas forcément intéressants dans l'optique de constitution de lexiques multilingues mais précieux dans la masse d'informations susceptible d'être alignée pour identifier les cas particuliers que nous souhaitons prendre en charge :

```
<document celex="IP-08-2065" lang="fr">
<h1> <a name="Heading4">
<p align="right">
</document>
```

TABLEAU 10 – Indices de forme dans le source HTML

La mise en correspondance de ces chaînes de caractères ne va pas de soi, elle est autant sujette à variation que l'usage d'un mot ou d'un de ses synonymes. Néanmoins l'appariement de ces unités constitue autant d'indices supplémentaires pour déterminer sans ressource extérieure si les documents contiennent des inversions et/ou des supressions, autrement dit pour ancrer notre alignement de zones.

4.4 LES CHAÎNES DE CARACTÈRES RÉPÉTÉES DE LONGUEUR MAXIMALE

Notre travail se situe dans la lignée de ceux de Cromières, nous procédons à une recherche de n-grammes de caractères en contexte, indépendamment de leur taille. Si l'on peut opposer à cette unité d'information un manque d'ergonomie interprétative, celle-ci présente néanmoins plusieurs avantages :

– elle permet de capturer par le même mécanisme : des expressions figées, des racines de mots, des indices de formes ;
– elle est indépendante de la langue, elle permet donc de couvrir un large éventail de langues sans module spécifique ;
– statistiquement comparables, elle permet de calculer des fréquences d'apparition et d'estimer leur distribution et la régularité avec

laquelle plusieurs unités co-occurentes dans les mêmes parties du texte ;
– elle est facile à repérer sur le plan informatique.

La notion de N-grammes de caractères est déjà utilisée pour l'identification d'auteurs (Jardino, 2006), l'identification de la langue (Dunning, 1994), l'analyse de l'oral, la catégorisation de textes (Damashek, 1995), la classification numérique multilingue de documents (Biskri et Delisle, 2001) ou encore la recherche d'informations (Majumder *et al.*, 2002; Mcnamee et Mayfield, 2004). Cependant, à notre connaissance, il n'existe qu'une tentative de Cromières (2006) pour appliquer une telle méthode à l'alignement multilingue. Cromières réalise un alignement sous-phrastique par calcul de coefficients de corrélation entre des N-grammes de caractères. Si, dans les applications de TAL évoquées ci-dessus, les n-grammes de caractères ont un nombre de caractères constants défini a priori, ce sont généralement des bi-grammes ou des tri-grammes de caractères (4-grammes ou 5-grammes dans le cas de Mcnamee et Mayfield (2004)), chez Cromières leur taille n'est pas prédéfinie.

Les systèmes d'alignement et d'extraction d'information au sens large passent généralement par une segmentation en mots. Mais la question du statut du mot se pose.

En TAL, le mot est généralement décrit comme un segment de discours compris entre deux espaces et/ou ponctuation. Or ce mot graphique, au travers des langues, recouvre des réalités très diverses d'un point de vue sémantique. En outre, certains systèmes d'écriture ne marquent pas les frontières du mot par des espaces, c'est le cas notamment en chinois.

Le concept de mot est donc complexe. Son statut dépend en fait du point de vue adopté : lexical ou graphique. Ces deux points de vue ne sont pas toujours en correspondance (cf. tableau 1).

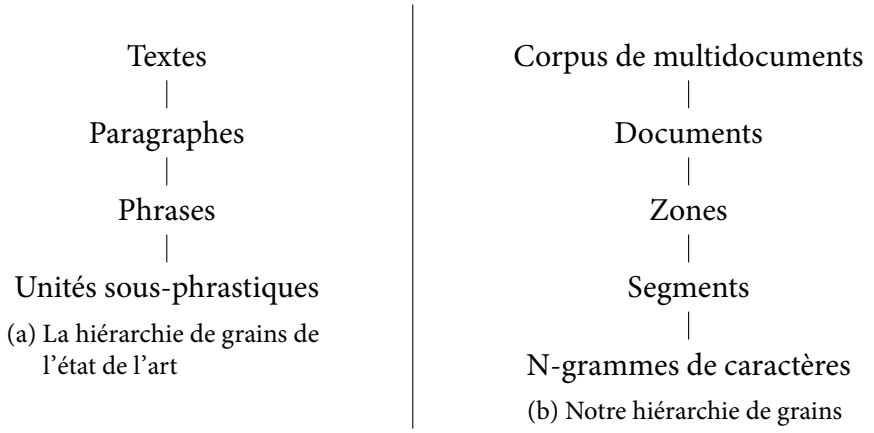
Cette question est d'autant plus complexe que l'on a à traiter des *mots polylexicaux* (ou complexes) à savoir « toute unité composée de deux mots simples ou mots dérivés préexistants [...] les mots polylexicaux (ou complexes) peuvent être soudés (et alors, du point de vue informatique, ils peuvent être assimilés à des mots simples) [...] ou comporter un séparateur »[1]. La forme graphique d'une unité lexicale composée tient de propriétés intralangues. Elle dépend des particularités morphologiques de flexions et de dérivations de chaque langue.

Au regard de ces caractéristiques morphologiques, le mot graphique n'apparaît pas suffisamment universel pour répondre au besoin de comparativité d'un système multilingue d'alignement et d'extraction d'information et qui plus est sans ressource. À cause des variations flexionnelles, nous nous fions aux chaînes de caractères plus qu'aux mots. Ce qui, pour l'humain correspond au même sens, se calcule davan-

1. G. Gross (1996) cité par Neveu (2004)

tage en terme de même forme pour la machine. Ainsi, nous prévoyons un découpage en contexte de N-grammes de caractères [2] pour faire émerger des correspondances que ne révèle pas un découpage en mots.

4.5 LES MULTIZONES



(a) La hiérarchie de grains de l'état de l'art

(b) Notre hiérarchie de grains

FIGURE 11 – Hiérarchie de grains

En corrélation avec le parallélisme présumé de la narration, l'état de l'art de l'alignement à gros grain s'appuie sur une délimitation forte des paragraphes (via la mise en page) et faible des phrases (via la ponctuation) (figure 11a).

Or nous l'avons vu, la phrase comme le mot, peut recouvrir une réalité sémantique différente d'une langue à l'autre. L'opération traduisante, réalisée par l'humain et visant à interpréter le sens d'un document donné dans une langue source et à produire un document sémantiquement équivalent dans une ou plusieurs langues cibles, peut donner lieu à des modifications dans l'organisation interne des différents volets. Cette possibilité intervient tant au niveau microscopique qu'au niveau macroscopique. Les figures 5 et 6 présentent deux cas de traductions différents du point de vue de l'ordre macroscopique, co-présents dans un même multidocument disponible en trois langues, français, anglais et allemand, repris dans la figure 12 (page 58) présentant de façon simplifiée les multidocuments des figures 5 et 6 :

À DROITE, l'alignement entre les volets allemand et anglais montre le cas d'un maintien de l'ordre ;

À GAUCHE, le cas d'inversions massives de plusieurs zones de textes entre le volet français et le volet anglais (et par conséquent allemand) du même multidocument.

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

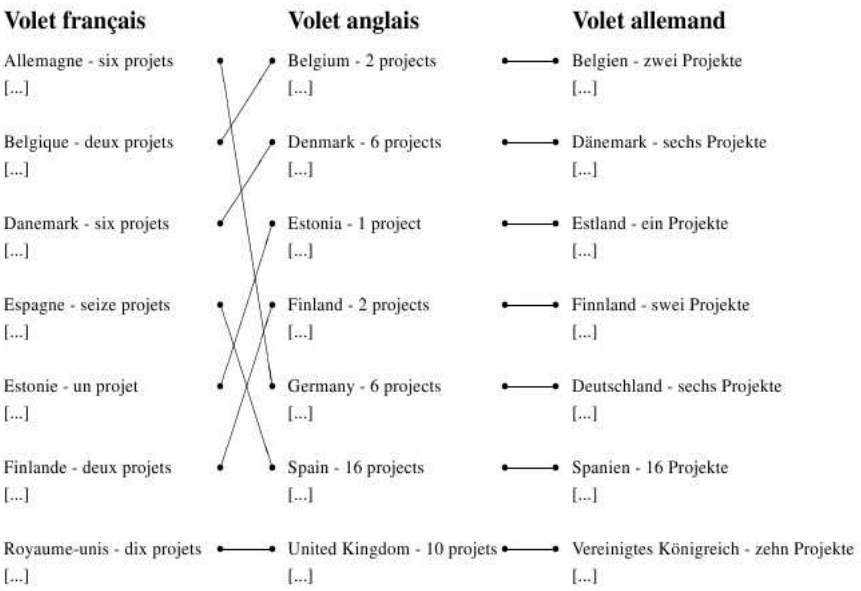


FIGURE 12 – Maintien de l'ordre et inversions entre les différents volets d'un multidocument (communiqué de presse IP/05/1157 de l'Union Européenne) en anglais, français et allemand contenant des paragraphes triés par ordre alphabétique. Nous utilisons les [...] pour symboliser le contenu d'un paragraphe, dont nous ne conservons ici que le début soit le nom du pays dont il traite.

Ainsi, dans le premier cas, selon notre hiérarchie de grains présentée dans la figure 11b, nous considérons qu'il existe deux zones parallèles (une *bi-zone*), c'est-à-dire traduites de façon globalement littérale, correspondant dans chaque langue aux documents dans son ensemble. Tandis que dans le deuxième cas, nous considérons qu'il existe plusieurs zones entre lesquelles il existe un parallélisme, plusieurs bi-zones. On dit de ces traductions qu'elles sont *asynchrones*. L'ordre macroscopique n'est pas systématiquement maintenu d'un volet à un autre, ce type d'inversion apparaît par exemple lorsqu'un résumé présent au début d'un volet est traduit à la fin d'un ou de plusieurs autres, quand une suppression de zone de textes intervient ou que les paragraphes sont triés par ordre alphabétique (figure 13). Ceci constitue un obstacle majeur aux méthodes d'alignement qui reposent sur une hypothèse de parallélisme et qui traitent comme objet de départ des documents traduits dans leur intégralité. Nos travaux s'orientent vers la délimitation automatique de ce grain intermédiaire, entre le document et les unités sous-phrastiques, grain défini en contexte dans un traitement bilingue et non de façon ad hoc. Cette bi-zone est constituée de deux zones, une dans chaque langue, elles-mêmes constituées de caractères pouvant recouvrir plusieurs réalités en contexte : du document à la chaîne de caractères en passant par le paragraphe, la phrase, la proposition, l'expression ou le

mot. Une *bi-zone* est donc le résultat de la mise en correspondance de deux zones de textes de deux langues différentes. Deux zones seront alignées si elles révèlent un maximum de liens, autrement dit si elles maximisent le parallélisme.

Nous avons présenté dans ce chapitre les concepts originaux à la base de notre méthode résolument orientée analyse textuelle : le multidocument, la collection de multidocument, le document et sa mise en forme matérielle, les chaînes de caractères répétées de longueur maximale et les multizones. Nous présentons dans le chapitre 5 l'exploitation que nous en faisons dans notre méthode.

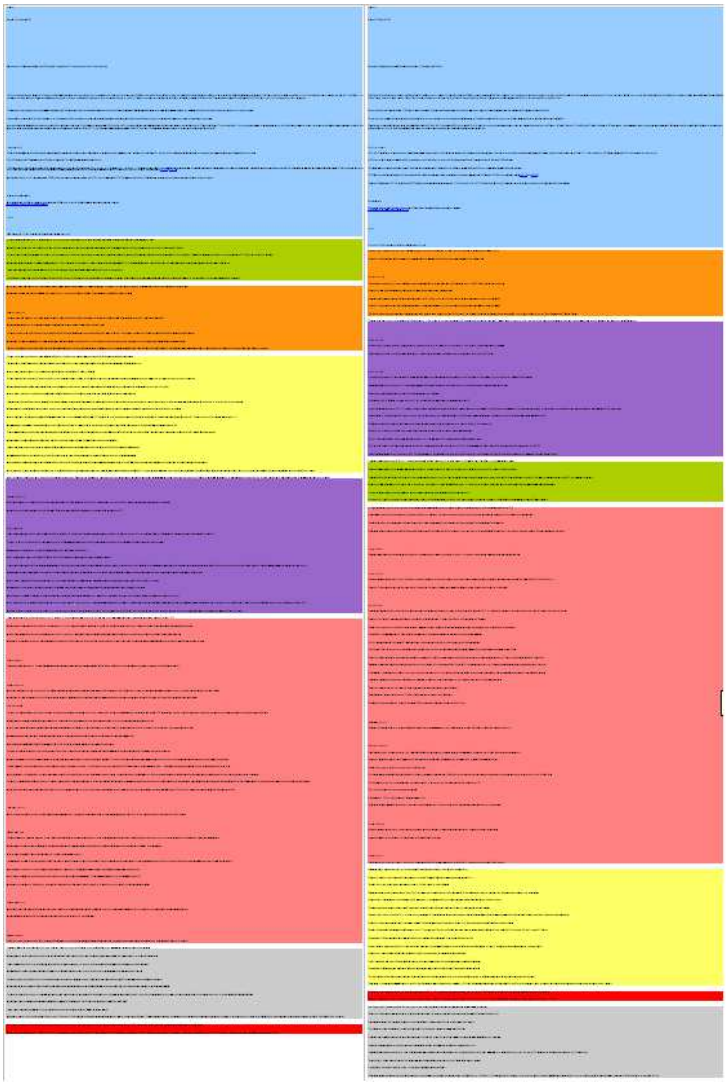


FIGURE 13 – Multizones FR-EN du même communiqué IP/05/1157.

5

UNE MÉTHODE TEXTUELLE GUIDÉE PAR LE MODÈLE

Notre méthode d'alignement est orientée analyse textuelle. Elle prévoit les problèmes de comparativité liées tant à l'activité du traducteur qu'aux différences entre les langues. La difficulté inhérente aux méthodes d'alignement endogènes est de savoir par quels alignements commencer, a fortiori lorsqu'elles sont appliquées sur des corpus multilingues et potentiellement bruités. Un alignement endogène ne peut être que progressif. C'est-à-dire qu'il ne peut que se situer dans un cadre itératif, alignant soit de façon ascendante, soit de façon descendante. De façon ascendante, en appariant d'abord les cognats, chaînes de caractères identiques entre plusieurs langues et en définissant à partir d'eux, des zones dont la taille sera progressivement étendue. De façon descendante, en mettant progressivement en correspondance des zones de texte sémantiquement équivalentes, à l'intérieur desquelles, nous recherchons à nouveau des multizones plus petites. Nous faisons le choix d'une méthode descendante qui n'impose pas le parallélisme mais recherche et calcule en contexte les zones de textes où il existe. Il existe un continuum entre des équivalences linguistiques répertoriées dans les ressources dictionnairiques, i.e. les appariements, et les équivalences traductionnelles observables en contexte, i.e. des alignements. Notre méthode d'alignement endogène prend en considération ce continuum et propose de l'exploiter dans le traitement homogène, multilingue et multiéchelle d'une collection de multidocuments.

SOMMAIRE

5.1 CARACTÉRISTIQUES DE LA MÉTHODE

5.1.1 *Une méthode descendante*

Notre méthode (figure 14) est descendante et s'attaque au problème de la détection de parallélisme suivant la hiérarchie de grain (figure 11b) : Document ⇒ Zone ⇒ Segment ⇒ N-grammes de caractères. Nous pouvons résumer ses caractéristiques principales en quelques points :

- le processus d'analyse prend en entrée des multidocuments ;
- le premier objectif est de proposer des outils de diagnostic de parallélisme : synchrones ou asynchrones et le cas échéant de détection en contexte des zones qui maximisent le parallélisme à l'intérieur de chaque multidocument, l'objectif second est celui d'un alignement lexical de ces zones ;
- elle comporte deux étapes intermédiaires servant d'amorces :
  - établir des correspondances multilingues de chaînes de caractères à partir d'une collection de multidocuments ;
  - les utiliser pour définir la similarité de segments de textes de niveau supérieur.

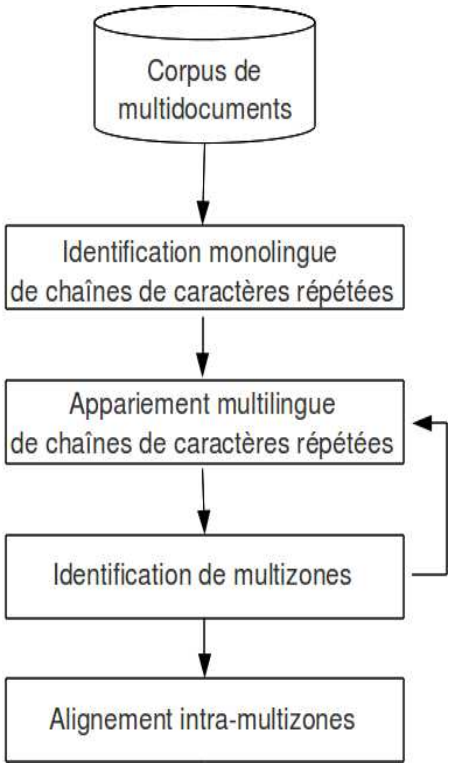


FIGURE 14 – Chaîne de traitement

Ainsi partant du principe que des différences entre les volets existent, même là où on ne les attend pas, nous proposons un relâchement des contraintes de parallélisme intra-multidocument, visant à diagnostiquer

en contexte les zones des documents à l'intérieur desquelles le parallélisme existe. Pour cela, nous faisons l'hypothèse que la co-présence de chaînes de caractères peut suffire à retrouver des zones sémantiquement équivalentes.

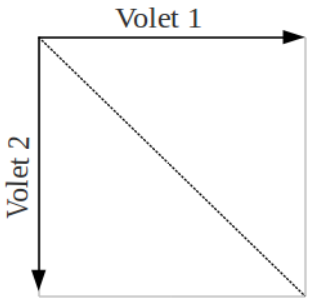
De façon théorique, nous présentons dans la section suivante les modèles d'alignement de zones qui nous guident. Car si l'on peut affirmer que tout n'est pas toujours présent ou dans l'ordre, partant du principe que les documents sont effectivement traductions, tous les cas de distorsion de la diagonale ne peuvent pas être envisagés.

5.1.2 *Différents types d'alignement de zones*

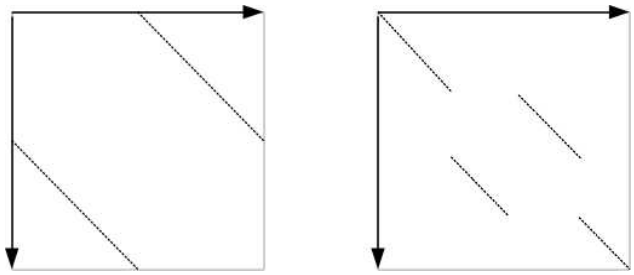
Nous présentons dans la figure 15 les différents attendus en matière de visualisation de la structure des multidocuments.

La figure 15a, page 65 présente le cas idéal d'une traduction globalement synchrone où la structure est la même dans les deux volets. Chaque point sur la diagonale représente des alignements d'unités aux mêmes positions dans les deux volets. Ainsi, la présence de la diagonale complète signifie que les volets ne présentent ni inversion, ni suppression. Nous avons une seule multizone équivalente au multidocument. L'alignement de zones est de type (1 : 1). Les figures 15b (page 65) et 15c (page 65), quant à elles, sont asynchrones, dans un cas tout n'est pas dans le même ordre et dans l'autre tout n'est pas présent. La figure 15b, présente deux cas d'inversions. La première est simple, elle correspond à l'interversion de deux zones de textes du volet 1 dans le passage au volet 2. Ce type d'interversion correspond à celle présentée dans la figure 16, page 66. La seconde est multiple, plusieurs zones du volet 1 subissent un déplacement dans le passage au volet 2, c'est le cas que nous avons pu observer au travers des volets français et anglais du communiqué IP/05/1157 présentées dans la figure 6, page 22. La figure 15c présente trois cas de suppressions respectivement au début, au milieu et à la fin, comme nous avons pu l'observer à travers l'exemple du communiqué IP/05/473 présenté dans la figure 8, page 25.

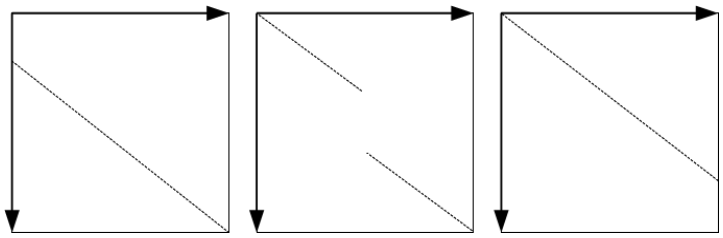
Ces figures correspondent à un attendu observable et définissable à l'œil nu. Dans nos expériences, nous établissons un diagnostic automatique définissant si les multidocuments sont : synchrones, asynchrones ou si le diagnostic n'est pas établi : « indéfini ». Le diagnostic indéfini est un diagnostic intermédiaire donnant lieu à un nouveau traitement automatique du multidocument. Il pourra par exemple être plongé dans une nouvelle collection plus grande ou de documents thématiquement proches ou en cas d'échec donner lieu à une observation manuelle. Il pourrait s'avérer que des documents identifiés comme traductions via leur url ne le soit pas en réalité. Une fois le diagnostic posé, nous répétons



(a) Deux volets synchrones.



(b) Modèles d'inversions globale (à gauche) et locale (à droite).



(c) Modèles de suppressions respectivement situées au début, au milieu et à la fin (de gauche à droite).

FIGURE 15 – Modèles des différents types d'alignement de zones.

les mêmes opérations sur les multizones ainsi détectées pour parvenir cette fois à un alignement lexical.

5.2 ALIGNEMENT DE ZONES

5.2.1 *Recherche de multizones*

La méthode repose sur la recherche de multizones, des portions de documents globalement sémantiquement équivalents entre les volets d'un multidocument. Elles peuvent correspondre à tout ou partie d'un multidocument. Le multidocument est une multizone donnée a priori. C'est-à-dire que nous savons d'emblée que ses différents volets ont globalement le même sens et que par conséquent, il existe entre eux ce que nous appelons des *faisceaux* de liens sémantiques à différents niveaux (figure 16).

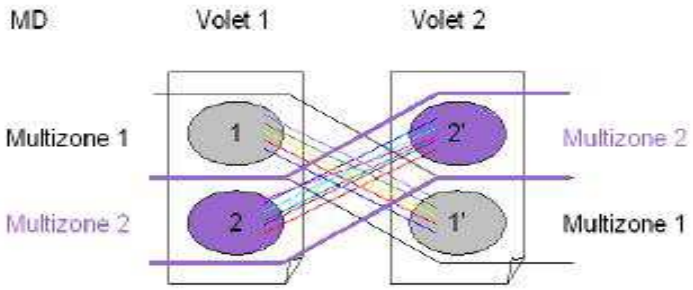


FIGURE 16 – Multizones et interdépendances entre les grains

Partant de cette connaissance qu'un volet d'un multidocument équivaut globalement aux autres volets, nous cherchons à faire émerger des multizones, c'est-à-dire que loin de supposer l'ordre ou le désordre entre les volets, nous cherchons à le constater, à le calculer. Calculer les multizones en contexte permet de garantir le bon déroulement de l'alignement. Nous n'intégrons pas de bruit. Si les indices ne convergent pas, les alignements ne sont pas considérés comme bons.

Dans la figure 17, nous observons cinq multizones. Observons les deux zones entourées de vert, il existe un faisceau de liens qui convergent, autrement dit il y a à un certain niveau un parallélisme entre ces deux zones. Les chaînes de caractères « verre » en FR et « glass » en EN notamment y apparaissent et permettent de le révéler. Ces deux zones constituent ce que nous appelons des multizones.

5.2.2 *Calcul des multizones : entre alignement et appariement*

Dans un document, chaque zone se distingue des autres zones du document par une liste et une densité de populations. Nous appelons population l'ensemble des occurrences d'une suite de N-grammes de caractères répétés dans une langue, nous appelons appariement la mise en correspondance de ces populations. Nous appelons individu, une occurrence d'un N-gramme d'une de ces populations et nous appelons alignement la mise en correspondance de deux de ces individus. Ces différences nous permettent de calculer la correspondance entre des zones équivalentes.

Dans l'exemple de la figure 18, les populations C et D sont toutes les 2 uniquement présentes dans le multidocument 1 et comportent le même nombre d'individus. Les populations A et B présentent les mêmes effectifs sur la collection. Mais alors que la population A est présente dans les multidocuments 1 et 3, la population B apparaît elle dans les multidocuments 1 et 2. Ainsi, les meilleurs candidats pour l'appariement avec ces deux populations dans la collection sont respectivement : (A, A',A'') et (B, B',B''). En outre, la population A apparaît dans les multidocuments 1 et 3 avec la population E, mais seulement dans le multidocument 1

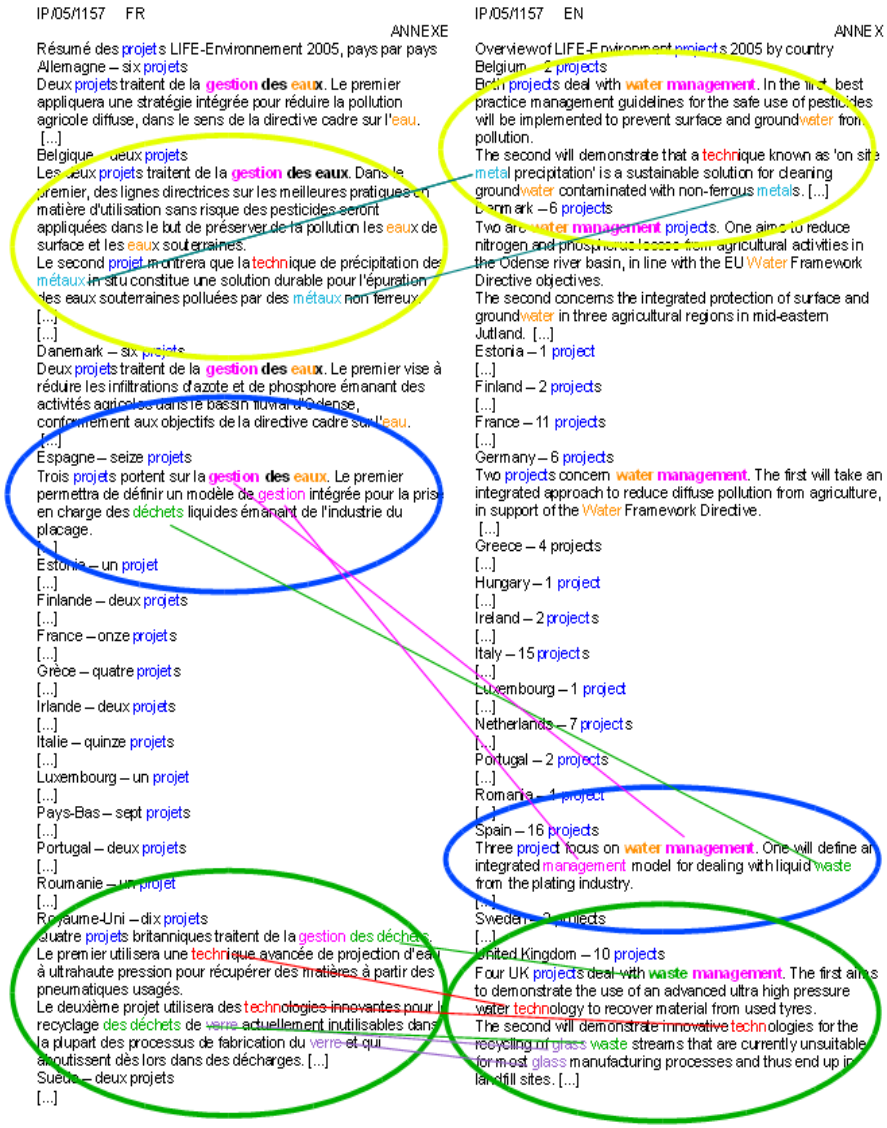


FIGURE 17 – Détection de multizones

avec la population F, tandis que la population B est co-présente avec la population F dans les multidocuments 1 et 2. Les répartitions sur la collection des populations A, B, E et F servent pour l'alignement des populations C et D, respectivement avec (C' et C'') et (D' et D'').

La méthode que nous proposons est descendante et repose sur les hypothèses suivantes (figure 19) :

- dans une collection de multidocuments, un volet dans une langue équivaut au moins partiellement aux autres volets dans les autres langues du multidocument ;
- dans une collection de multidocuments, un n-gramme de caractères d'une langue partage avec ses équivalents dans les autres langues, tout ou partie de sa liste de multidocuments. Autrement

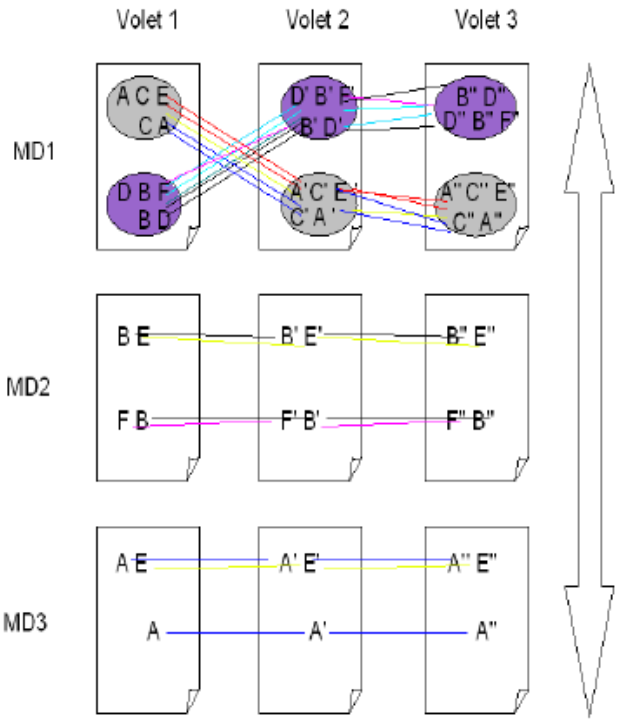


FIGURE 18 – Détection de multizones via la collection de multidocuments

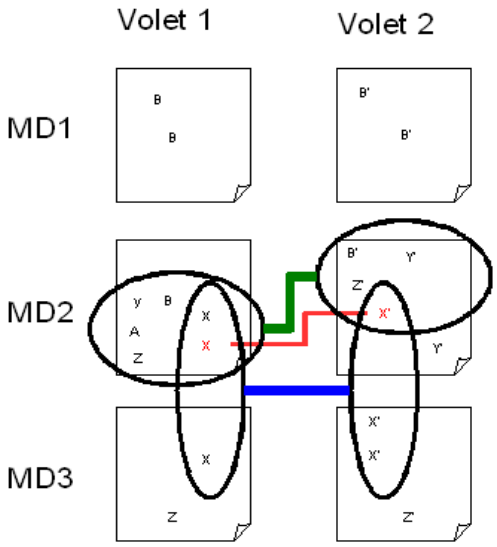


FIGURE 19 – Multizones : entre alignement et appariement (bleu : appariement de populations, vert : alignement de zones, rouge : alignement d'individus)

dit si deux n-grammes ne sont pas partagés par au moins un mul-

tidocument, s'ils ne partagent aucun contexte, ils ne peuvent pas être sémantiquement équivalents ;
– il existe un partitionnement optimal de la collection qui met en évidence des zones sémantiquement équivalentes de tailles variables, pour lesquelles le nombre d'alignements est maximisé.

Partant du fait que le multidocument est une multizone de taille maximale, un volet équivaut globalement aux autres volets, nous cherchons à faire émerger des multizones. Au lieu de supposer l'ordre ou le désordre entre les volets d'un multidocument, celui-ci sera calculé en contexte. Ainsi à la façon du système K-vec (Fung et Church, 1994) vu précédemment, nous nous attachons à la comparaison de vecteurs d'effectifs d'unités textuelles. Cependant, à la différence de celui-ci, ces unités sont dénombrées par document de la collection, et non par portion de document. La seule position d'une unité que nous considérons est la position « document » : telle unité en français est présente *n* fois dans tel document en français et *n* fois dans tel autre document français et ne l'est pas dans tel autre.

L'analyse de chacun des multidocuments de la collection est faite avec l'aide d'une collection de multidocuments, tirée du corpus (voir section 3.3). Considérons une collection de quatre multidocuments (Md1, Md2, Md3 et Md4) en deux langues (l1 et l2) et la répartition sur la collection de trois individus (un en langue 1 et les autre en langue 2) :

Individus	Effectifs par document			
	Md1	Md2	Md3	Md4
Individu $_{l1}$	4	6	12	1
Individu $_{l2}$	4	7	10	1
Individu $_{l2}$	4	3	20	0

TABLEAU 11 – Vecteurs d'effectifs par document de trois individus dans une collection de multidocuments

En ne considérant pas les positions des individus à l'intérieur des volets des multidocuments de la collection, nous parvenons néanmoins à révéler des similitudes entre eux et à les aligner sans imposer le parallélisme entre ces volets : l'« Individu $_{l1}$ » présente davantage de similitude de répartition sur la collection avec le premier « Individu $_{l2}$ » qu'avec le second « Individu $_{l2}$ ».

En amont du processus d'alignement, la collection nous sert également à délimiter et sélectionner les chaînes de caractères présentant un intérêt pour l'appariement. Notre critère de délimitation des chaînes étant la répétition, nous nous servons de la collection pour la favoriser. Ainsi nous ne conservons que les chaînes de caractères d'au moins deux

occurrences dans la collection, ces occurrences peuvent néanmoins être dans le même document.

Notre stratégie globale d'alignement est multiéchelle, c'est-à-dire qu'elle sera la même à tous les grains. Ainsi, la notion de zone mise en évidence précédemment pourra recouvrir plusieurs réalisations concrètes : du document lui-même au n-gramme en passant par le paragraphe, la phrase ou la proposition. Plus les zones seront petites, plus l'inertie intramultizone devra être minimisée, au profit de l'inertie intermultizone, plus, notamment, les ajouts et les suppressions de zones seront déterminants dans l'alignement.

5.3 APPARIEMENT ENDOGÈNE DE CHAÎNES DE CARACTÈRES RÉPÉTÉES

Si l'alignement monolingue peut s'appuyer sur une similitude de graphie (Bourdaillet et Ganascia, 2007), l'alignement multilingue ne peut s'en contenter. Il doit donc établir des similitudes entre les chaînes répétées dans chacune des langues sur un autre critère. Selon nos observations, un découpage en N-grammes de caractères répétés permet de faire émerger les facteurs communs nécessaires.

5.3.1 *Capacité des N-grammes de caractères à révéler des correspondances monolingues*

Pour un document donné dans une langue, une segmentation en N-grammes de caractères met en évidence des facteurs communs qu'un découpage en N-grammes de mots ne révèle pas.

Prenons l'exemple d'un échantillon de document français et de sa traduction en finnois.

Nous cherchons les N-grammes de mots répétés d'un échantillon de document en français :

FR Donner aux collectivités les moyens de développer les transports en commun. La Commission européenne a adopté aujourd'hui une proposition révisée d'un règlement qui contribuera au développement de services publics de transport en commun.

⇒ **3 N-grammes de mots sont répétés.**

Nous cherchons les N-grammes de caractères répétés (ici, plus de 3 caractères, espaces compris) du même échantillon :

FR Donner aux collectivités les moyens de développer les transports en commun. La Commission européenne a adopté aujourd'hui une proposition révisée d'un règlement qui contribuera au développement de services publics de transport en commun.

⇒ **5 N-grammes de caractères sont répétés.**

Nous cherchons les N-grammes de mots répétés d'un échantillon de document en finnois :

FI Paikallisviranomaisille tarjotaan keinot joukkoliikenteen kehittämiseen. Euroopan komissio hyväksyi tänään tarkistetun ehdotuksen asetukseksi jolla edistetään julkisten joukkoliikennepalvelujen kehittämistä.

⇒ **0 N-gramme de mots répété.**

Nous cherchons les N-grammes de caractères répétés (ici, plus de 3 caractères, espaces compris) du même échantillon :

FI Paikallisviranomaisille tarjotaan keinot joukkoliikent een kehittämiseen. Euroopan komissio hyväksyi tänään tarkistetun ehdotuksen asetukseksi, jolla edistetään julkisten joukkoliikennepal velujen kehittämistä.

⇒ **6 N-grammes de caractères sont répétés.**

Ainsi, en nous attachant aux chaînes de caractères répétées, nous souhaitons capturer par le même mécanisme des unités qui s'étendent sur moins d'un mot comme sur un ou plusieurs mots :

- des expressions répétées plus longues que des mots, détectant ainsi le figement ;
- des racines de mots se répétant en général avec plus de constance que les formes fléchies, notamment dans les langues morphologiquement riches et/ou agglutinantes ;
- des indices de forme (en général des parties de balises HTML) pas nécessairement intéressants dans l'optique de constitution de lexiques multilingues mais des éléments précieux comme points d'ancrage pour l'alignement.

Outre l'augmentation du nombre d'unités répétées, nous pouvons également considérer la nature de ces derniers : il nous apparaît qu'un découpage en N-grammes de caractères en favorisant la répétition met davantage de segments signifiants en évidence.

LANGUE	MOTS	CHAÎNES DE CARACTÈRES
fr	transport, transports, transporter, transportation	transport-

Tableau 12 – Mise en évidence de la chaîne de caractère commune à quatre mots formés par dérivation

Ici, même en mettant en œuvre pour les N-grammes de mots, un traitement type singulier/pluriel suffisant dans le cas de la flexion de

'transport'/ 'transports' , toutes les équivalences ne pourraient pas être révélées, c'est le cas notamment de la dérivation 'développer'/ 'développement' (cf. également tableau 12). L'usage dans ces cas est de faire appel à des dictionnaires, mais ceci a un coût, en termes de construction, de maintenance et donc d'extension du système à de nouvelles langues, auquel l'extraction de N-grammes de caractères n'est pas soumise.

5.3.2 *Capacité des N-grammes de caractères à mettre en évidence des correspondances multilingues*

Le problème de l'alignement multilingue est un problème de similarités et de différences de sens, graphie et répartition. Les facteurs communs monolingues, d'ordre graphique, précédemment révélés, mettent en évidence des segments de textes sémantiquement proches. Celles-ci peuvent à leur tour servir à révéler des similarités multilingues de répartition. Entre deux langues, des formes différentes mais sémantiquement équivalentes ont des répartitions semblables entre deux documents traductions l'un de l'autre.

Entre deux documents traductions l'un de l'autre, l'écart entre les effectifs de ***N-grammes de caractères*** *sémantiquement équivalents est inférieur à l'écart entre les effectifs des* ***N-grammes de mots*** *graphiques sémantiquement équivalents. L'alignement des mots graphiques échoue d'autant plus que les langues comparées sont morphologiquement différentes.*

LANGUE	MOTS GRAPHIQUES SIGNIFIANT « TRANSPORT » ET LEUR EFFECTIF
fr	transports (3), transport (3)
es	transporte (5), transportes (1)
el	μεταφορών (3), μεταφορέας (1), μεταφορές (1), μεταφορέα (1)

Tableau 13 – Liste des mots graphiques signifiant « transport » dans un échantillon de textes en fr, es et el, et leur effectif.

Ici, comme en témoigne le tableau 13, les écarts d'effectifs entre des mots alignés dans un échantillon sont déjà considérables. Or si l'on s'intéresse désormais aux répétitions de chaînes de caractères, on s'aperçoit qu'il existe dans chaque langue une sous-chaîne commune à l'ensemble des équivalents sémantiques de « transport ».

Cette sous-chaîne commune apparaît donc comme un moyen de comparaison des langues susceptible de passer à l'échelle à moindre coût. Les écarts d'effectifs entre les mots partiellement ou intégralement

LANGUE	CHAÎNES DE CARACTÈRES RÉPÉTÉES SIGNIFIANT "TRANSPORT"	EFFECTIFS
fr	transport- (3+3)	6
es	transporte- (5+1)	6
el	μεταφορ- (3+1+1+1)	6

Tableau 14 – Chaînes de caractères (d'au minimum 3 caractères) communes aux mots signifiant « transport » dans le même échantillon de textes en fr, es et el et leur effectif respectif.

équivalents se trouvent lissés. La mise en correspondance de séquences de caractères sémantiquement équivalentes en contexte entre plusieurs langues sera facilitée, le schéma d'alignement ne pouvant plus être que de l'ordre du 1 pour 1 ou du 0 pour 1, en cas d'absence de traduction. Prenons par exemple, les différentes occurrences d'un signifié tel que « collectivités » en finnois : « paikallisviranomaisille », « paikallisviranomaisten », « paikallisviranomaiset », « paikallisviranomaisilla », seront rapportées à la séquence de caractères « paikallisviranomai* », plus longue sous-chaîne commune. Ce travail en chaînes de caractères a pour effet de lisser les différences de fréquences de ces équivalents, engendrées dans ce cas par la nature flexionnelle du finnois.

5.3.3 *Incapacités des N-grammes de caractères*

Nous présentons dans cette section, trois limites à la segmentation-alignement de N-grammes de caractères. Celles-ci trouvent une solution via la mise en place d'un traitement informatique spécifique et/ou adapté :

- les mots lexicaux ou polylexicaux dont une ou plusieurs lettres changent, dans le cas de diphtongaison comme celle du verbe « contar » en espagnol, aux premières personnes du présent : « cuento », « cuentas », « cuenta » (i.e. skip-grams pour Mcnamee et Mayfield (2004) ou SFM Séquences Fréquentes Maximales avec possibilité d'avoir un *gap* entre les mots de la séquence pour Doucet (2004)). Ici, sans autre traitement, l'alignement de N-grammes de caractères ne permet pas de révéler davantage qu'un alignement basé sur des N-grammes de mots.
- le risque de mettre en rapport des chaînes de caractères non liées au niveau du mot, entre « transport » et « transparence » par exemple.
- la surgénération de chaînes répétées « inintéressantes » dans le but de construction de ressources lexicales par une méthode

d'alignement. Le fait de supposer que tout N-gramme de caractères d'une langue puisse être aligné avec n'importe quel N-gramme dans une autre langue nous permet de trouver beaucoup d'associations mais impose de fixer des règles pour parcourir ce très grand espace de recherche. Nous avons résolu ce problème en comparant les positions de N-grammes de fréquences similaires.

5.4 DE L'ALIGNEMENT DE ZONES À L'ALIGNEMENT INTRA-MULTIZONES

Nous considérons dans cette section le cas particulier de documents courts (1 à 2 pages), comme c'est le cas des communiqués de presse qui constituent notre corpus. Une fois les zones maximisant le parallélisme identifiées, le principe est de reprendre un alignement intra-multizones des individus qui les composent en favorisant le parallélisme. Ainsi, suivant la taille des zones composant ces multizones, un appariement détecté au moyen de la collection peut y apparaître ou non, répété ou non. Dans le cas où il est répété, nous considérons que la première occurrence d'un N-gramme de caractères en langue L1 apparié grâce à la collection à un N-gramme de caractère de la langue 2 est aligné avec la première occurrence de ce dernier dans la multizone et le deuxième avec le deuxième.

Ainsi nous regroupons dans un même corpus, les multidocuments synchrones et les multizones des documents asynchrones alignées pour calculer l'alignement intra-multizones. Dans cette dernière, nous pourrons présupposer le parallélisme puisque celui-ci aura été mesuré à l'étape précédente. À cette étape, l'espace de recherche se situe autour de la diagonale. Des stratégies devront être prévues pour aligner au mieux les zones résiduelles des multidocuments asynchrones, i.e. les zones n'ayant pu faire l'objet d'un alignement par manque d'information sur leur contenu ou par absence d'équivalent. De façon générale, le diagnostic devra pouvoir identifier le type exact de structure auquel correspond la traduction.

Dans ce chapitre 5, nous avons présenté les principes d'une méthode descendante sans présupposé de parallélisme. Cette méthode propose un relâchement des contraintes de parallélisme et vise à diagnostiquer en contexte les zones à l'intérieur desquelles le parallélisme existe.

Troisième partie

MISE EN ŒUVRE, ILLUSTRATIONS, ÉVALUATION

6

MISE EN ŒUVRE

Les travaux de mise en œuvre présentés dans ce chapitre sont le fruit de plusieurs rencontres et collaborations. Tout d'abord, avec Loïs Rigouste, au sein de notre lieu de stage, la société Pertimm, nous avons spécifié et développé les principes de calcul des populations sur une collection de multidocuments. Puis, à l'Université de Caen, Romain Brixtel a adapté à nos objets ses outils d'analyse et de visualisation de bi-documents, plaçant ainsi notre problématique dans le domaine du traitement d'image. De là nous avons été amenée à solliciter les connaissances et les compétences de Régis Clouard, spécialiste du traitement d'image de l'équipe Image du laboratoire GREYC de l'Université de Caen. Cette dernière collaboration nous a permis d'obtenir des outils capables d'analyser automatiquement les images que nous avions désormais à analyser. Le traitement de ces images reflétant l'appariement entre deux volets pose les bases d'un diagnostic automatique du parallélisme entre des bi-documents et par là d'un alignement de multidocuments sans présupposé de parallélisme.

SOMMAIRE

6.1 APPARIEMENT ENDOGÈNE DE POPULATIONS DE N-GRAMMES DE CARACTÈRES RÉPÉTÉS DANS UN CORPUS MULTILINGUE AU FORMAT HTML

Dans cette section, nous décrivons les expérimentations que nous avons faites en matière d'amorce fréquentielle en vue d'un alignement de multidocuments. Notre premier objectif consiste à obtenir de façon endogène et indépendante des langues une série de points de comparaison entre deux volets : des appariements. Pour mettre en œuvre les principes précédemment évoqués, nous avons implémenté les étapes de calcul de populations de N-grammes de caractères et d'appariement de ces populations. Les meilleurs appariements sont utilisés dans la phase suivante pour la création des matrices de points [1].

6.1.1 *Calcul des populations de N-grammes de caractères*

Les populations sont déduites d'un tableau de suffixes (Crochemore *et al.*, 2007; Kärkkäinen et Sanders, 2003). Ce dernier permet de calculer la liste des chaînes de caractères répétées de longueur maximale, c'est-à-dire les chaînes monolingues répétées qui ne sont pas incluses au sein d'autres répétitions de même effectif. De façon empirique, dans une démarche d'amorce, nous ne considérons que les chaînes de longueur égale ou supérieure à 5 caractères.

Le tableau 15 présente des exemples de populations extraites d'une collection de multidocuments. Celles-ci ne font par nécessairement directement sens pour l'humain. Les chaînes de caractères de ces populations s'étendent selon le cas sur moins d'un mot, plus d'un mot, voire sur plusieurs mots. Étant donné que nous prenons en compte la structure et le contenu par la même méthode, ces chaînes de caractères peuvent naturellement être ou contenir des morceaux de balises HTML. Certaines sont des hapax de documents mais sont répétées dans la collection. L'appariement de ces dernières constitue un ancrage robuste pour la suite. En revanche, certaines ne correspondent pas à la langue annoncée, comme c'est le cas du deuxième exemple en grec, ce qui témoigne d'ores et déjà de l'intrusion d'extraits de documents dans d'autres langues que celle dans laquelle les documents sont étiquetés.

Nous trions ces populations par effectif décroissant, puis à effectif égal par longueur des chaînes. Considérant nos hypothèses de travail, effectuer ce classement des populations est là encore une manière de rapprocher des unités potentiellement alignées. Le tri sur les effectifs des chaînes répétées sur l'ensemble du corpus fait que la méthode n'est plus sensible aux inversions locales et que statistiquement les décalages

1. Les outils permettant la création de ces matrices sont disponibles ici : `http://code.google.com/p/zone-align/`

LANGUE	POPULATION (effectif dans la collection)	POSITIONS N°Md : offset normalisé
en	'**Commission**' (319)	4:81% 10:5% 16:40% 14:32%[…]
	' **neighbouring countries**' (6)	4:66% 10:12% 11:9% 12:6% 12:15% 16:73%
	'. < /*p* >< *p* > *Thefourthproject*' (5)	10:90% 10:47% 10:44% 11:78% 11:81%
	'**ber 2004.** ' (2)	3:80% 36:99%
fr	'**ir les c**' (7)	4:47% 10:16% 11:12% 12:20%[…]
	'**s environnementaux**' (5)	11:26% 11:5% 11:4% 12:1% 12:2%
	' **projet concerne la** ' (4)	10:91% 10:62% 10:93% 11:80%
fi	'**n elvytyssuunnitelman** ' (2)	36:1% 36:2%
	' **elektroniikkalaitteissa**' (2)	10:30% 10:35%
de	'**. Bei dem ersten wird ein** ' (2)	5:15 33:24
	'**ng und Werbung,** ' (2)	56:38% 51:79%
el	'' **Οι προτάσεις που** ' (2)	64:28% 60:10%
	' **departing from an** ' (2)	52:74% 52:74%

TABLEAU 15 – Exemple de populations extraites d'une collection de multidocuments en français, anglais, finnois, allemand et grec. Chaque ligne fournit pour une chaîne de caractère répétée : la langue, la **'chaîne'**, son (effectif) dans la collection et les positions de chacun de ses individus sous la forme : numéro de multidocument : offset dans le volet, normalisé sur 100.

d'effectifs peuvent se compenser. Les effectifs monolingues des populations sont notre premier critère de classement pour trouver des candidats à l'appariement. Les populations similaires d'une langue à l'autre ont la caractéristique d'apparaître approximativement le même nombre de fois dans une langue donnée. Ce critère pris isolément est naturellement insuffisant pour proposer des appariements. L'observation des effectifs ne peut à lui seul être un indicateur fiable d'appariement, cette phase de classement devra nécessairement être suivie d'un calcul de distance puisque, comme le souligne Zimina (2006, p.4) : « Lorsqu'il s'agit de mots dotés d'un large éventail de sens dans le corpus, les correspondances lexicales entre les volets forment un réseau complexe et la comparaison des effectifs totaux des formes graphiques ne constitue pas toujours une bonne indication pour l'appariement ».

Nous introduisons donc dans la section 6.1.2, pour chaque population, une étude des positions dans la collection des individus qui les composent, ou plus précisément une étude de leur effectif par document.

6.1.2 *Appariement de N-grammes de caractères répétés à partir de ventilation similaire sur la collection*

Nous avons donc en sortie de l'étape précédente une liste de populations triée par effectif monolingue. Afin de limiter l'explosion combinatoire d'un calcul exhaustif entre toutes les chaînes répétées maximales,

nous comparons les chaînes d'effectifs proches. En tout état de chose, les chaînes en dessous du seuil que nous nous fixons sont nécessairement d'effectifs proches. Pour conduire des tests d'appariement plus poussés, nous faisons passer une fenêtre glissante sur cette liste et, pour chaque position de la fenêtre, nous testons l'appariement du dernier élément avec tous ceux qui le précèdent. Pour une fenêtre de taille F [2], on aura donc calculé une distance sur les positions dans la collection (selon une méthodologie que nous précisons ci-dessous) entre une population et les 2F-2 populations les plus proches de la liste (F-1 au-dessus et F-1 en-dessous). Même s'ils pourraient se révéler intéressants pour d'autres applications, nous ignorons ici les couples constitués de populations de même langue. Ils pourraient servir à révéler les couples dont les apparitions sont fortement corrélées. Nous nous concentrons sur les liens interlingues.

langue	N-gramme	effectif dans la collection	effectif par volet			
			$volet_1$	$volet_2$	[...]	$volet_{200}$
el	'_αερολιμέν'	(23)	4	2	[...]	3
fr	'aéroports'	(21)	4	2	[...]	2

Tableau 16 – Exemple de répartitions de deux N-grammes de caractères en grec et en français. Les espaces sont représentés par le caractère « _ ».

Ainsi, nous calculons les appariements entre chaînes de caractères de langues différentes, en prenant en compte des similitudes de répartitions sur l'ensemble des bi-documents. Un exemple de répartitions par volet de deux N-grammes de caractères est donné dans le tableau 16.

Pour calculer les appariements, nous utilisons une distance L1 normalisée, elle consiste à faire pour deux N-grammes de caractères (s_1 et s_2) de deux langues différentes (l1 et l2), le rapport entre la somme des différences d'effectifs par document et la somme des effectifs des deux N-grammes dans la collection de bi-documents dans ces langues.

$$distance\mathrm{L1}(s_1, s_2) = \frac{\sum_{doc} |effectif(s_1, volet_{l1}) - effectif(s_2, volet_{l2})|}{effectif\_corpus(s_1) + effectif\_corpus(s_2)}$$

Ce calcul de distance génère des appariements entre deux populations de N-grammes de caractères avec une distance située entre 0 et 1.

2. Dans nos expériences, nous avons essayé plusieurs tailles de fenêtres différentes, typiquement entre 100 et 10000. Plus la collection est grande, plus la fenêtre doit l'être aussi, afin d'être sûre de comparer les N-grammes d'effectifs proches. Plus on arrive dans les faibles effectifs, plus il y a de candidats à comparer. Pour une collection de 40 multidocuments, une fenêtre de 40 suffit.

Les meilleurs appariements ont une distance de 0. Cette distance fait l'hypothèse que certains termes sont globalement traduits de la même manière au travers des documents en relation de traduction et qu'ils ont donc une répartition analogue calculable. Cette distance ne prend en considération les positions des individus qu'en terme de présence/absence dans les différents volets. Plus précisément nous comparons des populations via leur vecteur d'effectifs par volet dans chaque langue, sans tenir compte des positions des individus à l'intérieur des volets [3]. Les deux propriétés principales de cette distance sont donc de :

- calculer des correspondances fortement généralisées dans une collection de multidocuments ou multizones, des correspondances bi-univoques ou quasi bi-univoques.
- être insensible aux différences d'ordres entre les volets et aux suppressions locales de zones de textes.

Nous donnons quelques exemples d'appariements ainsi calculés dans le tableau 17, page 82.

Les résultats de cette étape corroborent notre intuition qu'apparier des populations de chaînes de caractères à l'intérieur d'une collection de documents est une piste prometteuse. Ils prouvent qu'il existe bien des populations bi-univoques statistiquement identifiables. Dans l'annexe A page 123, nous présentons une expérience d'évaluation quantitative des résultats de l'opération d'appariement par rapport à des dictionnaires. Dans le chapitre 7, nous évaluerons s'ils sont en quantité suffisante pour permettre un diagnostic du parallélisme entre les volets d'un multidocument. Cette évaluation extrinsèque passe par la projection des appariements révélés sur des matrices de points qui font par la suite l'objet d'un traitement d'image.

3. Les offsets présentés dans le tableau 15 stockés au moment du calcul des populations ne nous servent pas au moment du calcul de distance. Ils ne sont stockés que pour permettre un retour au texte. Ils nous permettent de tracer les liens entre les segments des volets (voir figure 20, page 85).

distance : 0.000
fr **' l'enseignement'** (4) : 4, 4, 31, 31
en **' teaching'** (4) : 4, 4, 31, 31

distance : 0.000
fr **'ette année, la '** (4) : 4, 7, 21, 34
en **' year, th'** (4) : 4, 7, 21, 34

distance : 0.000
fr **'es chiffres '** (4) : 3, 15, 24, 26
en **' figures '** (4) : 3, 15, 24, 26

distance : 0.000
fr ' **migratoires** ' (4) : 27, 30, 30, 30
en '**e migrati**' (4) : 27, 30, 30, 30

distance : 0.000
fr 'embre 2005 <h1> <a name="Heading' (25) : 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 21, 22, 24, 26, 28, 30, 32, 31, 33, 34, 35, 36, 37, 39
en 'ember 2005 <h1> <a name="Heading4"' (25) : 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 21, 22, 24, 26, 28, 30, 32, 31, 33, 34, 35, 36, 37, 39

distance : 0.000
de ' **the obligation** ' (2) : 53, 53
es ' **Member States to** ' (2) : 53, 53

distance : 0.000
de '>   C' (2) : 53, 53
es ' **de las compañías** ' (2) : 53, 53

distance : 0.053
el ' **this Regulation**' (9) : 52, 52, 52, 52, 52, 52, 52, 52, 52
fr ' width="12.' (10) : 42, 52, 52, 52, 52, 52, 52, 52, 52, 52

distance : 0.053
el ' "><b>H E ' (9) : 48, 45, 50, 68, 71, 72, 73,77, 79
fr ' "><b>L ' (10) : 48, 45, 50, 68, 71, 72, 73, 77, 78, 79

distance : 0.053
el ' παχυσαρκίας' (9) : 56, 56, 56, 56, 56, 56, 56, 56, 56
fr ''**obésité** ' (10) :56, 56, 56, 56, 56, 56, 56, 56, 56, 56

distance : 0.064
fr **' Parlement '** (25) : 1, 2, 2, 2, 2, 5, 6, 7, 7, 7, 7, 7, 7, 7, 12, 16, 16, 17, 17, 17, 19, 19, 19, 21, 27, 34
en **' European Parliament'** (22) : 1, 2, 2, 5, 6, 7, 7, 7, 7, 7, 7, 7, 12, 16, 16, 17, 17, 17, 19, 19, 19, 21, 27

distance : 0.083
fr **' Parlement '** (25) : 1, 2, 2, 2, 2, 5, 6, 7, 7, 7, 7, 7, 7, 7, 12, 16, 16, 17, 17, 17, 19, 19, 19, 21, 27, 34
en **' European Pa'** (23) : 1, 2, 2, 5, 6, 7, 7, 7, 7, 7, 7, 7, 12, 16, 16, 17, 17, 17, 19, 19, 19, 21, 27, 37

distance : 0.080
fr **'s aér'** (26) : 2, 7, 7, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 17
en **' airp'** (24) : 7, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10

TABLEAU 17 – Appariements de populations de chaînes de caractères répétées dans la collection. Chaque groupe de 3 lignes présente : ligne 1, la distance qui a été calculée entre deux chaînes de caractères sur la collection, elle se situe entre 0 et 1 ; lignes 2 et 3, respectivement pour la chaîne 1 et la chaîne 2 : la langue, la **'chaîne'**, son (effectif) dans la collection et la liste de numéros de multidocument dans lesquels elle apparaît.

6.2 APPARIEMENT ET ALIGNEMENT DE ZONES

Dans cette section, nous présentons les travaux réalisés en matière de détection de multizones. Ils comportent un travail préparatoire de création de matrices de points à partir des appariements préalablement détectés, une détection de multizones via un traitement de ces matrices et une phase de diagnostic établi en fonction des multizones révélées. La chaîne de traitement est illustrée au travers du tableau 18.

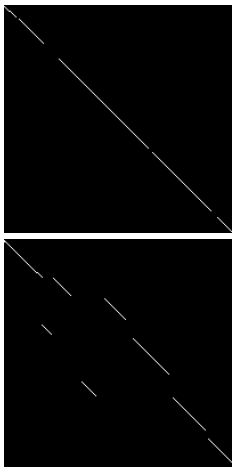
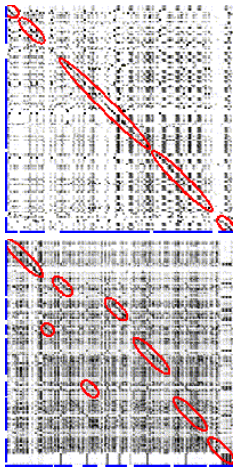
Matrice	Image binarisée	Segments de droites détectés	Diagnostic
			bi-document synchrone
			bi-document asynchrone

TABLEAU 18 – Traitement effectué sur chaque matrice. La première ligne présente le traitement effectué sur un bi-document danois-allemand (le communiqué de presse IP/05/489 de l'UE). La seconde présente le traitement effectué sur un bi-document anglais-français (le communiqué de presse IP/05/1157 de l'UE). Les images de droites illustrent la détection de multizones. Les segments de droites sont mis en évidence par des ellipses rouges, leurs projections sur les axes apparaissent en bleu.

Chacune des étapes est détaillée dans les sous-sections qui suivent.

6.2.1 *Travail préparatoire pour la détection de multizones : création de matrices de points*

La phase d'appariement constitue une amorce grâce à laquelle nous trouvons des segments de volets présentant des similitudes. Grâce à ces segments, nous révélons des zones de volets, autrement dit des grains supérieurs, présentant des similitudes : des multizones. Un segment de volet correspond à une portion de volet définie en pourcentage. Dans notre hiérarchie de grains (voir figure 11, page 57), il se situe entre la

zone et le N-gramme de caractères. Ainsi, une zone peut comprendre plusieurs segments et un segment plusieurs N-grammes de caractères.

Une matrice représente sous forme de points l'appariement entre les N-grammes de caractères de deux volets d'un multidocument. Tous les liens correspondant à un appariement de deux N-grammes de caractères calculé à partir de la collection et actualisé dans ce multidocument y sont pris en compte.

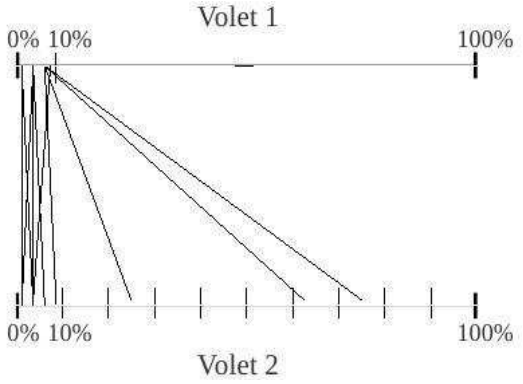
Chaque axe de nos matrices, axe horizontal et axe vertical, correspond à un des deux volets du bi-document à diagnostiquer. Il y a autant de points sur une ligne d'un axe que de segments de volet définis en paramètre. Les segments de texte peuvent se chevaucher, il ne s'agit pas d'une partition. Nous autorisons un chevauchement de nos segments pour éviter une segmentation trop abrupte de nos volets. Un segment est une sous-partie d'un volet que nous exprimons relativement à la taille du volet. Pour la même segmentation, S =(s^1_1,…,s^1_n), appliquée à deux volets, nous obtenons une matrice de similarité de taille n×n.

De façon empirique, nous choisissons pour traiter les communiqués de presse de notre corpus, une segmentation en 200 segments correspondant à 1% du document. Ces segments se chevauchent donc, S =(s_1 = [0, 0.01], s_2 = [0.005, 0.015], s_3 = [0.01, 0.02]…) pour chacun des deux volets. C'est en fonction de la répartition des segments similaires sur toute la matrice que nous calculons le parallélisme entre deux documents. Comme l'illustre la figure 20, deux segments sont considérés comme similaires lorsqu'ils maximisent le nombre de liens qui les relient.

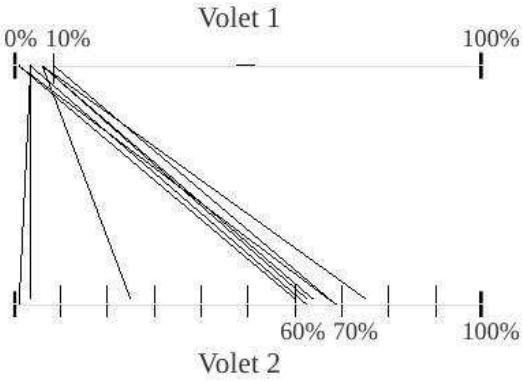
La figure 20 illustre la répartition et la densité des appariements de N-grammes de caractères entre un segment du volet 1 et les segments de même taille du volet 2. Dans notre exemple, les segments ne se chevauchent pas et correspondent chacun à un intervalle d'offset de 10% du volet. Les traits reliant les segments des volets symbolisent des appariements obtenus lors de l'étape décrite dans la section précédente et entrant dans la fourchette de distances voulues (typiquement entre 0 et 0.1). Un N-gramme de caractères présent dans le segment qui s'étend de 0 à 10% du volet 1 se voit attribuer autant de liens que le N-gramme de caractères qui lui est apparié est répété dans les segments du volet 2. Les appariements ainsi reportés mettent en évidence que dans la figure 20a, le segment 0-10% du volet 1 partage plus d'appariements avec le segment 0-10% qu'avec les autres segments du volet 2 tandis que dans la figure 20b, ce même segment partage plus de liens avec le segment 60%-70% du volet 2.

Pour calculer cette similarité entre deux segments, nous utilisons la fonction de score suivante :

$$score(s_1, s_2) = \frac{nb\_liens(s_1, s_2)}{max\_liens(s_1)}$$



(a) Segments similaires synchrones.



(b) Segments similaires asynchrones.

FIGURE 20 – Appariement directionnel entre les segments de deux volets.

$nb\_liens(s_1,s_2)$ représente le nombre d'appariements ayant une distance inférieure à 0.1 mettant en jeu des N-grammes de caractères inclus dans les segments 1 et 2, $max\_liens(s_1)$ représente le nombre de liens maximum entre le segment 1 et tous les segments de s_2 [4]. Pour éviter de supposer le parallélisme, nous considérons donc l'ensemble des liens possibles entre les occurrences des N-grammes appariés sans se focaliser sur un espace de recherche précis.

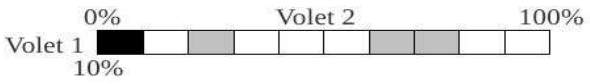
Segments(s_2)	[0]	[0.05]	[0.1]	[0.15]	[0.2]	[…]	[0.75]	[0.8]	[0.85]	[0.9]	[0.95]
Nombre de liens	14	3	0	0	0	[…]	0	0	2	0	0

TABLEAU 19 – Illustration de $max\_liens(s_1)$, max_liens vaut ici 14, le maximum sur la ligne

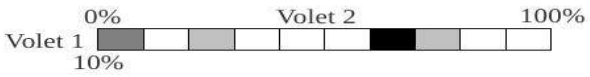
4. Ainsi la création des matrices est directionnelle. Nous n'obtenons pas le même rendu en comparant : langue 1 → langue 2 ou langue 2 → langue 1.

Dans la figure 19 (p.85), nous prenons pour illustrer $max\_liens(s_1)$, la distribution entre un segment donné du volet 1 s'étendant de 0 à 10% du document (ici : [0, 0.1]) avec chacun des segments du volet 2. Chaque ligne représente un segment sans chevauchement avec les autres, chacun correspondant à 5% du volet 2, 20 fenêtres en tout.

Étant donnée la méthode de construction des matrices précédemment décrite, nous pouvons dire que plus un point de la matrice est noir, plus les segments qui le composent sont similaires, i.e. plus il existe de liens issus de l'étape d'appariement décrite dans la section 6.1.2. La figure 21 présente les lignes de matrice correspondant aux deux types d'appariement de segments présentés dans la figure 20.



(a) Ligne de matrice correspondante à la figure 20a



(b) Ligne de matrice correspondante à la figure 20b

FIGURE 21 – Coloration d'une ligne de matrice.

Les matrices présentent donc différents niveaux de gris. Une similarité maximale est représentée par un pixel noir. Plus un pixel est blanc, plus les segments associés sont différents suivant notre fonction de similarité.

Ainsi, si deux documents sont traduits de façon globalement littérale, alors une diagonale se dessine de l'angle supérieur gauche à l'angle inférieur droit de la matrice. Une diagonale brisée signifie au contraire l'existence d'inversions dans l'ordre de la traduction.

Ainsi, la question qui subsiste est celle de la détection automatique des segments de droites autrement dit des multizones que nous observons sur ces matrices. Nous présentons dans la section 6.2.2 les étapes du traitement réalisé sur ces images.

6.2.2 *Détection des multizones à partir des matrices*

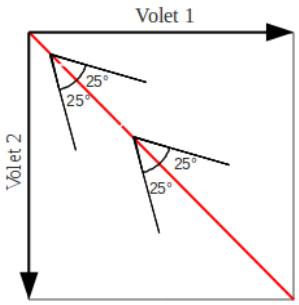
Le problème de la détection des multizones, en l'occurrence ici des bizones entre deux volets, est désormais ramené à un problème de traitement d'image et plus précisément de détection de segments de droites.

Les outils dont nous nous servons pour traiter les images font appel à la suite d'outils Pandore [5], élaborée au sein de l'équipe Image du laboratoire GREYC de l'Université de Caen. Ils ont été développés par Régis Clouard. En collaboration avec lui, nous les avons utilisés sur nos objets.

Les étapes de lecture des matrices générées à partir de la phase d'appariement sont les suivantes :

1. sélection des points de l'image initiale qui peuvent entrer dans la composition d'une diagonale. La sélection des points d'intérêt utilisés pour détecter les lignes utilise un seuillage fixe. Un seuil fixe est possible ici, car les images sont des images artificielles. La valeur seuil a été fixée de façon empirique à celui le plus proche de la perception humaine. Les niveaux de gris vont de 0 à 255, nous ne conservons que ceux au dessus de 127. Il reste ici beaucoup de points candidats ;
2. utilisation de la transformée de Hough qui retourne la droite qui contient le plus de points de l'image précédente ;
3. dilatation de cette droite pour avoir une épaisseur de 3 pixels, soit 3 segments de documents afin de palier les micro décrochements de diagonale ;
4. filtrage des points de l'image initiale pour ne garder que les points sous la droite dilatée ;
5. mise en relation des points qui ont une distance inférieure à une distance minimum donnée en paramètre pour construire le plus grand segment de droite possible ;
6. conservation du segment de droite le plus long ;
7. suppression des points de l'image de points candidats, qui sont couverts par ce segment. On empêche ainsi que ces points entrent en jeu dans une autre diagonale. Nous souhaitons de cette façon obtenir le meilleur recouvrement des zones. Celui dans lequel il n'y a pas de recouvrement des segments et donc des projections. Les multizones se contraignent mutuellement ;
8. réitération de ce processus jusqu'à épuisement des candidats, c'est-à-dire jusqu'à ne plus trouver de diagonale suffisamment longue pour être pertinente. La longueur minimum est fixée à 8 pixels.

5. http://www.greyc.ensicaen.fr/~regis/Pandore/index-fr.html



La recherche de segments de droites est guidée par un modèle. Seules les droites avec au maximum un angle entre +25˚et -25˚par rapport à la diagonale ont été considérées. Nous utilisons deux méthodes de détection des segments de droites. La première fortement contrainte présuppose le parallélisme. Elle permet de détecter des segments de droites ayant la même inclinaison que la diagonale parfaite, une inclinaison de 45˚. Nous l'appelons la méthode « petit angle ». En cas de détection insuffisante avec la première, nous utilisons la deuxième méthode qui offre une relaxation des contraintes. Elle permet d'étendre l'espace de recherche aux segments de droites ayant une inclinaison située entre +25˚et -25˚par rapport à la diagonale. Nous l'appelons la méthode « grand angle ».

Ainsi, la première méthode nous permet de détecter les volets à la fois *quasi-synchrones et quasi-bijectifs* dans lesquels globalement ce qui est présent dans l'un l'est dans l'autre et dans le même ordre, et les volets *asynchrones*, c'est-à-dire les volets présentant le même contenu mais avec des différences d'ordre notables dans la structure. La seconde permet, quant à elle, l'identification de volets globalement dans le même ordre mais avec une différence de contenu. Il s'agit de volets *synchrones non bijectifs*, présentant une ou des zones supprimées (ou ajoutées) d'un volet à l'autre volet. Ainsi, à ce stade, nous ne prenons pas en charge le cas de figure de deux volets différents à la fois du point de vue de l'ordre et du contenu.

La taille des matrices que nous créons à partir de notre corpus de communiqués de presse est de 200×200. Il est évidemment possible de changer la taille de l'image, notamment pour traiter des documents plus longs, mais il faut que le contenu soit toujours à la même échelle (notamment, la distance minimale entre les points d'un même segment de droite, l'épaisseur des segments de droites). Le programme fonctionne avec des a priori sur la taille des objets à l'intérieur (points, lignes), mais pas avec les dimensions de l'image.

6.2.3 *Diagnostic de parallélisme*

À l'issue du traitement présenté dans la section 6.2.2, nous disposons d'images sur lesquelles les segments de droites sont mis en évidence par des ellipses (en rouge dans les images du tableau 18, page 83) et leur longueur projetée sur les axes correspondant à chacune des deux langues (en bleu dans les images du tableau 20, page 90). L'analyse de ces matrices nous fournit les informations chiffrées suivantes :

- le nombre total de segments de droites découverts ;
- la longueur totale des segments de droites découverts ;
- les coordonnées des segments de droites découverts ;
- le nombre de segments de droites situés sur la diagonale ;
- la longueur totale des segments de droites situés sur la diagonale ;
- le nombre de segments de droites situés hors de la diagonale ;

- la longueur totale des segments de droites situés hors de la diagonale ;
- la longueur des projections de ces segments de droites dans chacune des langues ;
- la longueur de la diagonale ;
- le ratio (longueur des segments de droites détectés/longueur de la diagonale).

Ces informations nous servent à établir un diagnostic de parallélisme entre les volets représentés. Ce diagnostic de synchronicité permet de reconnaître trois types de bi-documents synchrones, asynchrones avec inversion ou asynchrones avec suppression ou indéfinis. Ainsi, en fonction de la longueur et de la position des segments de droites découverts, nous établissons un diagnostic de synchronicité entre les volets. Si la différence de longueur en valeur absolue entre un des segments de droites détectés pour un des volets (dimension x) est supérieure à 3 par rapport à son équivalent dans l'autre volet (dimension y), nous reconnaissons ce bi-document comme asynchrone avec suppression. Á partir des coordonnées (x,y)(x',y') de chaque segment, nous établissons que si entre deux segments consécutifs x_n est inférieur à y_{n-1}-1, alors nous sommes face à un bi-document asynchrone avec inversion. Enfin, si la longueur totale des segments de droites détectés est inférieure à 20% de la diagonale, nous ne nous prononçons pas sur la nature du parallélisme qui lie les volets observés. Ces documents font alors l'objet d'un nouveau traitement. Plusieurs solutions sont à notre disposition : utiliser la méthode « Grand angle » présentée dans la section 6.2.2, les plonger dans une nouvelle collection plus grande ou thématiquement homogène ou changer la taille de la matrice. Dans les autres cas, le bi-document est reconnu comme synchrone.

Retour aux textes

Les coordonnées des segments calculés à partir des matrices nous servent, quant à elles, à réaliser un retour aux volets, autrement dit à visualiser les multizones ainsi détectées, les alignements de zones. Á ce stade, nous pouvons d'ores et déjà mentionner un des effets de la méthode. La méthode repère correctement des cœurs de zones mais moins bien les frontières. Les frontières de zones peuvent présenter un décalage de plusieurs caractères, voir plusieurs mots. Ceci s'explique par le fait que nous utilisons ici les coordonnées des segments compris dans les segments de droites détectées et non les coordonnées des N-grammes appariés se situant à l'intérieur.
Nous présentons dans la section 7.4 des retours aux textes sur des documents asynchrones correctement attribués par notre système.

Ellipses et projections

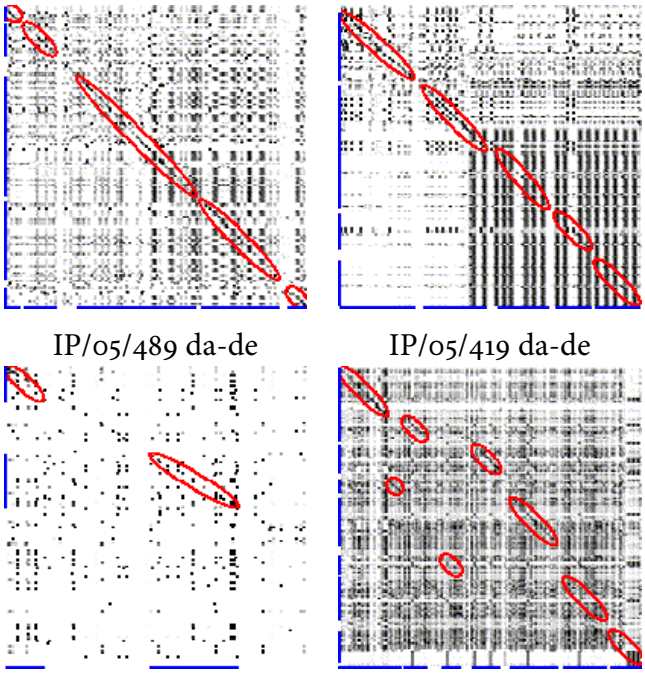


TABLEAU 20 – Ellipses et projections des segments de droites sur les axes des multidocuments

Ce chapitre nous a permis de décrire les étapes d'appariement et de construction de nos matrices. Dans le chapitre 7, nous en faisons l'évaluation sur la tâche d'alignement de zones de documents traduits. Le chapitre 7 présente les résultats que nous obtenons en matière de diagnostic de parallélisme sur plusieurs collections de multidocuments. Nous y présentons également les domaines de validité de notre méthode.

7

RÉSULTATS ET ÉVALUATION SUR LA TÂCHE D'ALIGNEMENT DE ZONES

Dans ce chapitre, nous allons éprouver les modèles de traductions attendus définis dans le chapitre 5 dans plusieurs dimensions. Pour cela, nous faisons varier les dimensions suivantes :

- proximité des langues ;
- collection de multidocuments thématiquement proches ou non ;
- multidocuments avec ou sans leur mise en forme matérielle.

Ainsi nous définissons les domaines de validité de notre méthode de détection et d'alignement de zones.

SOMMAIRE

7.1 MODÈLES ET IMAGES OBTENUES

7.1.1 *Modèles envisagés et images obtenues*

Dans cette section, nous comparons l'attendu que nous avions en matière de visualisation de phénomènes textuels entre des volets de multidocuments comparés deux à deux. Les images que nous obtenons sont en accord avec les modèles proposés au chapitre 5. Pour illustration, nous mettons les images obtenues et les modèles envisagés en vis-à-vis dans le tableau 21.

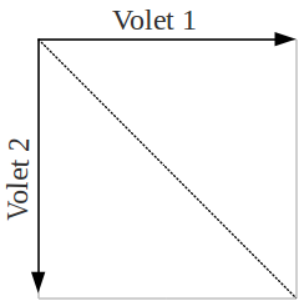
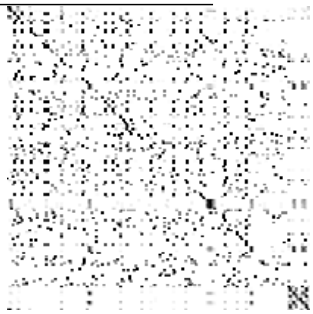
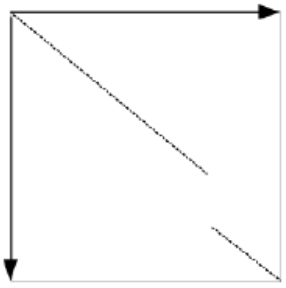
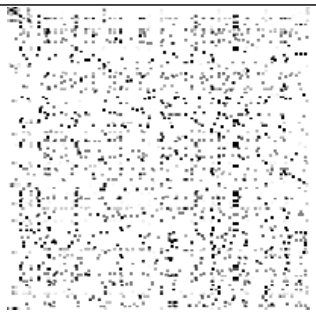
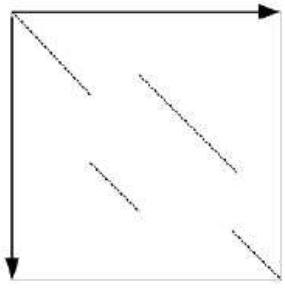
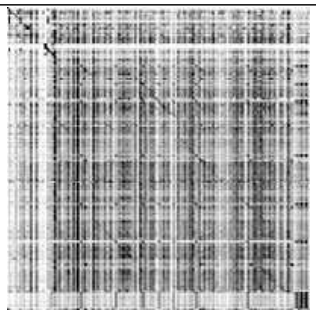
Modèles	Matrices
IP/05/1451 el-fr : Volets synchrones	
	
IP/05/473 en-fr : Volets asynchrones avec suppression	
	
IP/05/1157 en-fr : Volets asynchrones avec inversions locales	
	

TABLEAU 21 – Panel des matrices obtenues en vis à vis avec les modèles définis au chapitre 5.

7.1.2 *Images obtenues et émergence d'un nouveau modèle*

En observant nos matrices à l'œil nu, nous avons constaté l'existence d'un motif récurrent, une sorte de matrice dans la matrice. En retournant aux documents, nous avons constaté que ce motif décrivait des zones de textes dans une autre langue que les deux attendues, dans au moins un des deux volets. Pour des raisons fortuites ou structurelles, oubli ou défaut de traducteurs, des zones de textes de certains volets n'ont pas fait l'objet d'une traduction. Au travers de la collection, le volume de traduction de chaque volet diffère. Contrairement à nos attentes, les volets ne sont pas tous monolingues. Nous illustrons ce nouveau cas de figure dans le tableau 22 par deux exemples de communiqués de presse,

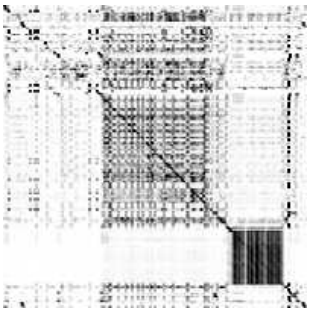
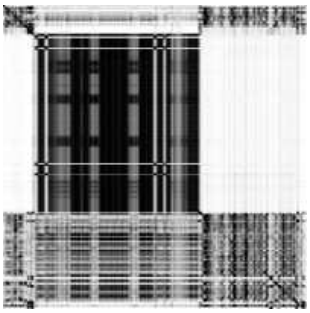
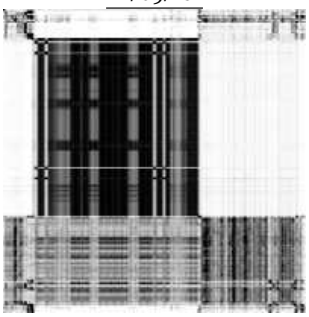
Cas de volets multilingues		
IP/05/182		
		
Volets en-fr	Volets de-fr	Volets da-de
IP/05/181		
		
Volets en-fr	Volets de-fr	Volets el-fr

TABLEAU 22 – Nouveau modèle : cas de multilinguisme intra bi-document.

Dans le tableau 22, le multidocument IP/05/181 [1], par exemple, se compose d'un volet anglais monolingue (en), d'un volet français bilingue présentant deux zones en français, l'introduction et les annexes, séparées par un tableau en anglais (fr-en-fr) et tous les autres volets sont

1. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/181&format=HTML&aged=1&language=ES&guiLanguage=en

bilingues (autres langues, en, en). Le multidocument IP/05/182 [2], quant à lui, comprend entres autres : un volet anglais terminant par une zone en français, un volet français alternant français-anglais-français, des volets danois et allemand trilingues, respectivement danois-anglais-français et allemand-anglais-français.

Les carrés visibles au centre de ces images reflètent des zones avec une forte densité de liens. Des zones non traduites entre deux documents présentent naturellement beaucoup plus d'alignements qu'entre des zones traduites.

Ce phénomène de non tradution n'est pas marginal, nous l'avons constaté sur plusieurs dizaines de multidocuments de nos collections. Une identification automatique de ce cas laisse envisager des opérations de contrôle a posteriori des traductions.

7.2 RÉPARTITIONS DES DIFFÉRENTS DIAGNOSTICS SUR LES COLLECTIONS

7.2.1 *Corpus d'évaluation*

Dans cette section, nous présentons les résultats obtenus sur 6 collections de 40 multidocuments en 7 langues (cf.chapitre 3). Ces collections sont tirées de l'ensemble des communiqués de presse de l'Union Européenne entre 2004 et 2009. 213 multidocuments différents observés au total, certains multidocuments faisant partie de plusieurs collections.

- Collection 1, 2 et 3 : Après une identification sur le corpus complet des documents disponibles dans les 7 langues que nous souhaitions traiter, nous avons constitué des multidocuments de 7 langues chacun, 495 en tout. Les multidocuments sont donc équilibrés du point de vue des langues. Pour constituer les collections de 40 multidocuments nous avons regroupé dans des dossiers des multidocuments par paquets de 40, au fil de leur numérotation (collection 1 : md 1 à md 40, collection 2 : md 41 à 80…) ;
- Collections « transport », « santé » et « téléphone ». Une des stratégies utilisée pour améliorer la qualité des matrices est de plonger les multidocuments non diagnostiqués dans des collections de documents thématiquement proches. L'idée est de maximiser les chances de rencontrer des correspondances bi- ou quasi- univoques. Les collections « transport », « santé » et « téléphone » ont été constituées en exploitant des expressions régulières sur les mots des thèmes voulus en français.

Les collections 1, 2 et 3 ont été traitées *avec* et *sans* leur mise en forme matérielle afin de mesurer l'impact de la mise en forme sur nos

2. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/182&format=HTML&aged=1&language=EN&guiLanguage=en

résultats. Une synthèse des résultats sur les 6 collections est présentée dans le tableau 23.

Les diagnostics sont bilingues. Ils sont réalisés sur les couples de langues suivants : fr-de, fr-el, fr-en, fr-es, fr-fi et de-da. Ces couples nous permettent de constater l'impact de la proximité des langues sur le diagnostic. Ainsi pour chaque collection, nous obtenons 240 matrices : 6 couples × 40 mds.

Chaque collection a été analysée avec les deux méthodes : « Petit angle » et « Grand angle ». Ainsi 1440 correspond au total des collections 1, 2 et 3, soit 240 matrices × 3 collections × 2 méthodes.

7.2.2 *Synthèse des résultats sur notre corpus d'évaluation*

Une synthèse des résultats obtenus en matière de diagnostic de synchronicité des multidocuments est présentée dans le tableau 23. Ce tableau présente les résultats obtenus par chacune des deux méthodes *Petit Angle* et *Grand Angle* sur l'ensemble des collections. Ceci permet d'évaluer la capacité de chacune des deux méthodes à décider, étant entendu que la méthode Grand Angle n'est pas prévue pour diagnostiquer les cas d'inversion et de synchronicité.

Ce tableau montre que le taux de décision est important, partant de 64% pour les collections traitées sans leur mise en forme matérielle et allant jusqu'à 97% sur les documents dans des langues proches exploités avec leur mise en forme. Ces résultats nous permettent de valider nos hypothèses de départ :

- l'appariement entre des langues proches donne de meilleurs résultats que celui entre des langues éloignées. Les taux de décisions entre ces deux contextes présentent un écart de 13% sur le total des collections 1, 2 et 3. L'usage du lexique est différent d'une langue à l'autre. Le finnois par exemple comportera beaucoup plus d'occurrences que son équivalent en français qui sera alternativement remplacé, ici par un synonyme, ici par un pronom… En d'autres termes, nous aurons plus de difficultés à apparier des langues différentes de ce point de vue là. Les différences morphologiques étant, quant à elles, lissées par l'usage des N-grammes de caractères qui permet de traiter par la même méthode des langues riches ou pauvres morphologiquement ;
- analyser un multidocument par le prisme d'une collection de multidocuments thématiquement proches améliore également les résultats jusqu'à +3% de décisions prises. Ceci s'explique par la diminution du nombre d'hapax par document. Un hapax de document pourra être répété à d'autres endroits de la collection, ce qui nous permet obtenir les informations nécessaires à son appariement, et à son alignement ultérieur ;
- traiter les documents avec leur mise en forme donne lieu à de meilleurs résultats : +10% de décisions prises. Le parti pris original

		total	Indécisions		Décisions		Petit Angle (PA)					Grand Angle (GA)				
							Décisions PA	Synchrones	Asynchrones (avec inv)	Asynchrones (avec sup)	Indécisions	Décisions GA	Synchrones	Asynchrones (avec inv)	Asynchrones (avec sup)	Indécisions
Collections avec Mise en Forme Matérielle	**Corpus :**	**1440**	**122**	**8,47%**	**1318**	**91,53%**	**665**	**601**	**41**	**23**	**55**	**393**	**333**	**60**	**260**	**67**
	Collection 1 :	480	39	8,13%	441	91,88%	223	201	12	10	17	218	120	22	76	22
	Collection 2 :	480	52	10,83%	428	89,17%	212	187	18	7	28	216	105	23	88	24
	Collection 3 :	480	31	6,46%	449	93,54%	230	213	11	6	10	219	108	15	96	21
	Couples proches :	**720**	**21**	**2,92%**	**699**	**97,08%**	**352**	**317**	**15**	20	**8**	**347**	**197**	14	**136**	**13**
	fr-es	240	7	2,92%	233	97,08%	118	108	4	6	2	115	74	1	40	5
	fr-en	240	5	2,08%	235	97,92%	118	106	5	7	2	117	65	5	47	3
	de-da	240	9	3,75%	231	96,25%	116	103	6	7	4	115	58	8	49	5
	Couples éloignés :	**720**	**101**	**14,03%**	**619**	**85,97%**	**313**	**284**	**26**	3	**47**	**306**	**136**	46	**124**	**54**
	fr-el	240	40	16,67%	200	83,33%	99	89	10	0	21	101	51	16	34	19
	fr-de	240	25	10,42%	215	89,58%	109	100	7	2	11	106	50	11	45	14
	fr-fi	240	36	15,00%	204	85,00%	105	95	9	1	15	99	35	19	45	21
Collections avec Mise en Forme Matérielle	**Collection théma. :**	**1440**	**105**	**7,29%**	**1335**	**92,71%**	**671**	**603**	**46**	**22**	**49**	**664**	**277**	**82**	**305**	**56**
	Collection transport :	480	38	7,92%	442	92,08%	222	199	18	5	18	220	103	**27**	**90**	20
	Collection santé :	480	52	10,83%	428	89,17%	220	200	13	7	20	208	89	**25**	**94**	32
	Collection téléphone :	480	15	3,13%	465	96,88%	229	204	15	10	11	236	85	**30**	**121**	4
	Couples proches :	**240**	**20**	**2,85%**	**682**	**97,15%**	**354**	**316**	**18**	20	**6**	**346**	**167**	18	**161**	**14**
	fr-es	240	4	1,67%	236	98,33%	119	108	3	8	1	117	70	**2**	**45**	3
	fr-en	240	4	1,67%	236	98,33%	119	101	10	8	1	117	47	**11**	**59**	3
	de-da	240	12	5,00%	228	95,00%	116	107	5	4	4	112	50	**5**	**57**	8
	Couples éloignés :	**240**	**85**	**11,81%**	**635**	**88,19%**	**317**	**287**	**28**	2	**43**	**318**	**110**	64	**144**	**42**
	fr-el	240	29	12,08%	211	87,92%	107	98	9	0	13	104	40	**21**	**43**	16
	fr-de	240	21	8,75%	219	91,25%	110	101	8	1	10	109	35	**20**	**54**	11
	fr-fi	240	35	14,58%	205	85,42%	100	88	11	1	20	105	35	**23**	**47**	15
Collections sans MFM	**Corpus :**	**1440**	**338**	**25,13%**	**1102**	**74,87%**	**566**	**504**	**56**	**6**	**154**	**536**	**210**	**95**	**231**	**184**
	Collection 1 :	480	157	32,71%	323	67,29%	184	164	17	3	56	139	66	**36**	**73**	65
	Collection 2 :	480	150	31,25%	330	68,75%	186	169	16	1	54	144	68	**31**	**76**	65
	Collection 3 :	480	126	26,25%	354	73,75%	196	171	23	2	44	158	76	**28**	**82**	54
	Couples proches :	**720**	**81**	**11,25%**	**639**	**88,75%**	**327**	**306**	**17**	4	**33**	**312**	**148**	23	**141**	**48**
	fr-es	240	14	5,83%	226	94,17%	113	106	4	3	7	113	60	**5**	**48**	7
	fr-en	240	23	9,58%	217	90,42%	112	106	5	1	8	105	46	**9**	**50**	15
	de-da	240	44	18,33%	196	81,67%	102	94	8	0	18	94	42	**9**	**43**	26
	Couples éloignés :	**720**	**257**	**35,69%**	**463**	**64,31%**	**239**	**198**	**39**	2	**121**	**224**	**62**	72	**90**	**136**
	fr-el	240	93	38,75%	147	61,25%	74	61	11	2	46	73	23	**21**	**29**	47
	fr-de	240	69	28,75%	171	71,25%	90	79	11	0	30	81	26	**19**	**36**	39
	fr-fi	240	95	39,58%	145	60,42%	75	58	17	0	45	70	13	**32**	**25**	50

TABLEAU 23 – Synthèse des diagnostics obtenus sur plusieurs collections de multidocuments. Ils sont présentés en fonction du type de collections, de la méthode employée et des couples de langues observés.

de prendre en charge les documents avec cette mise en forme et de traiter la structure et le contenu par la même méthode ajoute visiblement des informations supplémentaires, assimilables à des cognats.

7.3 ÉVALUATION ET DISCUSSION DES RÉSULTATS

Évaluer ces résultats n'est pas une tâche triviale. Il n'existe pas de références pour évaluer la détection de multizones. La réalisation manuelle de cette référence est une tâche sinon subjective, au moins fastidieuse. À une collection, telles que nous les constituons, correspondent 240 bi-documents. Nous présentons dans les tableaux[3] 24 et 25 les mesures de précision, rappel et F-mesure obtenues à partir d'une référence constituée pour les collections 1,2,3 d'une part et sur les trois collections thématiques constituées à partir de notre corpus d'autre part. Une étude qualitative et quantitative des différents types de parallélisme entre les volets des différents bi-documents est fournie dans l'annexe B.

	Petit Angle				**Grand Angle**			
	Synchrones	Asynchrones		Total	Synchrones	Asynchrones		Total
		avec inversion	avec suppression			avec inversion	avec suppression	
Obtenus	601	41	23	665	333	60	260	653
Attendus	652	19	49	720	652	19	49	720
Correctement attribués	554	6	0	560	325	7	26	358
Précision	92,18%	14,63%	0,00%	84,21%	97,60%	11,67%	10,00%	54,82%
Rappel	84,97%	31,58%	0,00%	77,78%	49,85%	36,84%	53,06%	49,93%
F-mesure	88,43%	20,00%	0,00%	80,87%	65,99%	17,72%	16,83%	52,26%

TABLEAU 24 – Mesures de précision, rappel et F-mesure sur les collections 1,2,3 avec leur MFM. La référence sur les 720 bi-documents a été réalisée par nos soins.

	Petit Angle				**Grand Angle**			
	Synchrones	Asynchrones		Total	Synchrones	Asynchrones		Total
		avec inversion	avec suppression			avec inversion	avec suppression	
Obtenus	603	46	22	671	277	82	305	664
Attendus	678	16	26	720	678	16	26	720
Correctement attribués	572	5	0	577	270	12	12	294
Précision	94,86%	10,87%	0,00%	85,99%	97,47%	14,63%	3,93%	44,28%
Rappel	84,37%	31,25%	0,00%	80,14%	39,82%	75,00%	46,15%	40,83%
F-mesure	89,31%	16,13%	0,00%	82,96%	56,54%	24,49%	7,25%	42,49%

TABLEAU 25 – Mesures de précision, rappel et F-mesure sur les collections transport, santé et téléphone avec leur MFM. La référence sur les 720 bi-documents a été réalisée par nos soins.

3. Les résultats en couleur dans le tableau font chacun l'objet d'une présentation d'une partie des matrices les illustrant et qui ont servi au diagnostic.

Les expériences réalisées sur ces deux séries de collections montrent que la méthode *Petit Angle* offre un rappel entre 77 et 80% pour une précision entre 84 et 86%. La méthode *Grand Angle*, quant à elle, obtient un rappel entre 40 et 49% pour une précision entre 40 et 44%. Il faut rappeler à sa décharge que cette dernière méthode n'est pas prévue pour détecter les documents synchrones ou avec inversion. Si ses résultats sur les bi-documents avec inversion dépassent nos attentes en atteignant jusqu'à 44% de plus que la méthode *Petit Angle*, les résultats pour les documents synchrones correspondent bien eux à l'attendu, +35% de rappel par rapport à la méthode *Grand Angle* dans les deux séries de collections. Pour ce qui est des bi-documents avec suppression, la méthode *Grand Angle* répond bien à nos attentes en obtenant un rappel de 46 à 53%, meilleur pour les collections 1,2,3, contre 0% pour la méthode *Petit Angle*, cependant sa précision s'avère décevante, plafonnant à 10% pour les collections 1,2,3.

7.3.1 *Comparaison avec d'autres modèles*

Comparaison avec le modèle « tout synchrone »

Le tableau 26 donne à titre comparatif les résultats par rapport à une méthode baseline prenant comme hypothèse que tous les documents parallèles sont synchrones dans chacune de nos deux séries de collections.

	Synchrones	
	collections 1,2,3	collections thématiques
Obtenus	720	720
Attendus	652	678
Correctement attribués	652	678
Précision	90,56%	94,17%
Rappel	100%	100%
F-mesure	95,04%	97,00%

TABLEAU 26 – Mesures de précision, rappel et F-mesure sur les collections 1,2,3 et les collections thématiques avec leur MFM suivant l'hypothèse que tous les bi-documents sont synchrones.

Nos résultats sur les documents synchrones sont de 2 à 7% meilleurs que les résultats obtenus par cette méthode baseline.

Comparaison avec le modèle « synchrone par défaut »

Le tableau 27 donne à titre comparatif les résultats par rapport à une méthode considérant par défaut Vergne et Giguet (1998) que les documents parallèles sont synchrones dans chacune de nos deux séries de collections. Ainsi, le nombre de bi-documents synchrones correspond

à la somme des documents que nous avons définis comme étant synchrones et des bi-documents non diagnostiqués par nos deux méthodes dans chacune des deux séries de collections.

	Collections 1,2,3			Collections thématiques		
	Petit Angle	Grand Angle	Total	Petit Angle	Grand Angle	Total
Obtenus	656	400	1056	652	333	985
Attendus	652	652	1304	678	678	1356
Correctement attribués	609	392	1001	621	326	947
Précision	92,84%	98,00%	94,79%	95,25%	97,90%	96,14%
Rappel	93,40%	60,12%	76,76%	91,59%	48,08%	69,84%
F-mesure	93,12%	74,52%	84,83%	93,38%	64,49%	80,91%

TABLEAU 27 – Mesures de précision, rappel et F-mesure sur les collections 1,2,3 collections thématiques avec leur MFM en considérant par défaut les indéfinis comme synchrones.

Notre méthode se comporte aussi bien que si nous avions pris le parti de considérer par défaut les indécisions comme des bi-documents synchrones.

Ainsi, le système s'avère très précis et assez pertinent pour les documents synchrones. Mais les classes sont très déséquilibrées et les résultats sur les documents asynchrones sont moins satisfaisants. Les images liées à ces bi-documents sont présentées dans les tableaux des pages suivantes.

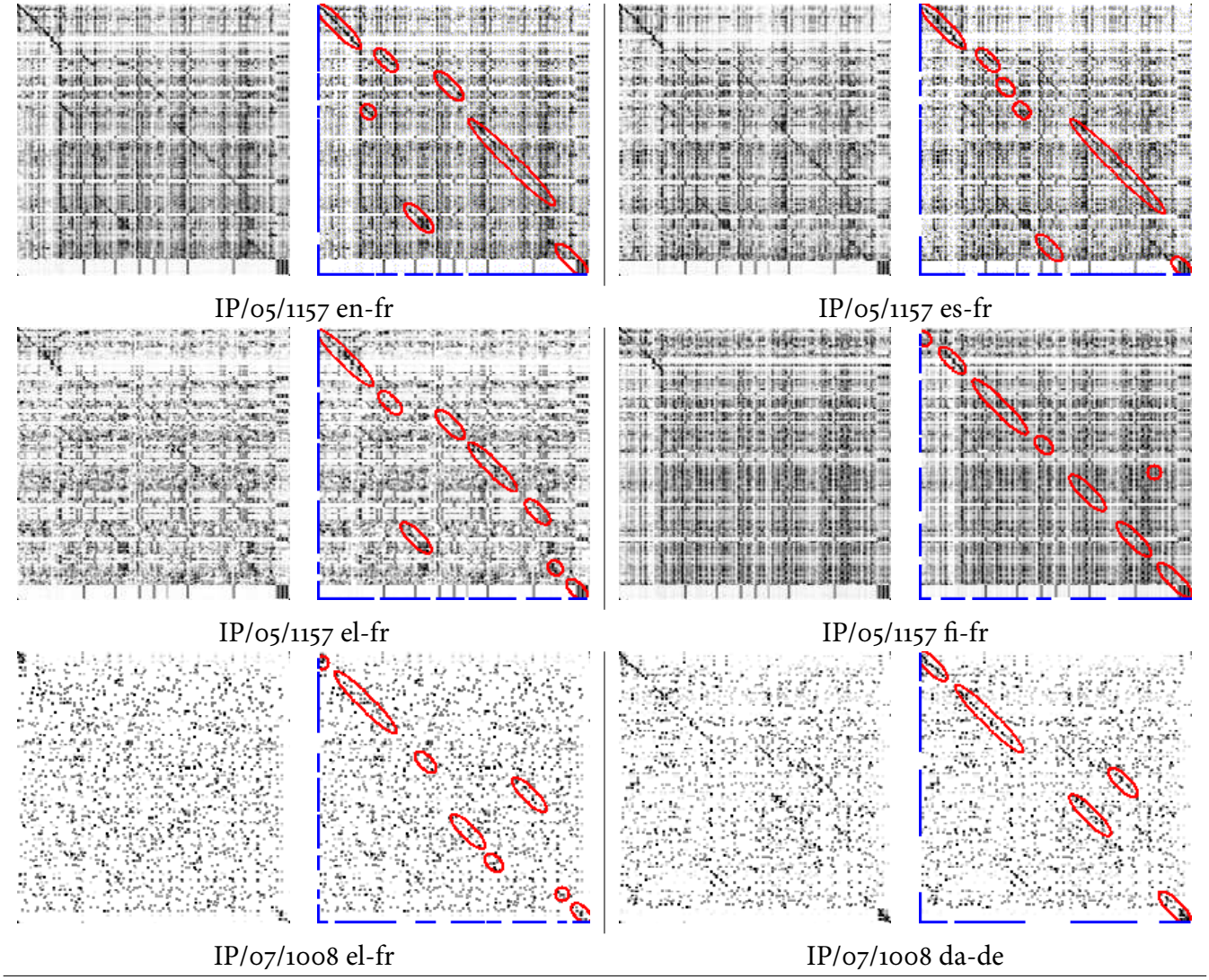
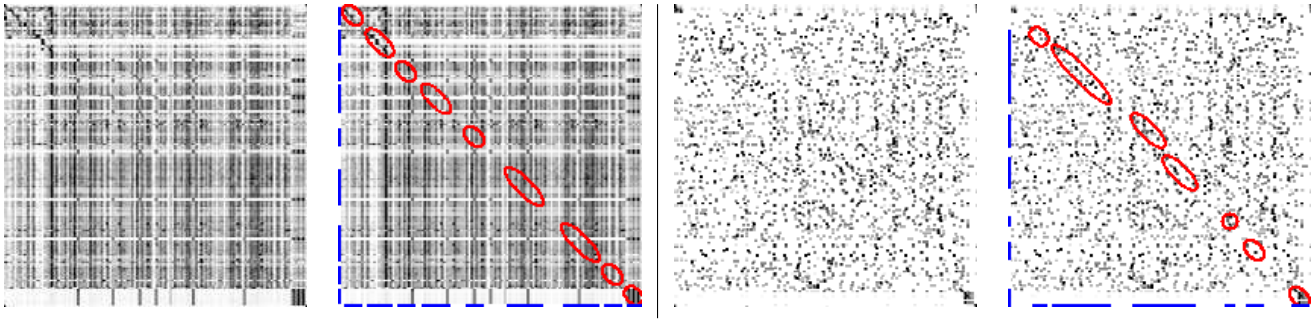
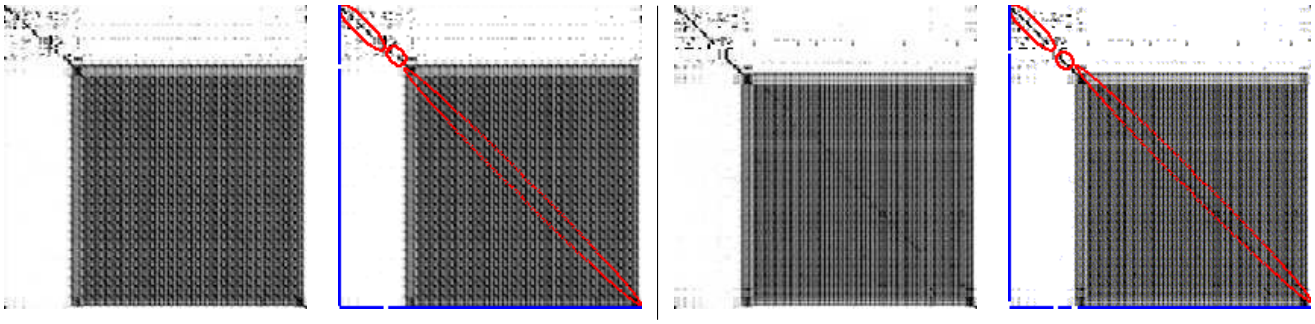


TABLEAU 28 – Les 6 bi-documents asynchrones avec inversion correctement attribués sur les collections 1,2,3 avec la méthode *Petit Angle* (voir tableau 24).



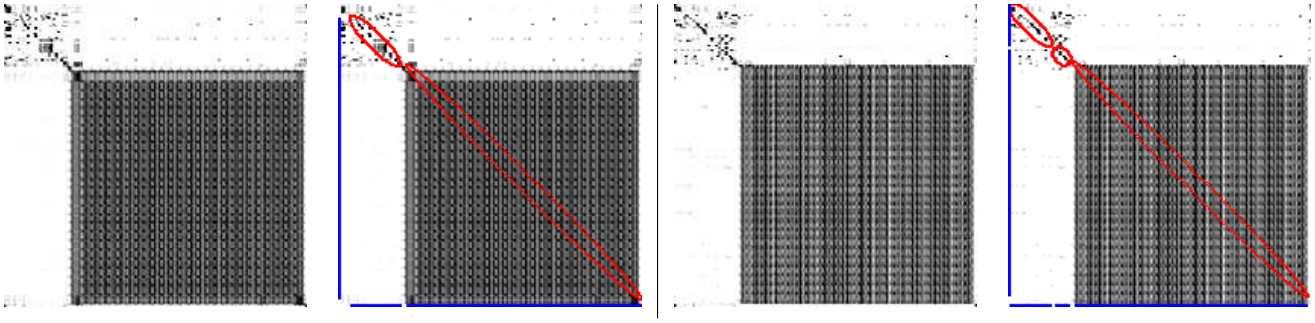
IP/05/1157 de-fr

IP/07/1008 fi-fr



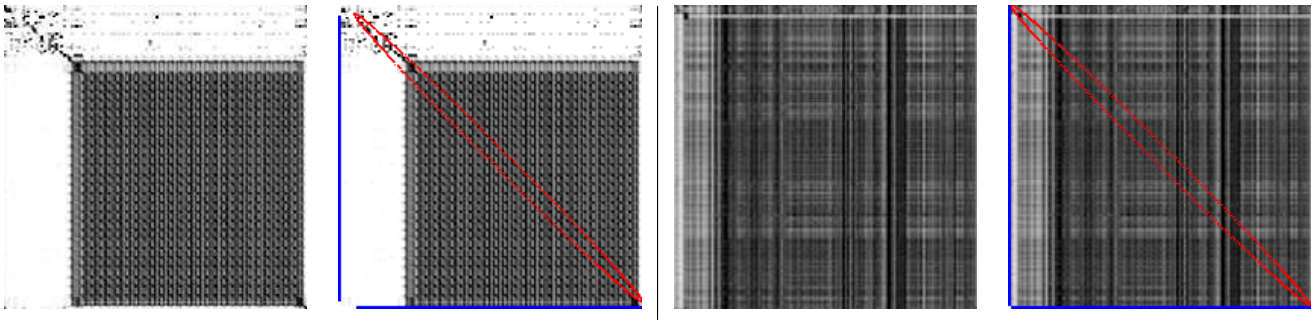
IP/06/1310 en-fr

IP/06/1310 es-fr



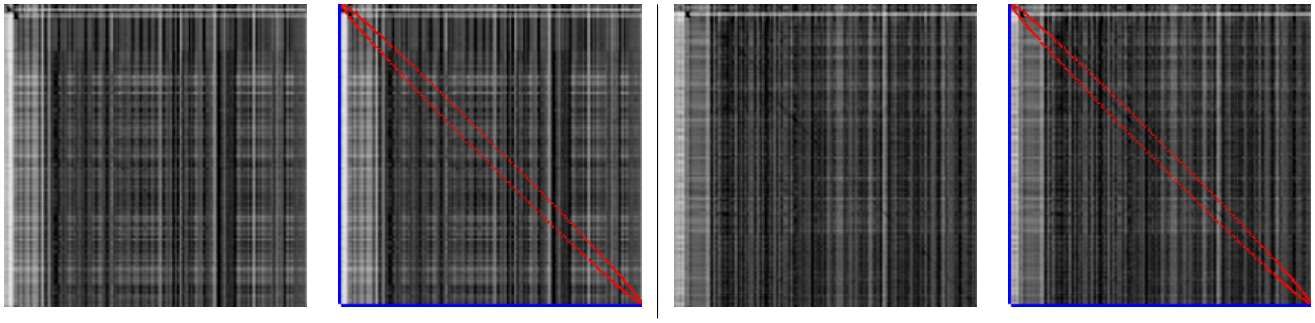
IP/06/1310 el-fr

IP/06/1310 de-fr



IP/06/1310 fi-fr

IP/10/1002 en-fr



IP/10/1002 es-fr

IP/10/1002 da-de

TABLEAU 29 – 10 bi-documents asynchrones avec inversion attendus mais non obtenus parmi les 19 sur les collections 1,2,3 avec la méthode *Petit Angle* (voir tableau 24).

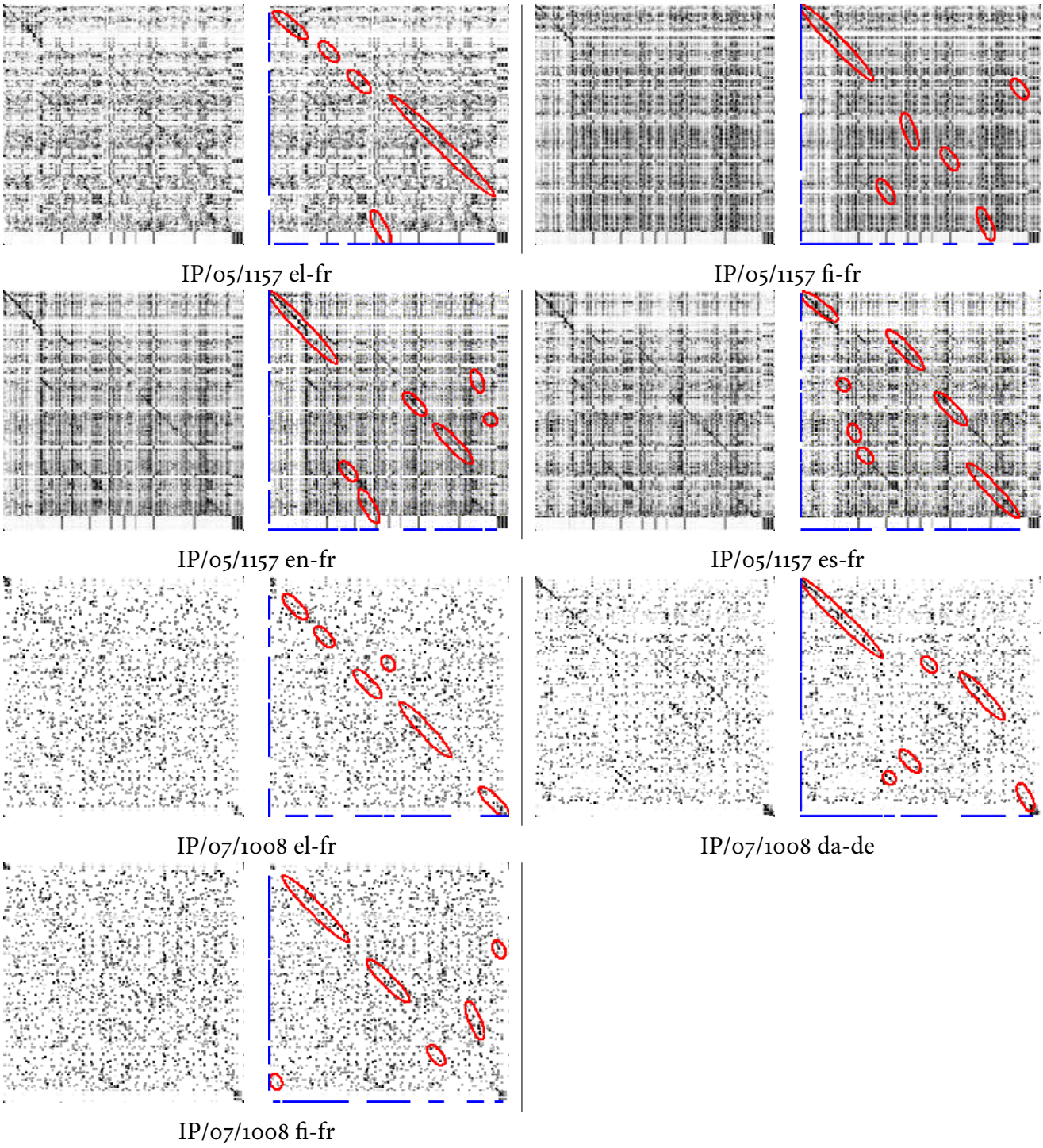


TABLEAU 30 – Les 7 bi-documents asynchrones avec inversion correctement attribués sur les collections 1,2,3 avec la méthode *Grand Angle* (voir tableau 24).

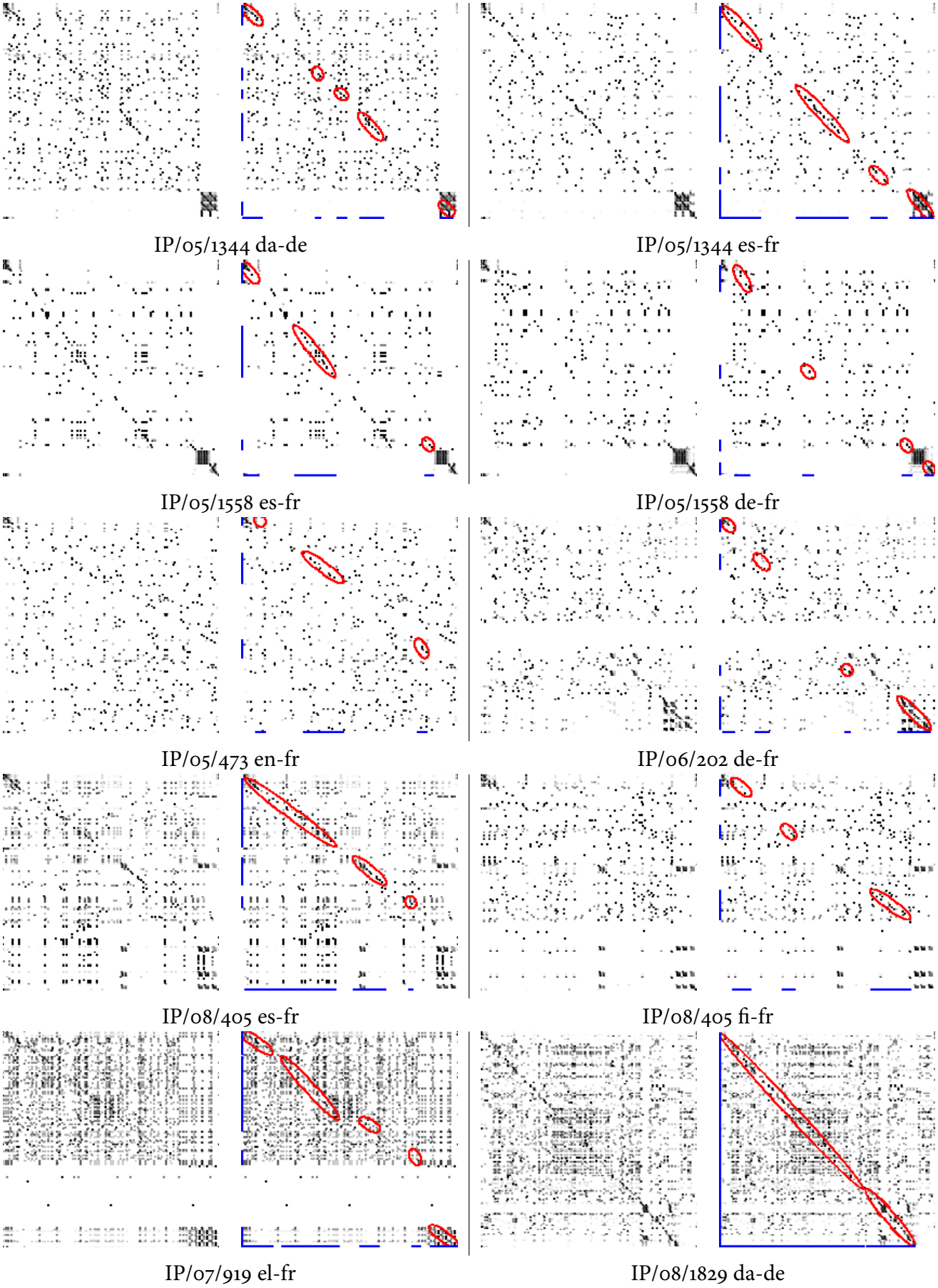
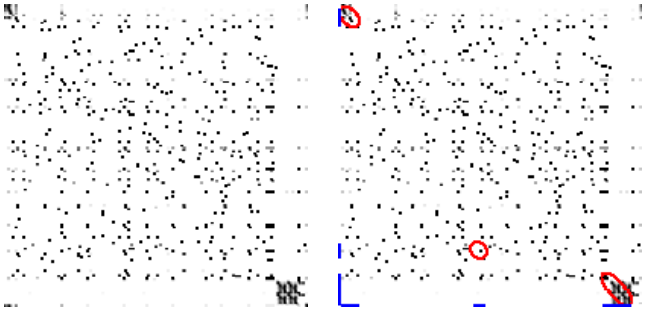
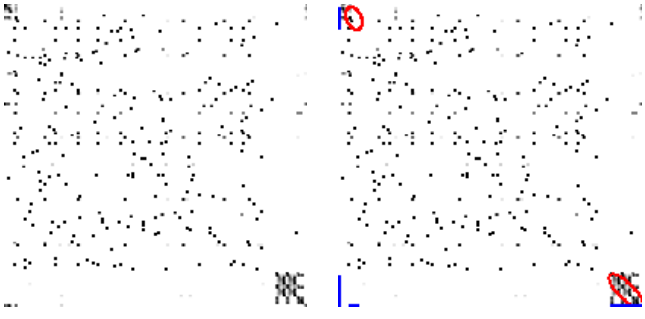


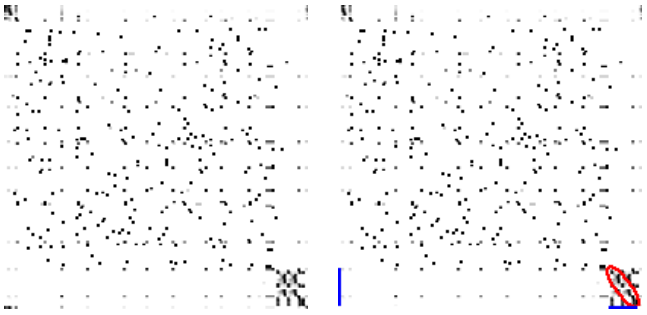
TABLEAU 31 – 10 bi-documents asynchrones avec suppression parmi les 26 correctement attribués sur les collections 1,2,3 avec la méthode *Grand Angle* (voir tableau 24).



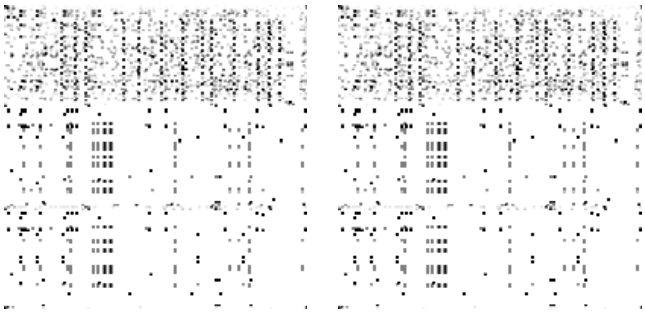
IP/05/1344 de-fr



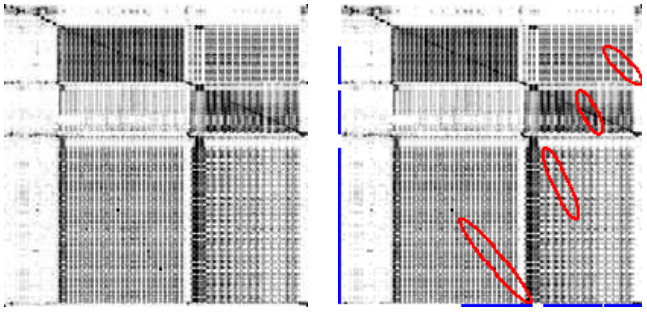
IP/05/1344 el-fr



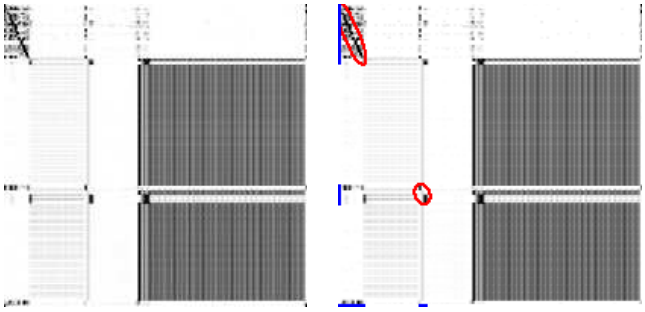
IP/05/1344 fi-fr



IP/05/1653 el-fr

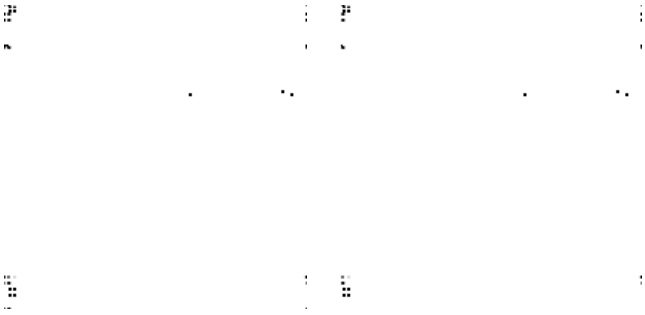


IP/07/1919 da-de

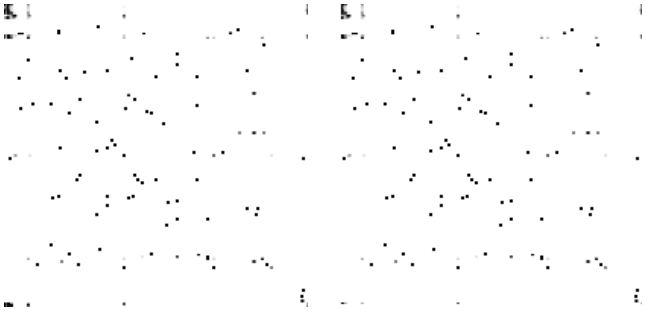


IP/07/1919 de-fr

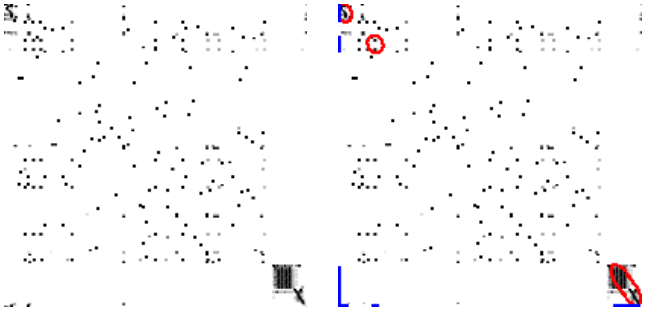
IP/07/1295 es-fr



IP/O6/751 el-fr



IP/05/1558 el-fr



IP/05/1558 fi-fr

TABLEAU 32 – 10 bi-documents asynchrones avec suppression non obtenus parmi les 49 attendus sur les collections 1,2,3 avec la méthode *Grand Angle* (voir tableau 24).

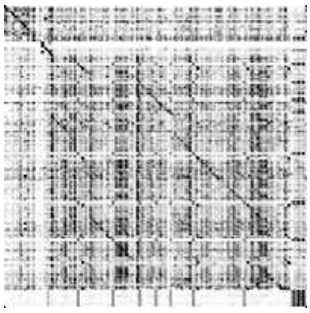
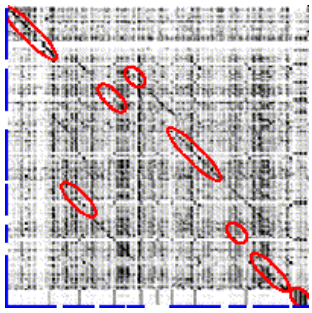
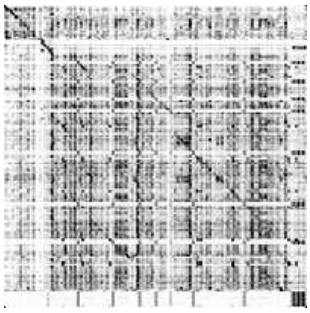
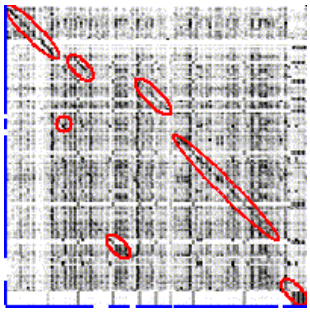
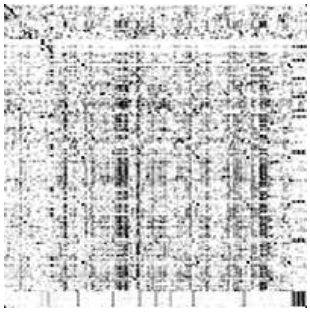
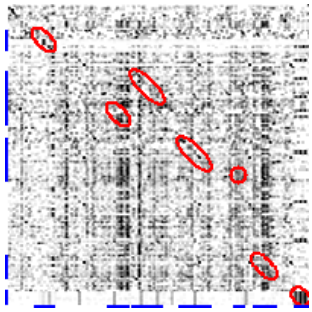
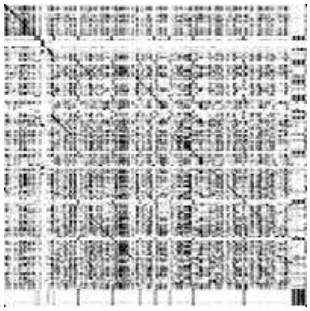
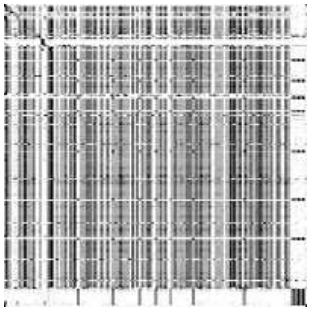
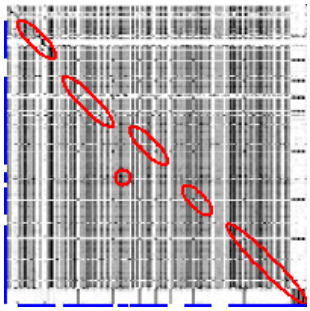
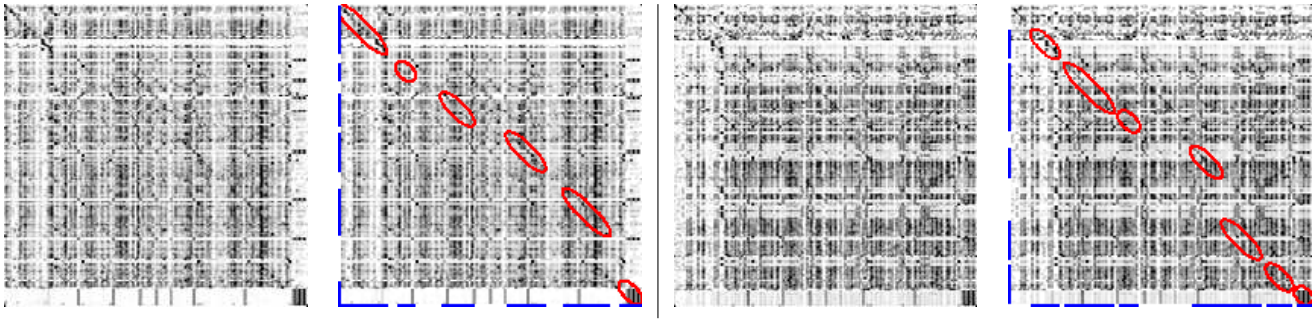
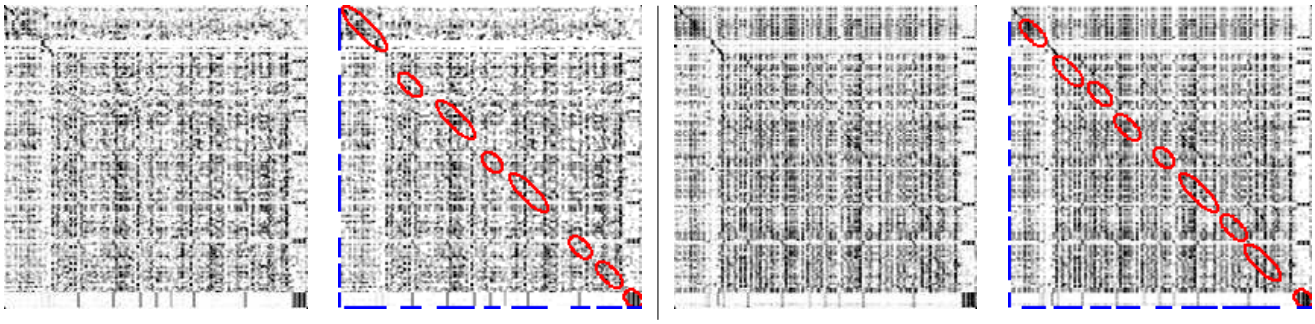
Collection transport	
  IP/05/1157 en-fr	  IP/05/1157 es-fr
  IP/05/1157 fi-fr	
Collection téléphone	
  IP/05/1157 en-fr	  IP/05/1157 de-fr

TABLEAU 33 – Les 5 bi-documents asynchrones avec inversion correctement attribués sur les collections thématiques avec la méthode *Petit Angle* (voir tableau 25).



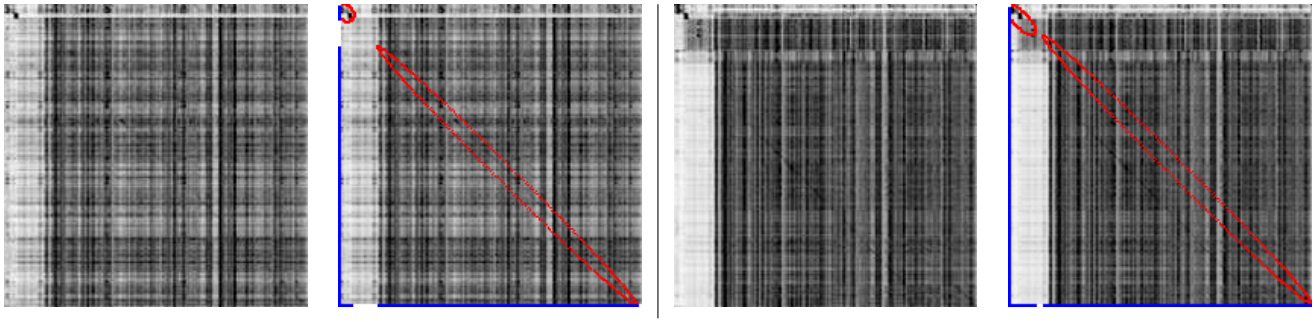
IP/05/1157 de-fr

IP/05/1157 el-fr



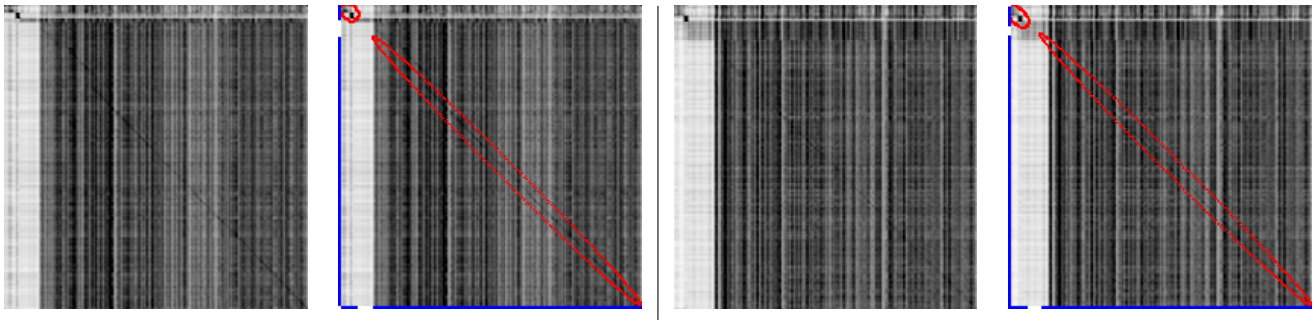
IP/05/1157 fi-fr

IP/05/1157 es-fr



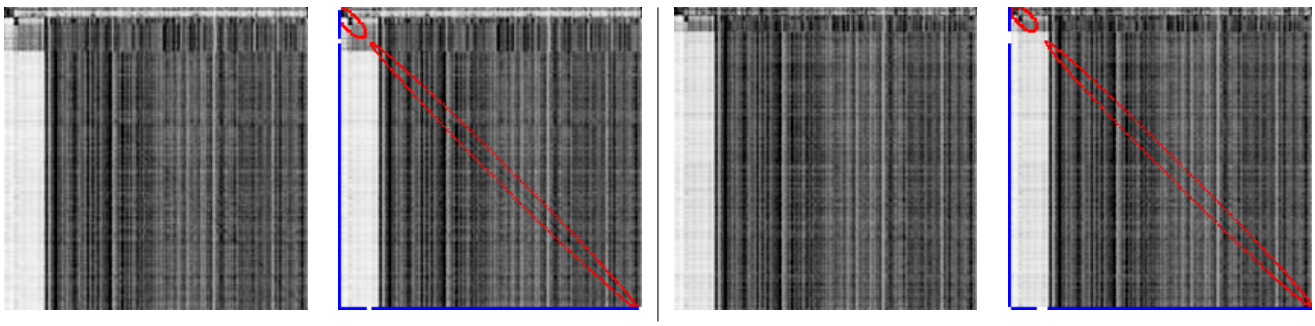
IP/10/1002 en-fr

IP/10/1002 es-fr



IP/10/1002 da-de

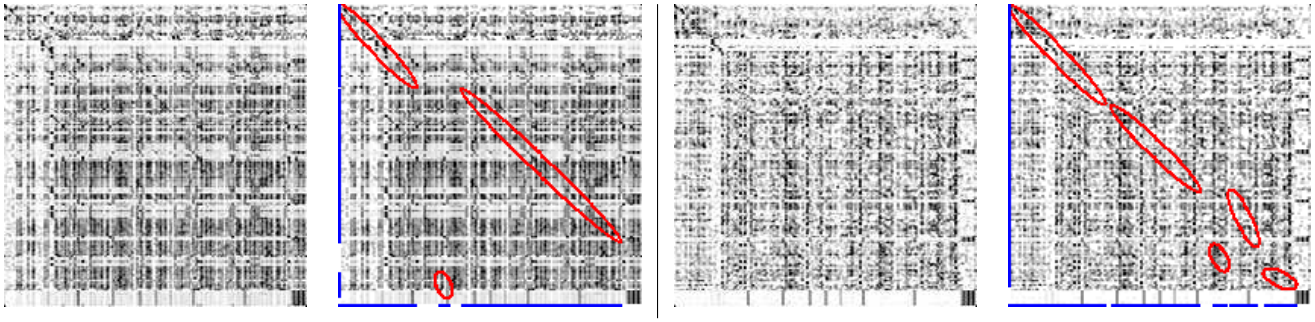
IP/10/1002 de-fr



IP/10/1002 el-fr

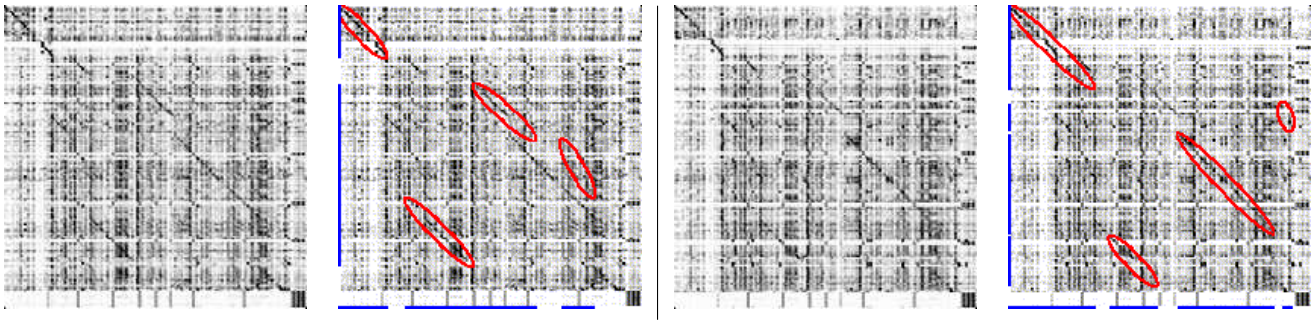
IP/10/1002 fi-fr

TABLEAU 34 – 10 bi-documents asynchrones avec inversion attendus mais non obtenus avec la méthode *Petit Angle* parmi les 16 des collections thématiques (voir tableau 25).



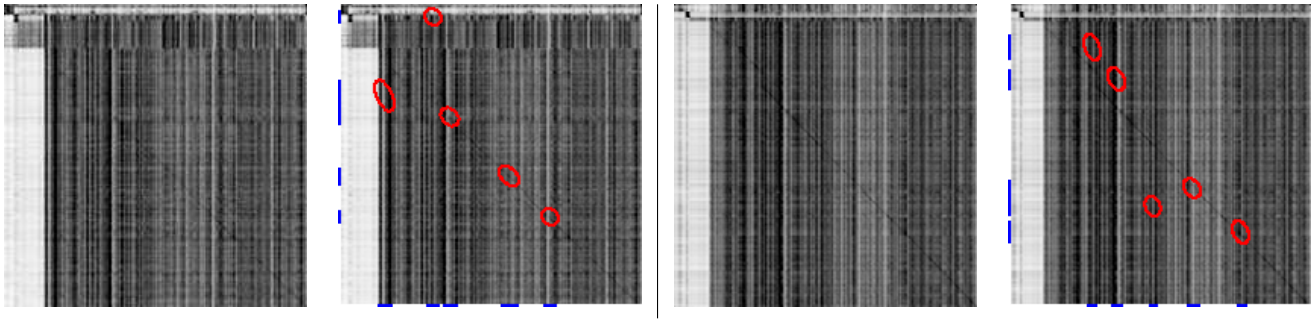
IP/05/1157 el-fr

IP/05/1157 fi-fr



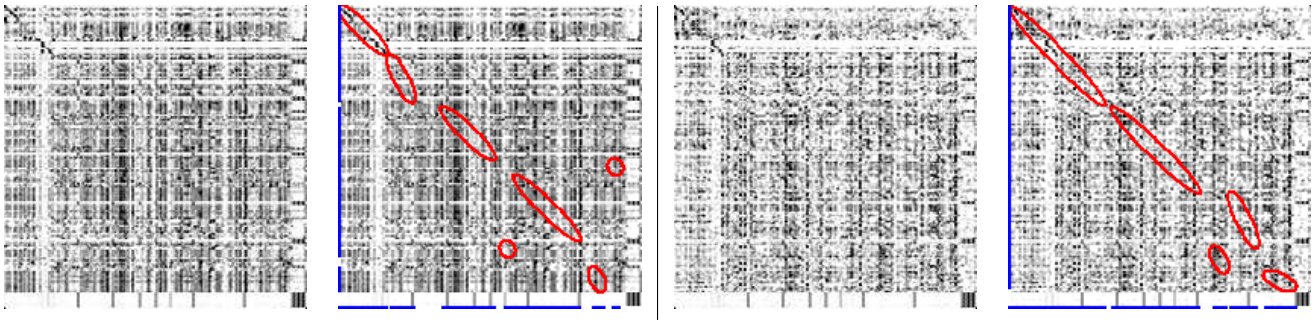
IP/05/1157 en-fr

IP/05/1157 es-fr



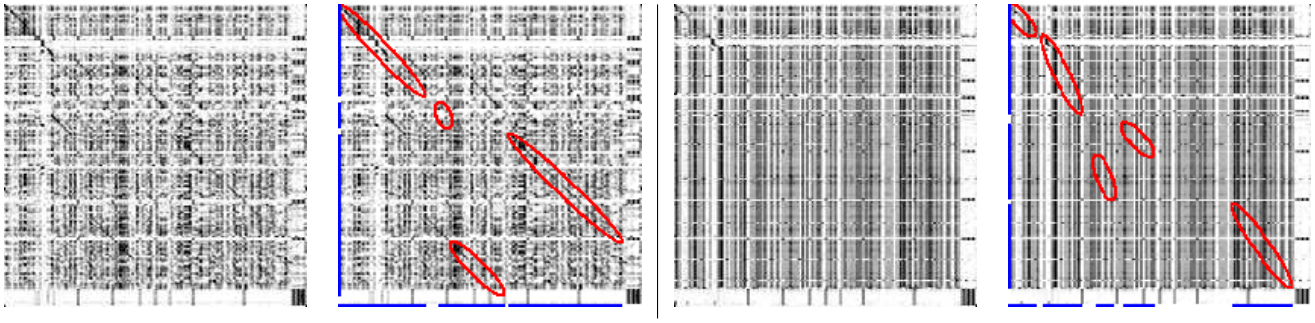
IP/10/1002 el-fr

IP/10/1002 da-de



IP/05/1157 el-fr

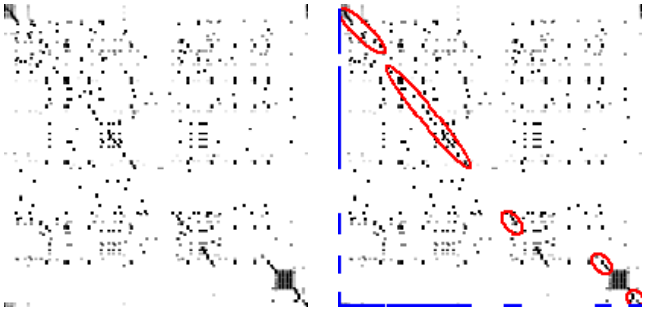
IP/05/1157 fi-fr



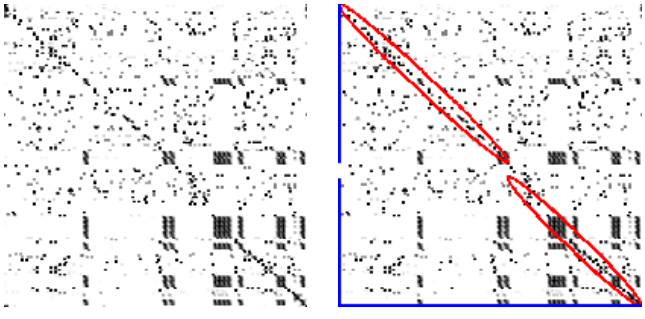
IP/05/1157 en-fr

IP/05/1157 de-fr

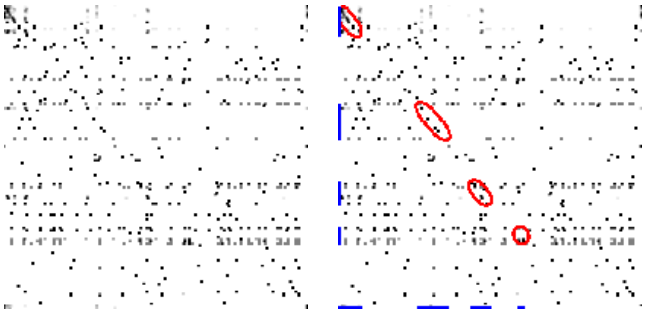
TABLEAU 35 – 10 bi-documents asynchrones avec inversion parmi les 12 correctement attribués sur les collections thématiques avec la méthode *Grand Angle* (voir tableau 24).



IP/05/1558 es-fr

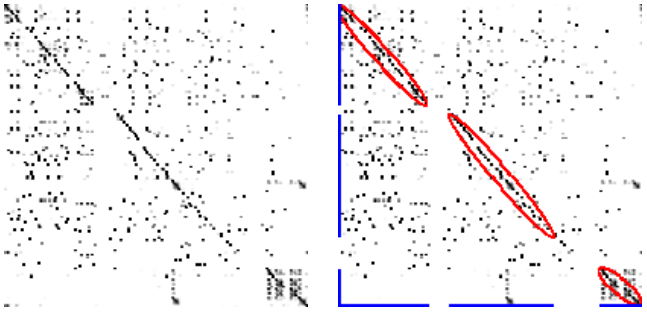


IP/06/1515 de-fr

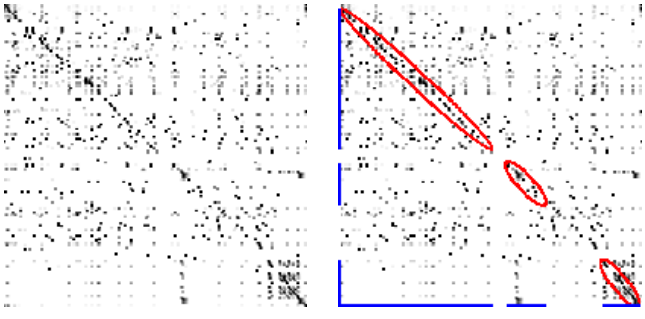


IP/06/527 de-fr

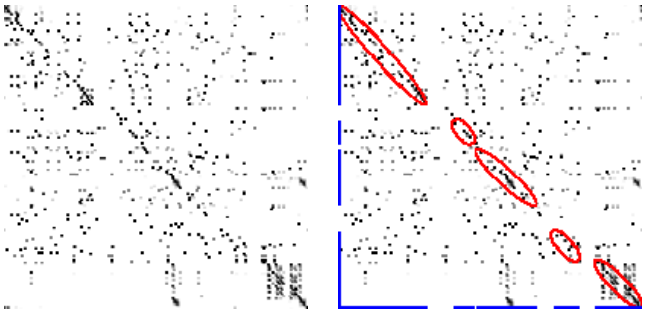
IP/06/527 da-de



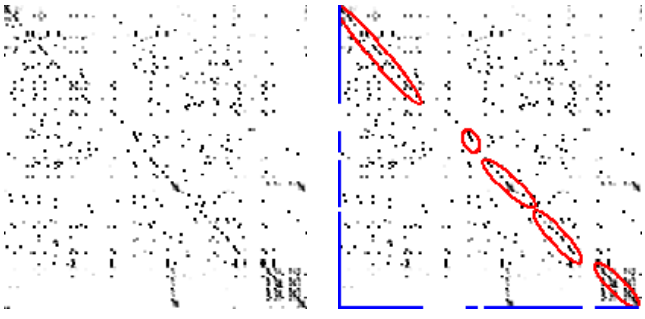
IP/07/139 es-fr



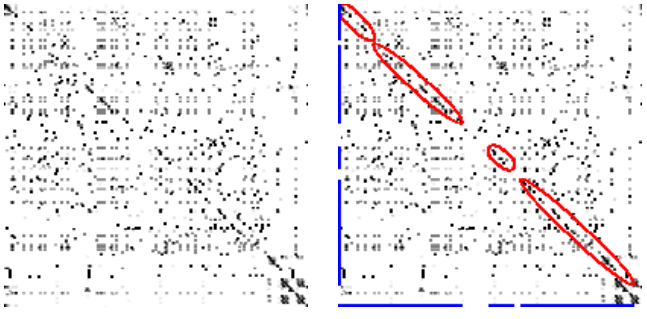
IP/07/139 da-de



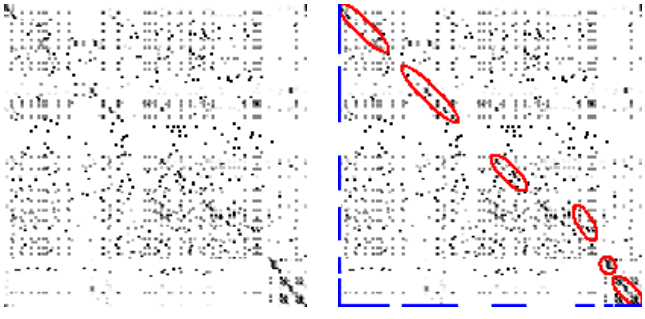
IP/07/139 en-fr



IP/07/139 el-fr



IP/05/1514 es-fr



IP/05/1514 da-de

TABLEAU 36 – 10 bi-documents asynchrones avec suppression parmi les 12 correctement attribués sur les collections thématiques avec la méthode *Grand Angle* (voir tableau 25).

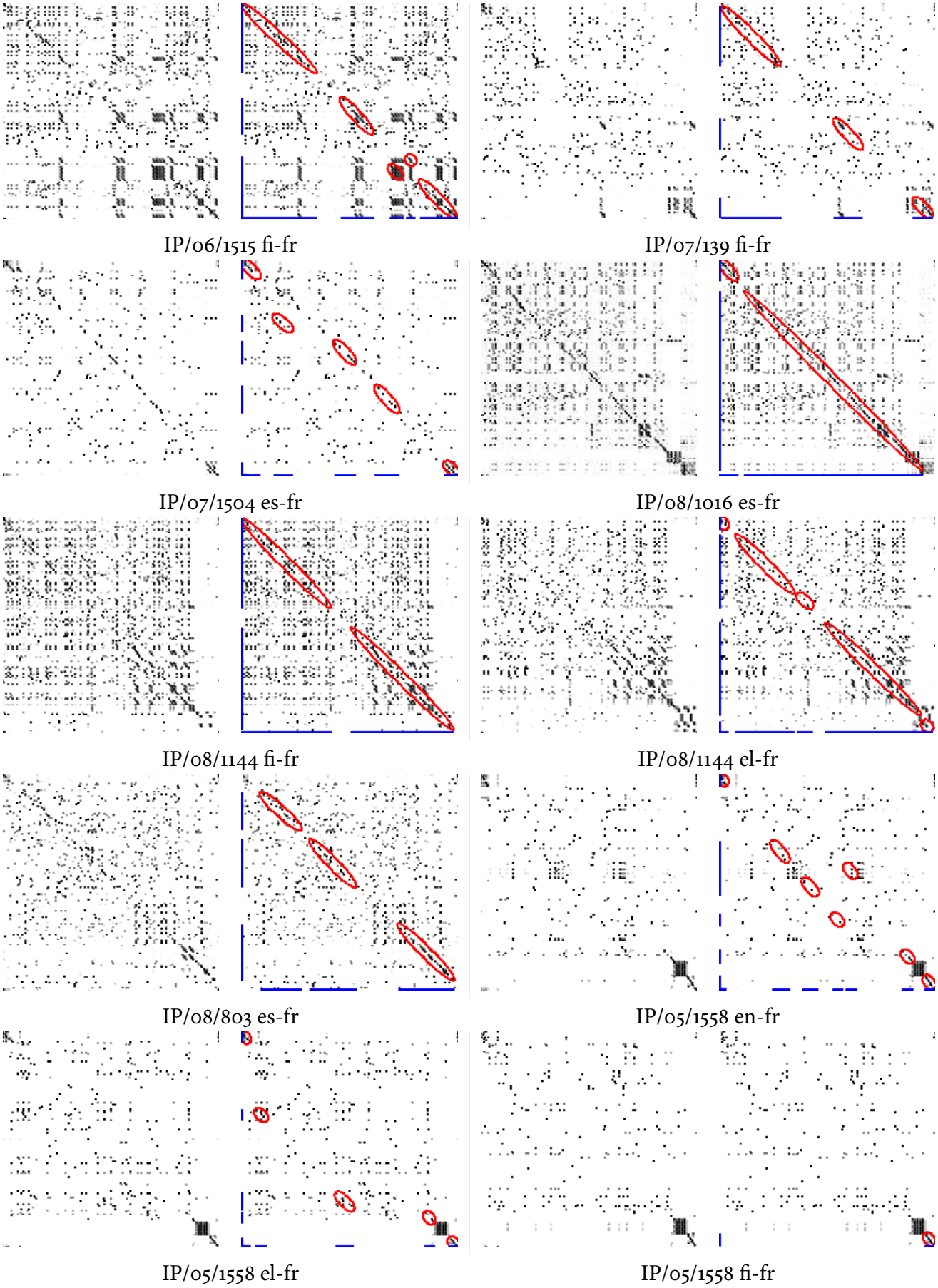


TABLEAU 37 – 10 bi-documents asynchrones avec suppression parmi les 26 attendus sur les collections thématiques avec la méthode *Grand Angle* (voir tableau 25).

L'observation de ces tableaux nous amène à plusieurs commentaires. Tout d'abord il convient de dire que la majorité de ces images offre à l'œil nu une idée claire des phénomènes engagés entre les deux volets concernés.

Nous plaçons donc principalement les difficultés dans les étapes ultérieures à la création des matrices :

- tout d'abord dans le traitement de ces images, certains segments de droites que nous souhaiterions voir isolés ne le sont pas ;
- enfin et c'est là la majorité des cas, dans le diagnostic que nous avons mis en œuvre. Des expériences de détection automatique des caractéristiques propres à chaque type sont en cours.

7.3.2 *Pourquoi des matrices restent indéfinies ? ou mal définies ?*

Entre 2 et 35% des matrices restent indéfinies selon la dimension observée. Ces matrices sont trop claires ou trop foncées pour permettre un diagnostic. Nous voyons plusieurs raisons à cela :

- les volets sont petits ou plus grands que la moyenne des communiqués, la taille que nous donnons à nos matrices n'est pas adaptée. Les segments de volets ne sont dans ce cas pas suffisamment significatifs ;
- les volets présentent une différence de taille significative lorsque par exemple, un des deux volets est quasi non traduit, c'est le cas notamment du volet grec du communiqué IP-06-751 et du volet espagnol IP-05-1653 ;
- les langues sont éloignées. Entre certains couples de langues, il existe moins de correspondances bi-univoques ou quasi bi-univoques.
- à l'inverse des volets présentant beaucoup de similarité notamment du fait de passages dans la même langue dans les deux volets (cf. cas de multilinguisme), la matrice est foncée, la détection des segments est délicate à réaliser.

7.4 ALIGNEMENT DE ZONES

Nous présentons dans cette dernière section les résultats en contexte de l'alignement de zones sur 5 documents asynchrones (3 suppressions, 2 inversions) correctement diagnostiqués.

Le tableau 38 illustre un cas de suppression dans un des deux volets, le volet fr, correspondant à environ un tiers du volet (2120 caractères). Si la suppression a bien été diagnostiquée, l'alignement de zones n'est lui que partiellement correct. Seule la multizone 2 correspond à l'attendu. Ce document fait partie des documents que nous présentions au chapitre 1 (p.25) et que nous annonçions au chapitre 3 vouloir être capable de traiter.

Le tableau 39 illustre un cas de suppression dans un des deux volets, le volet es, correspondant à 4 lignes (350 caractères). Les termes d'une aide apportée par l'Europe à la Bulgarie et la Roumanie n'ont pas fait l'objet d'une traduction en espagnol. L'alignement de zones est globalement correct.
Le tableau 40 illustre un cas de suppression dans un des deux volets, le volet fr, correspondant à environ 1000 caractères. Des balises type `</a><a name="_Hlt188932286">` n'apportant rien en terme de mise en forme ont été supprimées en français. L'alignement de zones est globalement correct.
Le tableau 41 illustre un cas de différences d'ordre entre les zones de textes de deux volets. L'ordre des paragraphes est différent d'un volet à l'autre. Tous les segments de droites de la matrice n'ont pas été mis en évidence, cependant l'alignement de zones découlant des segments isolés est globalement correct.
Le tableau 42 illustre un cas de différences d'ordre entre les zones de textes de deux volets. L'ordre des présentations des projets listés par pays respecte l'ordre alphabétique des noms des pays concernés. Tous les segments de droites de la matrice ont été mis en évidence, l'alignement de zones découlant des segments est globalement correct. Ce document fait partie des documents que nous présentions au chapitre 1 (p.22) et que nous annonçions au chapitre 3 vouloir être capable de traiter.

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

TABLEAU 38 – Alignement de zones entre les volets fr et en du communiqué IP/05/473 avec suppression détectée au travers de la collection 2 et de la méthode « Grand Angle ».

	IP/05/1344	
	fr	*es*
Multizone 1	gn="right"> Bruxelles, le 25 octobre 2005 La Bulgarie et la Roumanie se rapprochent de l'adhésion La Commission a adopté ce jour le rapport global de suivi relatif aux préparatifs de la Bulgarie et de la Roumanie en vue de leur adhésion à l'UE. Ce rapport montre que les deux pays ont bien progressé en la matière. Ils devraient être à même de satisfaire aux conditions d'adhésion à l'Union à la date prévue du 1^{er} janvier 2007, pour autant qu'ils consacrent tous leurs efforts à la mise en œuvre des réformes. La Commission continuera de suivre de près leurs préparatifs. Elle réexaminera la situation en avril-mai 2006, où elle pourrait recommander, au besoin, le report de l'adhésion à 2008 en cas d'impréparation manifeste de l'un des pays, voire des de	uselas, 25 de octubre de 2005 Bulgaria y Rumanía se aproximan a la adhesión La Comisión ha adoptado hoy el Informe Global de Seguimiento de 2005 sobre los preparativos de Bulgaria y Rumanía para su adhesión a la UE. El informe muestra que ambos países han avanzado bien en sus preparativos. Deberían poder cumplir los requisitos de adhesión a la UE en la fecha prevista de 1 de enero de 2007, siempre que dediquen todos sus esfuerzos a las reformas. La Comisión va a continuar siguiendo de cerca los preparativos de estos países. Volverá a examinar la situación en abril – mayo de 2006, momento en el que podría recomendar, en caso necesario, posponer la adhesión hasta 2008 en el caso de que, manifiestamente, u
Multizone 2	s rapports devraient être prises très au sérieux et servir d'aiguillon à l'accélération des réformes, si la Bulgarie et la Roumanie désirent être au rendez-vous de l'adhésion au 1^{er} janvier 2007». La Commission confirme que la Bulgarie et la Roumanie remplissent les critères politiques d'adhésion. Néanmoins, des efforts supplémentaires doivent être consentis, en vue notamment de renforcer l'État de droit, en améliorant la fonction publique et le système judiciaire et en luttant efficacement contre la corruption. La Bulgarie et la Roumanie satisfont à l'obligation d'être une économie de marché viable. Si la Bulgarie maintient le rythme actuel de son processus de réforme et si la Roumanie poursuit avec autant de vigueur la mise en œuvre de son programme de réformes structurelles, les deux pays devraient être en mesure de faire face à la pression concurrentielle et aux forces du marché à l'intérieur de l'Union. Ils ont continué à progresser dans l'adoption et la mise en œuvre de la législation de l'UE et sont bien avancés dans la plupart des domaines. Toutefois, la	stiones citadas en nuestros informes se deben tomar muy en serio y han de ser un incentivo para acelerar las reformas, si Bulgaria y Rumanía quieren estar preparadas para la adhesión el 1 de enero de 2007.» La Comisión confirma que Bulgaria y Rumanía cumplen los criterios políticos de adhesión. Sin embargo, deben hacer un esfuerzo adicional, en particular reforzar el Estado de derecho, mejorando la administración pública y el sistema judicial y luchando de manera efectiva contra la corrupción. Bulgaria y Rumanía cumplen el requisito de ser una economía de mercado viable. Si Bulgaria mantiene el ritmo actual de reformas y Rumanía prosigue con la misma determinación la aplicación de su programa de reformas estructurales, ambos países deberían poder hacer frente a la presión de la competencia y a las fuerzas del mercado dentro de la UE. Los dos países han seguido progresando en la adopción y la aplicación de la legislación de la UE. Han avanzado mucho en la mayoría de los diversos
Multizone 3	ons dans ces domaines particuliers d'ici au 1^{er} janvier 2007. L'an prochain, dans le courant des mois d'avril et mai, la Commission fera le point sur la situation. Elle pourrait alors recommander, si nécessaire, de différer l'adhésion de la Bulgarie ou de la Roumanie au 1^{er} janvier 2008 s'il existe un risque grave d'i	específicas de ahora al 1 de enero de 2007. El año próximo, en los meses de abril – mayo, la Comisión volverá a analizar la situación. Podría recomendar entonces, en caso necesario, retrasar la adhesión de Bulgaria o Rumanía hasta el 1 de enero de 2008 si existe un riesgo grave de que cualqui
Multizone 4	> pour la Roumanie : 1 155 millions d'euros. Pour de plus amples informations, consulter : MEMO/05/395 et MEMO/05/396 <a href="http://europa.eu.int/comm/enlargement/index.htm">http://europa.eu.int/co	tos importantes para enero de 2007. Para información adicional, véase : MEMO/05/395 et MEMO/05/396 <a href="http://europa.eu.int/comm/enlargement/index.htm">http://europa.eu.int/co

TABLEAU 39 – Alignement de zones entre les volets fr et es du communiqué IP/05/1344 avec suppression détectée au travers de la collection 1 et de la méthode « Grand Angle ».

	IP/08/405	
	fr	*fi*
Multizone 1	 <i> Vingt-sept «jeunes traducteurs», un par État membre de l'Union européenne, sont venus aujourd'hui à Bruxelles pour recevoir leur prix à l'issue du tout premier concours européen de traduction organisé à l'intention des écoles. M. Leonard Orban, commissaire europ	0. maaliskuuta 2008 <h1> EU-palkinnot lupaaville nuorille kääntäjille </h1> <h2> <i> Brysseliin saapuu tänään 27 nuorta kääntäjää – yksi jokaisesta EU-jäsenvaltiosta – noutamaan ensimmäisessä Euroopan laajuisessa koululaisten käännöskilpailussa heille myönn
Multizone 2	ment fier que le travail de nos traducteurs, souvent invisible mais indispensable à l'Union, soit aujourd'hui sous les projecteurs.» Outre la cérémonie de remise des prix qui aura lieu au siège de la </b	että tänään on näkyvästi esillä kääntäjien työ, joka jää usein huomaamatta mutta joka on EU :lle ratkaisevan tärkeää. Nuorille kääntäjille on järjestetty komission päätoimipaikassa pidettävän palkintojuhlan lisäksi vierailu Euroopan komission
Multizone 3	s n'importe quelle autre langue officielle de l'Union. Plus de 1 300 textes ont été reçus, représentant 134 combinaisons différentes de langue source et de langue cible. Les traducteurs de la DG Traduction ont noté les copies, contribuant ainsi à déterminer quelle était la meilleure traduction dans chaque État membre. Premier du genre, ce concours de traduction a été organisé par la Commission européenne à titre de projet pilote , le but étant de faire mieux connaître la place essentielle de la traduction dans la politique multilingue appliquée par la Commission. Il a également permis aux élèves de s'essayer au métier de traducteur	elle kielelle. Lähtötekstit käsittelivät vastuullista ja vaihtoehtoista matkailua. Kilpailuun lähetettiin yli 1 300 käännöstä. Lähtö- ja kohdekielten erilaisia yhdistelmiä oli 134. Käännöstoimen pääosaston kääntäjät arvioivat käännökset ja osallistuivat siten kunkin EU-jäsenvaltion voittajakäännöksen valitsemiseen. Tämä laatuaan ensimmäinen käännöskilpailu oli Euroopan komission pilottihanke , jolla haluttiin tuoda esiin kääntämisen keskeistä roolia komission noudattamassa monikielisyyspolitiikassa. Lisäksi koululaiset saivat hankkeessa tilaisuuden kok

TABLEAU 40 – Alignement de zones entre les volets fr et fi du communiqué IP/08/405 avec suppression détectée au travers de la collection 3 et de la méthode « Grand Angle ».

	IP/07/1008	
	da	*de*
Multizone 1	ght" > IP/07/1008 Bruxelles, den 4. juli 2007 <h1> Reformen af den fælles landbrugspolitik : Med vinreformen vil Europa kunne generobre tabte markedsandele [...]	<document celex="IP-07-1008" lang="de"> IP/07/1008 " Brüssel, den 4. Juli 2007 <h1> GAP-Reform : Weinreform wird Europa helfen, verlorene Marktanteile zurückzugewinnen [...]
Multizone 2	nsigten, at krisedestillation skal erstattes af to kriseforvaltningsforanstaltninger, som finansieres over de nationale rammebeløb. [...]	schafft. Die Dringlichkeitsdestillation würde durch zwei aus den nationalen Finanzrahmen finanzierte Maßnahmen für das Krisenmanagement ersetzt [...]
Multizone 3	altninger er bl.a. : salgsfremstød i tredjelande, omstrukturering og omstilling af vinbedrifter, støtte til grøn høst, nye kriseforvaltningsforanstaltninger, nemlig forsikring mod naturkatastrofer og dækning af de administrative omkostninger i forbindelse med oprettelse af sektorspecifikke gensidige fonde. Foranstaltninger til udvikling af landdistrikter : Mange af foranstaltningerne i forordningen om udvikling af landdistrikterne kan have interesse for vinsektoren, ikke mindst etablering af unge landbrugere, bedre markedsføring, erhvervsuddannelse, støtte til producentorganisationer, støtte til dækning af ekstraomkostninger og indkomsttab ved opretholdelse af kulturlandskaber samt førtidspensionering. For at tage højde herfor er det meningen, at der gradvist skal overføres penge til budgettet for udvikling af landdistrikterne. I 2009 bliver der således tale om 100 mio. EUR og fra 2014 om 400 mio. EUR. Disse penge skal øremærkes til vinproducerende områder. <	können. Die Entscheidung der Erzeuger, ihre Produktion zu steigern, wird davon abhängen, wieweit sie, das, was sie erzeugen, auch verkaufen können. Önologische Verfahren : Die Zuständigkeit für die Genehmigung neuer bzw. Änderung bestehender önologischer Verfahren wird auf die Kommission übertragen, die die von der OIV genehmigten önologischen Verfahren bewertet und in die Liste von genehmigten EU-Verfahren aufnimmt. Die EU genehmigt die Anwendung international bereits zugelassener önologischer Verfahren für die Herstellung von Wein, der zur Ausfuhr in diese Bestimmungsländer vorgesehen ist. Die Einfuhr von Most zur Weinbereitung und der Verschnitt von Weinen aus der EU mit eingeführten Weinen bleiben weiterhin verboten. Bessere Etikettierungsvorschriften : Das Konzept für Qualitätsweine aus der EU wird auf dem geografischen Ursprung basieren (in einer bestimmten Region erzeugter Qualitätswein). Weine mit geografischer Angabe werden unterteilt in Weine mit geschützter geografischer Angabe und Weine mit geschützter Ursprungsbezeichnung. Die Etikettierung wird den Bedürfnissen der Verbraucher entsprechen, indem sie vereinfacht wird und vor allem erstmals bei EU-Weinen ohne geografische Angabe die Angabe der Rebsorte und des Jahrgangs auf dem Etikett ermöglicht, um der Verbrauchernachfrage nach Rebsortenweinen Rechnung zu tragen. Nationale Finanzrahmen : Diese Finanzrahmen werden den Mitgliedstaaten die Möglichkeit geben, die Maßnahmen an ihre jeweilige Situation anzupassen. Die Mittelausstattung beträgt zwischen 634 Mio. EUR im Jahr 2009 und 850 Mio. EUR ab 2015. Der für jedes Land verf
Multizone 4	Ønologiske fremgangsmåder : Ansvaret for godkendelse af nye ønologiske fremgangsmåder eller ændring af de eksisterende fremgangsmåder overdrages til Kommissionen, der vil foretage en vurdering af de ønologiske fremgangsmåder, der er accepteret af OIV, og medtage dem på listen over accepterede fremgangsmåder i EU. EU vil tillade internationalt anerkendte fremgangsmåder med henblik på fremstilling af vin til eksport til de pågældende destinationer. Forbuddet mod fremstilling af vin af importeret most og blanding af vine fra EU med importerede vine opretholdes. Bedre etiketteringsregler : Begrebet EU-kvalitetsvine baseres på geografisk oprindelse (kvalitetsvin produceret i et bestemt dyrkningsområde). Vine med geografiske betegnelser opdeles i vine med beskyttede geografiske betegnelser og vine med beskyttede oprindelsesbetegnelser. Etiketteringen vil tage hensyn til forbrugernes behov. Den bliver således enklere, og navnlig tillades det for første gang at anføre druesort og årgang på etiketten for EU-vine uden geografisk betegnelse for at imødekomme forbrugernes efterspørgsel efter vine fremstillet af en enkelt druesort. Salgsfremstød og oplysning : Kommissionen vil gennemføre en resolut og ansvarlig kampagne for salgsfremme og oplysning. Hertil skal der afsættes et budget på 120 mio. EUR fra de nationale rammebeløb til salgsfremmende foranstaltninger uden for EU, hvor EU bidrager med 50% af finansieringen. Der vil blive gennemført nye oplysningskampagner in	fügbare Betrag wird anhand der Weinanbaufläche, der Erzeugung und der historischen Ausgaben berechnet. Mögliche Maßnahmen sind u.a. : Absatzförderung in Drittländern, Umstrukturierung/Umstellung von Rebflächen, Unterstützung für die grüne Weinlese, neue Maßnahmen zum Krisenmanagement wie z.B. Versicherung gegen Naturkatastrophen und Deckung der Verwaltungskosten für die Errichtung eines sektorspezifischen Fonds auf Gegenseitigkeit. Maßnahmen zur Entwicklung des ländlichen Raums : Viele Maßnahmen im Rahmen der Verordnung über die Entwicklung des ländlichen Raums könnten für den Weinsektor von Interesse sein, u.a. Niederlassung von Jungweinbauern, Verbesserung der Vermarktung, Berufsbildung, Förderung von Erzeugerorganisationen, Unterstützung zur Deckung der mit der Erhaltung von Kulturlandschaften verbundenen zusätzlichen Kosten und Einkommenseinbußen und Vorruhestand. Zu diesem Zweck würden Mittel auf die Maßnahmen zur ländlichen Entwicklung übertragen (von 100 Mio. EUR im Jahr 2009 bis 400 Mio. EUR im Jahr 2014). Diese Mittel wären den Weinbauregionen vorbe
Multizone 5	nter vil fremstille vin udelukkende af druer og ikke-subsideret most. EU's vinsektor EU har over 2,4 mio. bedrifter, der producerer vin, svarende til 3,6 mio. ha og 2% af EU's landbrugsareal. Vinproduktionen i 2006 [...]	n im Rahmen der Entwicklungsprogramme für den ländlichen Raum werden aufgestockt. Der Weinsektor der EU In der EU gibt es mehr als 2,4 Millionen weinerzeugende Betriebe mit einer Fläche von insgesamt 3,6 Mio. ha, das sind 2% der landwirtschaftlichen Fläche der EU. Im Jahr 2006 [...]

TABLEAU 41 – Alignement de zones entre les volets da et de du communiqué IP/07/1008 présentant une différence d'ordre des zones détectée au travers de la collection 1 et de la méthode « Petit Angle ».

	IP/05/1157	
	fr	*en*
Multizone 1	Bruxelles, le 19 septembre 2005 Environnement : la Commission subventionne 89 projets d'innovation dans 17 pays pour un montant de 71 millions d'euros <i>La Commission européenne a approuvé le financement de 89 projets innovants dans le domaine de l'environnement dans 17 pays, au titre du programme LIFE-Environnement 2005. [...] Pour plus de détails concernant chaque projet, consulter le site suivant : http://europa.eu.int/comm/environment/ life/project/index.htm ANNEXE Résumé des projets	/a> Environment : Commission supports 89 innovation projects in 17 countries with €71 million </h1> <i> The European Commission has approved funding for 89 environmental innovation projects in 17 countries under the LIFE-Environment programme 2005. [...] More information See the annex for a summary of the 88 projects funded under LIFE-Environment. More detailed information on each project is available at : <a href="http://europa.e
Multizone 2	r appliquera une stratégie intégrée pour réduire la pollution agricole diffuse, dans le sens de la directive cadre sur l'eau 1. Le second [...] Le second projet concerne le prétraitement de la laine dans la production de fil. L'objectif principal est de supprimer les émissions de composés organohalogénés absorbables (AOX) et de réduire sensiblement l'utilisation de produits chimiques dans le processus de nettoyage, grâce un procédé durable de prétraitement par plasma. Un projet porte sur la gestion des déchets e	ht"> ANNEX Overview of LIFE-Environment projects 2005 by country Belgium – 2 projects [...] Denmark – 6 projects [...] Estonia – 1 project [...] the fermentation of manure, processing of bio-gas into
Multizone 3	er les tôles laminées à froid. Un nouveau procédé basé sur la technologie sous vide à haute pression et n'utilisant pas de produits chimiques sera employé. Belgique – deux projets [...] Danemark – six projets [...] Espagne – seize projets Trois projets portent sur la gestion des eaux . Le premier permettra de définir un modèle d	tronic equipment, in line with EU legislation ^{[2] }, with a particular emphasis on rural areas. The second targets households, schools and day-care centres in Helsinki, with a view to increasing awareness and ensuring the amount of waste produced does not exceed 2003 levels. France – 11 projects [...] The sixth will substitute lead with o
Multizone 4	s variétés d'amandiers capables de résister à de telles conditions. Le troisième projet vise à définir un système de gestion durable de la viticulture de montagne, en vue de réduire les incidences de cette activité sur le paysage, les sols et les ressources en eau. Quatre projets traitent des technologies propres. [...] Le sixième projet démontrera qu'il est techniquement et économiquement possible d'appliquer un nouveau procédé à haute capacité pour séparer les alliages métalliques à pureté élevée (plus de 90%). Utilisé pour extraire le fer, l'aluminium et les métaux lourds contenus dans les véhicules hors d	to reduce diffuse pollution from agriculture, in support of the Water Framework Directive1. The second [...] The second concerns the pre-treatment of wool in yarn production. The main goal is the elimination of emissions of absorbable organic halides (AOX) and a significant decrease in the use of chemicals in the cleaning process, through a sustainable plasma pre-treatment process. One project addresses waste management</
Multizone 5	nouvelle technologie recourant à la fermentation du lisier, à la transformation du biogaz en énergie et en chaleur « écologiques » et à la séparation intégrale des composants recyclables et non recyclables. Finlande – deux projets [...] France – onze projets [...] Le quatrième projet vise à démontrer qu'il est techniquement possible de recourir à la technologie des ultrasons pour réduire la production de boues résiduaires dans les stations d'épuration des eaux usé	ng of cold rolled plates. A new chemical-free process will be used, based on high-pressure vacuum technology. Greece – 4 projects [...] Hungary – 1 project The project, covering water management, assesses the scale of arsenic contamination in groundwater in the southern part of Hungary. It will develop a pilot management plan, incorporating a new arsenic removal technology. Ireland – 2 projects [...] Italy – 15 projects [...] Netherlands – 7 projects [...] Portugal – 2 projects [...] Romania – 1 project [...] Spain – 16 projects [...] The third aims at defining
Multizone 6	ernier projet français concerne la gestion de la qualité de l'air. Il vise à mettre au point un échantillonneur d'air basé sur une nouvelle méthode de surveillance des pollens dans l'air. Au lieu de quantifier les grains de pollens selon leur morphologie, cette méthode reposera sur la mesure en ligne de l'antigénéité/l'allergénéité. Grèce – quatre projets [...] Hongrie – un projet [...] Irlande – deux projets [...] Italie – quinze projets [...] Luxembourg – un projet [...] Pays-Bas – sept projets [...] Portugal – deux projets [...] Roumanie – un projet [...] Royaume-Uni – dix projets [...] Le quatrième projet vise à réduire l'élimination des déchets hospitaliers non stériles dans les	g a mountain viticulture sustainable management system in order to reduce the environmental impacts of this activity on landscape, soil and water resources. Four projects deal with clean technologies. [...] The last project will demonstrate the technical and economic feasibility of a new high-capacity process to separate high purity metal alloys (>90%). Used for the separation of iron, aluminium and heavy metals from
Multizone 7	s incidences environnementales des activités économiques. Le premier vise à démontrer l'efficacité du recyclage de l'eau au moyen d'un nouveau réacteur de digestion aérobie des eaux usées. Le second projet concerne l'exploitation des friches industrielles pour la culture de biomasse à des fins énergétiques, la réhabilitation des terres endommagées et la production de chaleur et d'énergie à partir de sources d'énergie renouvelables. [...] Suède – deux projets [...] Directive 2002/95/CE du Parlement européen et du Conseil du 27 janvier 2003 relative à la limitation de l'utilisation de certaines substances dangereuses dans	re-use. A fourth project aims to reduce the disposal of non-sterile clinical waste in landfill sites and promote its use as a raw material for recycled products. Two projects seek to mitigate the environmental impact of economic activities. One will demonstrate the effectiveness of water recycling using a new reactor for aerobic digestion of wastewater. A second aims to re-use brownfield sites to grow biomass energy crops, restore damaged land, and generate heat and power from renewable energy sources. [...] Council Directive 1999/13/EC of 11 March 1999 on the limitation of em

TABLEAU 42 – Alignement de zones entre les volets fr et en du communiqué IP/05/1157 présentant une différence d'ordre des zones détectée au travers de la collection 1 et de la méthode « Petit Angle »

CONCLUSION ET PERSPECTIVES

Nous annoncions en introduction qu'une marge de progression dans le domaine de l'alignement de documents traduits semblait envisageable. Au regard du chemin parcouru, nous pouvons valider cette hypothèse de départ. Une voie est ouverte vers le traitement de documents traduits *réels*. Un tel résultat est le fruit d'une conjonction de connaissances linguistiques et de compétences informatiques en algorithmique du texte et en traitement d'images.

Le chapitre 1 nous a permis d'illustrer la complexité de la traduction en tant que produit de l'opération traduisante, opération empreinte à la fois de servitudes linguistiques et d'un travail de réécriture de la part des traducteurs. Deux phénomènes amenant chacun son lot de différences entre des documents traduits : différence de volume, ajout ou suppression, inversion…

Dans le chapitre 2, nous avons procédé à un tour d'horizon des méthodes existantes avec un intérêt particulier pour les façons de prendre en charge ces différences entre les documents traduits. Le constat qui en est ressorti est que l'hypothèse de parallélisme largement exploitée par l'état de l'art constitue un verrou au traitement de documents traduits réels.

Notre parti pris a dès lors été celui d'une méthode sans présupposé de parallélisme. Ainsi, dans le chapitre 3, nous avons formulé les grandes lignes de notre approche et présenté le corpus que nous souhaitions être capable de traiter, un corpus réel. Plus précisément dans les chapitres 4 et 5, nous avons successivement présenté les concepts originaux à la base de notre méthode : le multidocument, les collections de multidocuments, le document et sa mise en forme, les chaînes de caractères répétés et les multizones, avant de détailler la méthode à proprement parler.

Enfin, la troisième partie a permis de montrer qu'un travail interdisciplinaire alliant hypothèses linguistiques, algorithmique du texte et traitement d'image donnait des résultats d'ores et déjà prometteurs.

Les images que nous tirons des bi-documents offrent à l'**œil nu** une vision claire des stratégies de traductions. Ces images nous ont d'ailleurs permis de pointer d'autres réalités sur les traductions que celles communément envisagées : les permutations de zones importantes entre deux versions d'un même document, les suppressions de zones de textes et les cas de zones restées dans la langue source (volets multilingues).

Nous faisons le constat que l'identification **automatique** des documents asynchrones ne donne pas encore pleinement satisfaction. Certaines pistes susceptibles de mener à des améliorations de notre méthode sont déjà envisagées :

- affiner le diagnostic des matrices, permettant notamment de mieux diagnostiquer les bi-documents asynchrones et de capter les modèles émergents, comme le modèle multilingue. Sur ce point, une collaboration avec des chercheurs en fouille de données pourrait nous permettre de proposer de meilleurs combinaisons de critères de diagnostics ;
- ôter les seuils que nous avons fixés afin de procéder à un filtrage plus fin des appariements par des combinaisons de filtres moins indépendants des langues et des collections ;
- de la même manière, adapter automatiquement la taille des matrices en fonction des volets à traiter ;
- détecter plus finement les frontières de zones. Une des stratégies envisageables serait de partir d'unités prédéfinies comme l'alinéa ou la section et de les aligner selon notre méthode.

À terme, l'apport de la détection des multizones pourra être évalué en vérifiant que par cette méthode nous sommes désormais effectivement capable de traiter ce que l'on n'était pas capable de traiter : les multidocuments avec inversion ou suppression. Cette évaluation pourra dans un premier temps être réalisée sur la tâche d'alignement de phrases avant de l'être sur l'alignement d'unités sous-phrastiques. En outre, d'autres corpus comme l'Acquis Communautaire qui présente également des cas de suppressions ou de non traduction d'annexes, pourront également être testés.

Néanmoins ces travaux peuvent d'ores et déjà se placer dans le cadre d'une chaîne d'observation et de contrôle qualité de documents traduits. La détection de corpus de traductions synchrones ou non est également un champ d'utilisation de notre méthode qui atteint les 97% de décisions sur des corpus de langues proches.

Quatrième partie

ANNEXES

A

ÉVALUATION QUANTITATIVE DES APPARIEMENTS

Lorsque l'on examine les appariements obtenus, ils semblent cohérents et représentatifs de ce que l'on cherche. Pour aller plus loin, il est néanmoins souhaitable de valider à plus grande échelle. Une difficulté supplémentaire dans notre contexte est que les équivalents multilingues peuvent être aussi bien des expressions, des morceaux de mots que des balises HTML. Notre méthode s'appuie sur une des applications de l'algorithme : la constitution de dictionnaires. Si cet objectif est atteignable, nous devrions, en utilisant des dictionnaires existants disponibles sur Internet, trouver au sein de nos appariements des liaisons existantes dans les dictionnaires.

On observe aussi que de nombreuses chaînes mises en évidence sont des chaînes identiques d'une langue sur l'autre, ou cognats. À notre avis, cela constitue également un indice de bon fonctionnement de la méthode d'appariement dans la mesure où, rappelons-le, elle ne s'appuie que sur des informations de fréquence et de positions et ne fait aucun usage du contenu ou de la longueur des n-grammes.

Dans les deux cas (traductions ou cognats), nous envisageons avant l'évaluation une étape de reconstruction des mots : nous retournons aux textes pour trouver une liste de mots dans lequels interviennent les deux N-grammes appariés. Cette étape présente peu de difficultés théoriques et computationnelles dans la mesure où nous connaissons exactement les différentes occurrences de chaque population.

La figure 22 présente les évolutions des pourcentages de cognats et de traductions trouvés par notre méthode appliquée à un corpus bilingue anglais/français de 40 bidocuments (soit 80 textes). En abscisse est porté le nombre de mots vus par langue et en ordonnée le pourcentage de ces mots identiques (cognats) ou trouvés dans les dictionnaires de traduction. Il faut signaler qu'on ne peut rien dire sur les autres couples de mots, sinon qu'ils ne sont pas identiques et ne figurent pas dans le dictionnaire de traduction. En particulier, le repérage d'expressions multi-mots équivalentes, qui est un de nos objectifs, ne peut que partiellement être évalué par cette approche. Malgré l'imperfection de la méthode d'évaluation, les résultats sont positifs partant de 50% de cognats et 6% de traductions et se stabilisant autour de 6% pour les deux.

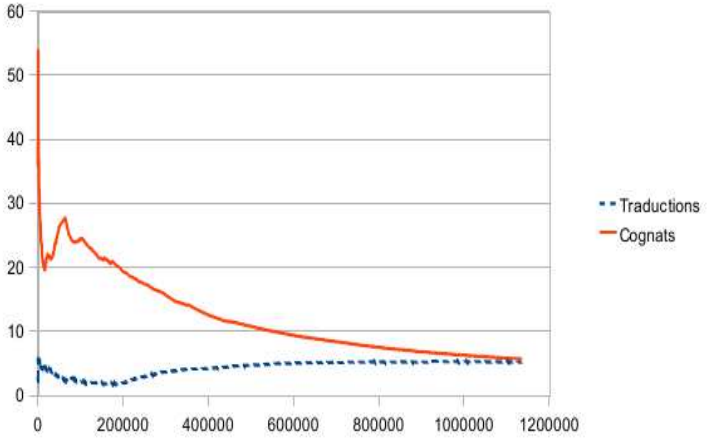


FIGURE 22 – Évolution des pourcentages de cognats (trait continu) et de traductions (pointillés) sur 40 md en français-anglais

B

ÉVALUATION MANUELLE DU PARALLÉLISME ENTRE LES VOLETS DES COLLECTIONS

Nous présentons ci-après les diagnostics de parallélisme entre les volets des collections de notre corpus établis à l'oeil nu par nos soins. Chaque tableau présente l'ensemble des bi-documents d'une collection. Chaque bi-document a fait l'objet d'un diagnostic : synchrone, asynchrone avec inversion ou asynchrone avec suppression et le cas échéant nous avons relevé des spécificités telles que :

- le type d'inversion ;
- le type de suppression ;
- la longueur des volets ;
- la présence de multilinguisme.

Les principes qui ont présidés l'attribution d'un diagnostic sont les suivants :

- seules les inversions et suppressions sur-phrastiques ont été prises en considération ;
- les suppressions de balises n'engageant pas de changement de rendu sont considérées comme des suppressions sur-phrastiques de même que les suppressions d'url.

S'il n'est pas toujours évident de faire la part des choses entre liberté du traducteur et contraintes éditoriales, nous pouvons néanmoins faire quelques remarques d'ordre qualitatif sur les documents présentant des différences d'ordre dans le discours ou des différences de contenu, des suppressions. Il ressort de cette étude que les inversions sont principalement dûes à des tris par ordre alphabétique dans le texte ou à l'intérieur de tableau (changeant ainsi l'ordre des lignes de ces derniers). Le cas de paragraphes inversés a également été rencontré sans pouvoir y relever de raison apparente. Pour ce qui est des suppressions, nous avons pu relever des cas de suppressions divers allant de la suppression de titres, de balises, de paragraphes, d'annexes, de tableaux à la suppression de l'intégralité du corps de certains volets.

COLLECTION	SYNCHRONES	ASYNCHRONES AVEC INVERSION	ASYNCHRONES AVEC SUPPRESSION
1	228 (95,00%)	5 (2,08%)	7 (2,92%)
2	223 (92,92%)	0 (0,00%)	17 (7,08%)
3	201 (83,75%)	14 (5,83%)	25 (10,42%)
Transport	229 (95,42%)	5 (2,08%)	8 (3,33%)
Téléphone	220 (91,67%)	5 (2,08%)	15 (6,25%)
Santé	231 (91,67%)	6 (2,50%)	3 (1,25%)

TABLEAU 43 – Étude quantitative des différents phénomènes répertoriés par collection (une collection = 240 bi-documents).

Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails
IP-05-1013	da-de	synchrone		IP-05-1224	da-de	synchrone		IP-05-1457	da-de	synchrone		IP-05-1011	da-de	synchrone	Petit document (1,4k - 1,7K)
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	Petit document (1,7K)
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	Petit document (2,6k - 1,7K)
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	Petit document (1,5k - 1,7K)
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	Petit document (1,7k - 1,7K)
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	Petit document (1,6k - 1,7K)
IP-05-1068	da-de	synchrone		IP-05-1097	da-de	synchrone		IP-05-1225	da-de	synchrone		IP-05-1473	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-1125	da-de	synchrone		IP-05-1155	da-de	synchrone		IP-05-1226	da-de	synchrone		IP-05-1490	da-de	synchrone	bilingues : tableau en anglais dans les 2 volets
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	idem
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	idem
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	fr bilingue : tableau en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	bilingues : tableau en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	idem
IP-05-1156	da-de	synchrone		IP-05-1233	da-de	synchrone		IP-05-1500	da-de	synchrone		IP-05-1157	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	inversion	Listes de projets triées par ordre alphabétique (en et fr)
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	inversion	idem
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	inversion	idem
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	inversion	idem
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	inversion	idem
IP-05-1169	da-de	synchrone		IP-05-1171	da-de	synchrone		IP-05-1239	da-de	synchrone		IP-05-1510	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-1175	da-de	synchrone		IP-05-1179	da-de	synchrone		IP-05-125	da-de	synchrone		IP-05-1514	da-de	suppression	de : suppression de balises < *aname*... >
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	fr : suppression de balises < *aname*... >
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-1189	da-de	synchrone		IP-05-1418	da-de	synchrone		IP-05-1391	da-de	synchrone		IP-05-1392	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-1208	da-de	synchrone		IP-05-1436	da-de	synchrone		IP-05-1272	da-de	synchrone		IP-05-1525	da-de	synchrone	bilingues annexe en anglais dans les 2 volets
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	bilingues annexe en anglais dans les 2 volets
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	bilingues annexe en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	fr : bilingue, annexe en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	bilingues annexe en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	bilingues annexe en anglais dans les 2 volets
IP-05-1217	da-de	synchrone		IP-05-1442	da-de	synchrone		IP-05-155	da-de	synchrone		IP-05-1344	da-de	suppression	de : suppression d'une url
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	de, el, es, fi, da :
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	suppression des dernières lignes
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	détaillant l'aide apportée
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	à la Bulgarie et la Roumanie
IP-05-1223	da-de	synchrone		IP-05-1451	da-de	synchrone		IP-05-130	da-de	synchrone		IP-05-1551	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	

TABLEAU 44 – Diagnostics manuels sur la Collection 1.

Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Détails
IP-05-1558	da-de	synchrone		IP-05-1573	da-de	synchrone		IP-05-1603	da-de	synchrone		IP-05-1653	da-de	synchrone	
	de-fr	suppression	fr : absence de		de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	suppression	deux tableaux		el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	suppression	présents dans		en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	suppression	les autres		es-fr	synchrone			es-fr	synchrone			es-fr	suppression	es : volet quasi non traduit
	fi-fr	suppression	volets		fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-1672	da-de	synchrone		IP-05-1673	da-de	synchrone		IP-05-1674	da-de	synchrone		IP-05-181	da-de	synchrone	tableau en anglais dans les 2 volets
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	tableau en anglais dans les 2 volets, annexe en français dans volet fr
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	tableau en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	tableau en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	tableau en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	tableau en anglais dans les 2 volets
IP-05-1679	da-de	synchrone		IP-05-225	da-de	synchrone		IP-05-231	da-de	synchrone		IP-05-182	da-de	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets trilingues
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	FAQ bilingue anglais-français les 2 volets = volets bi- et tri-lingues
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bi- et tri-lingues
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bilingues
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bi- et tri-lingues
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bi- et tri-lingues
IP-05-292	da-de	synchrone		IP-05-320	da-de	synchrone		IP-05-389	da-de	synchrone		IP-05-419	da-de	synchrone	da : tableau en anglais
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	el : tableau en anglais
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	es : tableau en anglais
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	fi : tableau en anglais
IP-05-32	da-de	synchrone		IP-05-384	da-de	synchrone		IP-05-445	da-de	synchrone		IP-06-202	da-de	synchrone	mini annexe en anglais dans les 2 volets, série de balises au milieu dans les 2
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	mini annexe en anglais dans les 2 volets, fr : suppression de la série de balises au milieu
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	mini annexe en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	idem
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	idem
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	idem
IP-05-473	da-de	synchrone	La partie	IP-05-459	da-de	synchrone		IP-05-460	da-de	synchrone		IP-06-634	da-de	synchrone	
	de-fr	synchrone	du communiqué		de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone	annonçant		el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	suppression	les perspectives		en-fr	synchrone			en-fr	synchrone			en-fr	suppression	en : suppression de quelques balises html à la fin
	es-fr	synchrone	n'existe		es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone	qu'en anglais		fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-489	da-de	synchrone		IP-05-513	da-de	synchrone		IP-05-544	da-de	synchrone		IP-06-751	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	el :volet quasi non traduit
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-55	da-de	synchrone		IP-05-572	da-de	synchrone		IP-05-599	da-de	synchrone		IP-07-1295	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	el : présence de deux tableaux inexistants dans les autres volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-606	da-de	synchrone		IP-05-628	da-de	synchrone		IP-05-663	da-de	synchrone		IP-07-919	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	fr : suppression d'une série de balises html au milieu
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-05-680	da-de	synchrone		IP-05-776	da-de	synchrone		IP-09-351	da-de	synchrone		IP-08-1829	da-de	suppression	da : suppression note de bas de page
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	el : suppression note de bas de page
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	suppression	en : suppression note de bas de page
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	es : suppression note de bas de page
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	fi : suppression note de bas de page

Tableau 45 – Diagnostics manuels sur la Collection 2.

Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails
IP-06-1006	da-de	synchrone		IP-06-1015	da-de	synchrone		IP-06-103	da-de	synchrone		IP-06-1008	da-de	synchrone	tableau en anglais dans les 2 volets
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	idem
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	idem
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	idem
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	idem
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	idem
IP-06-1177	da-de	synchrone		IP-06-1186	da-de	synchrone		IP-06-1219	da-de	synchrone		IP-07-165	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	fr : suppression de quelques balises html à la fin
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-1257	da-de	synchrone		IP-06-1275	da-de	synchrone		IP-06-1300	da-de	synchrone		IP-07-971	da-de	suppression	da : suppression d'une série de balises html
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	fr : idem
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	el : idem
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	suppression	en : idem
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	es : suppression d'une série de balises html et de plusieurs url
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-130	da-de	synchrone		IP-06-1313	da-de	synchrone		IP-06-1343	da-de	synchrone		IP-08-1110	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	fr : suppression d'une série de balises html au milieu
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-1356	da-de	synchrone		IP-06-1359	da-de	synchrone		IP-06-135	da-de	synchrone		IP-08-1324	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	fr : suppression de quelques balises html à la fin
IP-06-1384	da-de	synchrone		IP-06-1415	da-de	synchrone		IP-06-1236	da-de	synchrone		IP-08-178	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	fr : suppression de quelques balises html à la fin
IP-06-1059	da-de	synchrone		IP-06-1148	da-de	synchrone		IP-06-1149	da-de	synchrone		IP-06-1240	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-1154	da-de	synchrone		IP-06-1159	da-de	synchrone		IP-06-1174	da-de	synchrone		IP-06-1221	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	

Collection 3, suite sur la page 130

Communiqué	Bd	Diagnostic	Détails	Communiqué	Bd	Diagnostic	Détails
IP-08-1923	da-de	suppression	de : suppression d'une série de balises html au milieu	IP-08-771	da-de	synchrone	
	de-fr	synchrone			de-fr	suppression	fr : suppression d'un tableau à la fin
	el-fr	suppression	fr : suppression d'une série de balises html au milieu		el-fr	synchrone	
	en-fr	suppression	fr : suppression d'une série de balises html au milieu		en-fr	synchrone	
	es-fr	suppression	fr : suppression d'une série de balises html au milieu		es-fr	suppression	fr : suppression d'un tableau à la fin
	fi-fr	suppression	fr : suppression d'une série de balises html au début et au milieu		fi-fr	synchrone	
IP-08-405	da-de	synchrone		IP-08-439	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone	
	es-fr	suppression	fr : suppression d'une série de balises html à la fin		es-fr	suppression	fr : suppression d'une série de balises html à la fin
	fi-fr	suppression	fr : suppression d'une série de balises html à la fin		fi-fr	suppression	fi : pas de balises < *ahref...* > à la fin
IP-07-1008	da-de	inversion	paragraphes du détail de la proposition traduits dans un ordre différent				
	de-fr	synchrone					
	el-fr	inversion	idem				
	en-fr	synchrone					
	es-fr	synchrone					
	fi-fr	inversion	idem				
IP-07-1919	da-de	suppression	de : pas le tableau à la fin, ni les quelques lignes entre le 2ème et le 3ème tableau				
	de-fr	suppression	de : pas le tableau à la fin, ni les quelques lignes entre le 2ème et le 3ème tableau				
	el-fr	suppression	el : suppression de quelques lignes entre le 2ème et le 3ème tableau				
	en-fr	suppression	en : suppression de quelques lignes entre le 2ème et le 3ème tableau				
	es-fr	synchrone					
	fi-fr	suppression	fi : suppression de quelques lignes entre le 2ème et le 3ème tableau				
IP-06-1310	da-de	synchrone					
	de-fr	inversion	tri des lignes d'un tableau présentant des répartitions d'aides par Pays				
	el-fr	inversion	idem				
	en-fr	inversion	idem				
	es-fr	inversion	idem				
	fi-fr	inversion	idem				
IP-10-1002	da-de	inversion	annexes listant des projets par pays. de : seuls trois paragraphes concernant les pays germanophones sont traduits et placés en début d'annexe ; da : seul le paragraphe concernant le danemark est traduit. Les restes d'annexe sont en anglais				
	de-fr	inversion	annexes listant des projets par pays. de : seuls trois paragraphes concernant les pays germanophones sont traduits et placés en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais				
	el-fr	inversion	annexes listant des projets par pays.el : seuls deux paragraphes concernant les pays grecophones (2) sont traduits et placés en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais				
	en-fr	inversion	annexes listant des projets par pays. fr, seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais				
	es-fr	inversion	annexes listant des projets par pays. es : seul le paragraphe concernant l'Espagne est traduit et placé en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais				
	fi-fr	inversion	annexes listant des projets par pays.fi : seul le paragraphe concernant la Finlande est traduit et placé en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais				

TABLEAU 46 – Diagnostics manuels sur la collection 3.

Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Détails
IP-05-1097	da-de	synchrone		IP-05-1171	da-de	synchrone		IP-05-1224	da-de	synchrone		IP-05-1157	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	inversion	Listes de projets triées par ordre alphabétique
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	inversion	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	inversion	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	inversion	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	inversion	
IP-05-1457	da-de	synchrone		IP-05-155	da-de	synchrone		IP-05-156	da-de	synchrone		IP-05-1558	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	suppression	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	
IP-05-1672	da-de	synchrone		IP-05-231	da-de	synchrone		IP-05-489	da-de	synchrone		IP-05-419	da-de	synchrone	da : tableau en anglais
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	el : tableau en anglais
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	es : tableau en anglais
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	fi :tableau en anglais
IP-05-572	da-de	synchrone		IP-05-975	da-de	synchrone		IP-06-1186	da-de	synchrone		IP-05-181	da-de	synchrone	tableau en anglais dans les 2 volets
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	tableau en anglais dans les 2 volets, annexe en français dans volet fr
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	tableau en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	tableau en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	tableau en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	tableau en anglais dans les 2 volets
IP-06-1313	da-de	synchrone		IP-06-135	da-de	synchrone		IP-06-1384	da-de	synchrone		IP-05-182	da-de	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets trilingues
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	FAQ bilingue anglais-français les 2 volets = volets bi- et tri-lingues
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bi- et tri-lingues
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bilingues
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bi- et tri-lingues
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	FAQ bilingue anglais-français dans les 2 volets = volets bi- et tri-lingues
IP-06-1434	da-de	synchrone		IP-06-1590	da-de	synchrone		IP-06-1659	da-de	synchrone		IP-06-202	da-de	synchrone	mini annexe en anglais dans les 2 volets, série de balises au milieu dans les 2
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	mini annexe en anglais dans les 2 volets, fr : suppression de la série de balises au milieu
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	mini annexe en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	mini annexe en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	mini annexe en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	mini annexe en anglais dans les 2 volets
IP-06-1676	da-de	synchrone		IP-06-1709	da-de	synchrone		IP-06-1719	da-de	synchrone		IP-06-48	da-de	synchrone	da : annexe en anglais
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	fr : annexe en anglais
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	annexe en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	annexe en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	annexe en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	annexe en anglais dans les 2 volets
IP-06-1818	da-de	synchrone		IP-06-1862	da-de	synchrone		IP-06-442	da-de	synchrone		IP-06-527	da-de	suppression	de : suppression d'un paragraphe
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	idem
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-252	da-de	synchrone		IP-06-359	da-de	synchrone		IP-06-400	da-de	synchrone		IP-06-684	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-739	da-de	synchrone		IP-06-788	da-de	synchrone		IP-06-803	da-de	synchrone		IP-06-816	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	

Tableau 47 – Diagnostics manuels sur la Collection Transport.

Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Détails
IP-05-1217	da-de	synchrone		IP-05-1239	da-de	synchrone		IP-05-1514	da-de	suppression	de : suppression de balises < *aname*… >	IP-05-1157	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	inversion	Listes de projets triées par ordre alphabétique
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	inversion	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	inversion	
	es-fr	synchrone			es-fr	synchrone			es-fr	suppression	fr : suppression de balises < *aname*… >		es-fr	inversion	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	inversion	
IP-05-1603	da-de	synchrone		IP-05-544	da-de	synchrone		IP-05-901	da-de	synchrone		IP-06-1515	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	fr : absence d'une série de balises < *aname*… >
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	fr : absence d'une série de balises < *aname*… >
IP-07-435	da-de	synchrone		IP-07-453	da-de	synchrone		IP-07-668	da-de	synchrone		IP-07-870	da-de	suppression	da : suppression d'une série de balises html
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-08-1397	da-de	synchrone		IP-08-1492	da-de	synchrone		IP-08-425	da-de	synchrone		IP-08-1049	da-de	synchrone	annexes en anglais
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	annexes en anglais
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	annexes en anglais
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	annexes en anglais
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	annexes en anglais
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	annexes en anglais
IP-08-451	da-de	synchrone		IP-08-487	da-de	synchrone		IP-08-537	da-de	synchrone		IP-08-1129	da-de	synchrone	tableaux annexes multilingues
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	tableaux annexes multilingues
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	tableaux annexes multilingues
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	tableaux annexes multilingues
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	tableaux annexes multilingues
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	tableaux annexes multilingues
IP-08-1492	da-de	synchrone		IP-08-425	da-de	synchrone		IP-08-451	da-de	synchrone		IP-08-1144	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	suppression	+ une légende en grec
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	suppression	+ deux légendes en finnois
IP-08-718	da-de	synchrone		IP-08-487	da-de	synchrone		IP-08-537	da-de	synchrone		IP-08-1422	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	el : annexe en anglais
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-07-1177	da-de	synchrone		IP-07-1202	da-de	synchrone		IP-07-1227	da-de	synchrone		IP-06-386	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-07-696	da-de	synchrone		IP-08-1169	da-de	synchrone		IP-08-1276	da-de	synchrone		IP-06-1059	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-07-1445	da-de	synchrone		IP-07-1741	da-de	synchrone		IP-07-311	da-de	synchrone					
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone					
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone					
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone					
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone					
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone					

Collection Téléphone, suite sur la page 133

Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Détails
IP-08-803	da-de	synchrone						IP-08-618	da-de	synchrone	tableaux annexes en anglais				
	de-fr	synchrone							de-fr	synchrone	tableaux annexes en anglais				
	el-fr	synchrone							el-fr	synchrone	tableaux annexes en anglais				
	en-fr	synchrone							en-fr	synchrone	tableaux annexes en anglais				
	es-fr	suppression	+3 lignes dans l'annexe espagnol						es-fr	synchrone	tableaux annexes en anglais				
	fi-fr	synchrone							fi-fr	synchrone	tableaux annexes en anglais				
IP-07-1079	da-de	synchrone	annexe en anglais dans les 2 volets					IP-06-978	da-de	synchrone	série de balises < *ahref*... > identiques dans les deux volets au 2/3 des docs				
	de-fr	synchrone	annexe en anglais dans les 2 volets						de-fr	synchrone	série de balises < *ahref*... > identiques dans les deux volets au 2/3 des docs				
	el-fr	synchrone	annexe en anglais dans les 2 volets						el-fr	synchrone	série de balises < *ahref*... > identiques dans les deux volets au 2/3 des docs				
	en-fr	synchrone	annexe en anglais dans les 2 volets						en-fr	synchrone	série de balises < *ahref*... > identiques dans les deux volets au 2/3 des docs				
	es-fr	synchrone	annexe en anglais dans les 2 volets						es-fr	synchrone	série de balises < *ahref*... > identiques dans les deux volets au 2/3 des docs				
	fi-fr	synchrone	annexe en anglais dans les 2 volets						fi-fr	synchrone	série de balises < *ahref*... > identiques dans les deux volets au 2/3 des docs				
IP-08-1016	da-de	synchrone	da : annexe da/en					IP-07-139	da-de	suppression	de,da : suppression de balises < *aname*... >				
	de-fr	synchrone	fr : annexe fr/en						de-fr	suppression	de : suppression de balises < *aname*... >				
	el-fr	synchrone	fr : annexe fr/en, el : annexe en/el						el-fr	suppression	el : suppression de balises < *aname*... >				
	en-fr	synchrone	fr : annexe fr/en						en-fr	suppression	en : suppression de balises < *aname*... >				
	es-fr	suppression	suppression à la fin du volet espagnol + annexe es/en						es-fr	suppression	es : suppression de balises < *aname*... >				
	fi-fr	synchrone	fr : annexe fr/en, fi : annexe fi/en						fi-fr	suppression	fi : suppression de balises < *aname*... >				

Tableau 48 – Diagnostics manuels sur la Collection Téléphone.

Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Dét.	Communiqué	Bd	Diagnostic	Détails
IP-05-156	da-de	synchrone		IP-05-225	da-de	synchrone		IP-05-292	da-de	synchrone		IP-06-358	da-de	synchrone	annexes en anglais
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	annexes en anglais
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	annexes en anglais
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	annexes en anglais
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	annexes en anglais
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	annexes en anglais
IP-05-389	da-de	synchrone		IP-05-460	da-de	synchrone		IP-05-489	da-de	synchrone		IP-06-377	da-de	synchrone	petit bi-document (1,6K -1,8K)
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	petit bi-document (1,8K - 2K)
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	petit bi-document (2,9K - 2K)
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	petit bi-document (1,7K - 2K)
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	petit bi-document (1,7K - 2K)
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	petit bi-document (1,8K - 2K)
IP-05-513	da-de	synchrone		IP-05-606	da-de	synchrone		IP-05-808	da-de	synchrone		IP-06-527	da-de	suppression	de : suppression d'un paragraphe
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	suppression	de : suppression d'un paragraphe
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-1590	da-de	synchrone		IP-06-1659	da-de	synchrone		IP-06-1676	da-de	synchrone		IP-07-1498	da-de	synchrone	petit bi-document (1,9K - 2K)
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	petit bi-document (2K - 2,3K)
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	petit bi-document (3,6K - 2,3K)
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	petit bi-document (1,8K - 2,3K)
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	petit bi-document (2,2K - 2,3K)
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	petit bi-document (2K - 2,3K)
IP-06-396	da-de	synchrone		IP-06-400	da-de	synchrone		IP-06-442	da-de	synchrone		IP-07-1504	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	suppression	es : pas de titre
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-06-788	da-de	synchrone		IP-07-1449	da-de	synchrone		IP-07-1537	da-de	synchrone		IP-07-1783	da-de	synchrone	da : tableau en anglais
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	el : tableau en anglais
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	es : tableau en anglais
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	fi : tableau en anglais
IP-07-1543	da-de	synchrone		IP-07-1576	da-de	synchrone		IP-07-1663	da-de	synchrone		IP-07-1854	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-07-1720	da-de	synchrone		IP-07-1728	da-de	synchrone		IP-07-1761	da-de	synchrone		IP-07-1913	da-de	synchrone	tableaux en anglais dans les 2 volets
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	tableaux en anglais dans les 2 volets
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	tableaux en anglais dans les 2 volets
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	tableaux en anglais dans les 2 volets
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	tableaux en anglais dans les 2 volets
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	tableaux en anglais dans les 2 volets
IP-07-1968	da-de	synchrone		IP-07-202	da-de	synchrone		IP-07-204	da-de	synchrone		IP-07-387	da-de	synchrone	
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone			de-fr	synchrone	
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone			el-fr	synchrone	
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone			en-fr	synchrone	
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone			es-fr	synchrone	
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone	
IP-07-440	da-de	synchrone		IP-07-453	da-de	synchrone		IP-07-514	da-de	synchrone					
	de-fr	synchrone			de-fr	synchrone			de-fr	synchrone					
	el-fr	synchrone			el-fr	synchrone			el-fr	synchrone					
	en-fr	synchrone			en-fr	synchrone			en-fr	synchrone					
	es-fr	synchrone			es-fr	synchrone			es-fr	synchrone					
	fi-fr	synchrone			fi-fr	synchrone			fi-fr	synchrone					

Collection Santé, suite sur la page 135

Communiqué	Bd	Diagnostic	Détails
IP-10-1002	da-de	inversion	annexes listant des projets par pays. de : seuls trois paragraphes concernant les pays germanophones sont traduits et placés en début d'annexe ; da : seul le paragraphe concernant le danemark est traduit. Les restes d'annexe sont en anglais
	de-fr	inversion	annexes listant des projets par pays. de : seuls trois paragraphes concernant les pays germanophones sont traduits et placés en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais
	el-fr	inversion	annexes listant des projets par pays.el : seuls deux paragraphes concernant les pays grecophones (2) sont traduits et placés en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais
	en-fr	inversion	annexes listant des projets par pays. fr, seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais
	es-fr	inversion	annexes listant des projets par pays. es : seul le paragraphe concernant l'Espagne est traduit et placé en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais
	fi-fr	inversion	annexes listant des projets par pays.fi : seul le paragraphe concernant la Finlande est traduit et placé en début d'annexe ; fr : seuls les paragraphes concernant les pays francophones (2) sont traduits et placés en début d'annexe. Les restes d'annexe sont en anglais

Tableau 49 – Diagnostics manuels sur la Collection Santé.

BIBLIOGRAPHIE

Haneen Abudayeh : *Traduire l'émotion dans le discours politique*. Thèse de doctorat, Caen Basse-Normandie, 2010. (Cité à la page 7.)

Michel Ballard : À propos de l'erreur en traduction. *Revue des lettres et de traduction.*, 5:51–65, 1999. (Cité à la page 8.)

Cédric Becquey : Description, discussion, extension de la notion de parallélisme. http ://www.mae.u-paris10.fr/siteaci/aci/NiveauIII/parallelisme/notion.html, 2003a. URL http://www.mae.u-paris10.fr/siteaci/aci/NiveauIII/parallelisme/notion.html. (Cité à la page 28.)

Cédric Becquey : Le parallélisme. http ://www.mae.u-paris10.fr/siteaci/aci/NiveauII/parallelisme.html, 2003b. URL http://www.mae.u-paris10.fr/siteaci/aci/NiveauII/parallelisme.html. (Cité aux pages 28 et 30.)

Ismaïl Biskri et Sylvain Delisle : Les n-grams de caractères pour l'extraction de connaissances dans des bases de données textuelles multilingues. *In Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, 2-5 juillet*, Tours, France, 2001. URL http://www.uqtr.ca/~biskri/. (Cité à la page 56.)

Julien Bourdaillet et Jean-Gabriel Ganascia : Alignements monolingues avec déplacements. *In Actes des 14e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 303–312, Toulouse, France, 2007. (Cité aux pages 43 et 70.)

Romain Brixtel : *Alignement endogène de documents, une approche multilingue et multi-échelle*. Thèse de doctorat, Université de Caen/Basse-Normandie, 2011. (Cité aux pages 44, 45, 48 et 55.)

Romain Brixtel, Mathieu Fontaine, Boris Lesner, Cyril Bazin et Romain Robbes : Language-Independent clone detection applied to plagiarism detection. *In 2010 10th IEEE Working Conference on Source Code Analysis and Manipulation*, pages 77–86, Timisoara, Romania, septembre 2010. URL http://ieeexplore.ieee.org/Xplore/login.jsp?url=http.ieee.orgecision=-203. (Cité à la page 43.)

Romain Brixtel, Boris Lesner, Guillaume Bagan et Cyril Bazin : De la mesure de similarité de codes sources vers la détection de plagiat : le « Pomp-O-Mètre ». *In 7e Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication, 16-18 novembre*, page 8, Avignon, France, 2009. (Cité à la page 34.)

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer et Paul S. Roossin : A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, 1990. URL http://portal.acm.org/citation.cfm?id=92858.92860&coll=Portal&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité à la page 41.)

Peter F. Brown, Jennifer C. Lai et Robert L. Mercer : Aligning sentences in parallel corpora. *In Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176, Berkeley, California, 1991. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=981344.981366&coll=Portal&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité aux pages 32, 36, 37, 38 et 40.)

John C. Catford : *A Linguistic Theory of Translation : an Essay on Applied Linguistics*. Oxford University Press, London, 1965. (Cité à la page 8.)

Chirine Chamsine : La traduction des émotions. Mémoire de master conjoint franco-hellénique mention sciences du langage, spécialité sciences de la traduction : traductologie et sciences cognitives, Université de Caen Basse-Normandie, Caen, France, 2005. (Cité à la page 7.)

Jason S. Chang et Mathis H. Chen : An alignment method for noisy parallel corpora based on image processing techniques. *In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 297–304, Madrid, Spain, 1997. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=979617.979655&coll=GUIDE&dl=GUIDE&CFID=78470726&CFTOKEN=79586012. (Cité à la page 43.)

Stanley F. Chen : Aligning sentences in bilingual corpora using lexical information. *In Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, 1993. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=981574.981576&coll=GUIDE&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité aux pages 36, 38 et 40.)

Yun-Chuang Chiao, Olivier Kraif, Dominique Laurent, Thi Minh Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis et Wajdi Zaghouani : Evaluation of multilingual text alignment systems : the ARCADE II project. *In 5th international Conference on Language Resources and Evaluation*, Genoa/Italy, 2006. URL

http://hal.inria.fr/inria-00115670_v1/. (Cité aux pages 32 et 40.)

Kenneth Ward Church : Char_align : a program for aligning parallel texts at the character level. *In Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, page 1–8, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/981574.981575. ACM ID : 981575. (Cité aux pages 35, 36, 38, 43 et 48.)

Kenneth Ward Church et Jonathan Isaac Helfman : Dotplot : A program for exploring Self-Similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics*, 2(2):153–174, 1993. ISSN 10618600. URL http://www.jstor.org/stable/1390697. ArticleType : research-article / Full publication date : Jun., 1993 / Copyright © 1993 American Statistical Association, Institute of Mathematical Statistics and Interface Foundation of America. (Cité à la page 43.)

Guylaine Cochrane : Le foisonnement, phénomène complexe. *TTR : traduction, terminologie, rédaction*, 8(2), 2007. URL http://id.erudit.org/iderudit/037222ar. (Cité aux pages 8 et 10.)

Maxime Crochemore, Christophe Hancart et Thierry Lecroq : *Algorithms on Strings*. Cambridge University Press, 1 édition, 2007. ISBN 0521848997. (Cité à la page 78.)

Fabien Cromières : Sub-sentential alignment using substring co-occurrence counts. *In Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, pages 13–18, Sydney, Australia, 2006. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1557860. (Cité aux pages 42, 48 et 56.)

Ido Dagan et Ken Church : Termight : identifying and translating technical terminology. *In Proceedings of the fourth conference on Applied natural language processing*, ANLC '94, page 34–40, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/974358.974367. ACM ID : 974367. (Cité à la page 41.)

Ido Dagan, Kenneth W Church et William A Gale : Robust bilingual word alignment for machine aided translation. *In Proceedings of the Workshop on Very Large Corpora*, 1:1—8, 1993. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.4941. (Cité à la page 41.)

Béatrice Daille, Eric Gaussier et Jean-Marc Langé : Towards automatic extraction of monolingual and bilingual terminology. *PROCEEDINGS OF COLING 94*, pages 515—521, 1994. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.9536. (Cité à la page 41.)

Leyla Dakhli : Le multilinguisme est un humanisme. *La Vie des idées*, 2009. ISSN ISSN : 2105-3030. URL http://www.laviedesidees.fr/Le-multilinguisme-est-un-humanisme.html. (Cité à la page 5.)

Marc Damashek : Gauging similarity with n-Grams : Language-Independent categorization of text. *Science*, 267:843–848, 1995. (Cité à la page 56.)

Fathi Debili et Elyès Sammouda : Aligning sentences in bilingual texts : French-English and French-Arabic. *In Proceedings of the 14th conference on Computational linguistics - Volume 2*, pages 517–524, Nantes, France, 1992. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=992151&dl=GUIDE&coll=GUIDE&CFID=78336177&CFTOKEN=78125505. (Cité aux pages 34, 36 et 39.)

Ted Dunning : Statistical identification of language. Technical report MCCS 94-273, New Mexico State University, New Mexico, 1994. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.1958. (Cité à la page 56.)

Christine Durieux : Le foisonnement en traduction technique d'anglais en français. *Meta*, 35(1):55–60, 1990. ISSN 0026-0452. URL http://id.erudit.org/iderudit/002689ar. (Cité aux pages 8, 10 et 150.)

Hervé Déjean et Eric Gaussier : Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Alignement lexical dans les corpus multilingues(Numéro spécial), 2002. (Cité à la page 149.)

Jessica Enright et Grzegorz Kondrak : A fast method for parallel document identification. *In Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers on XX*, pages 29–32, Rochester, New York, 2007. Association for Computational Linguistics. URL http://webdocs.cs.ualberta.ca/~kondrak/papers/hlt07.pdf. (Cité à la page 33.)

Tomaz Erjavec, Nancy Ide, Vladimir Petkevic, Jean Véronis et Av. Robert Schuman : Multext-East : Multilingual text tools and corpora for central and eastern european languages. Technical Annex Cop 106, 1995. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.8485. (Cité à la page 32.)

Christian Fluhr, F Bisson et F Elkateb : Mutual benefit of sentence/-word alignment and crosslingual information retrieval. *In Parallel text processing : Alignment and use of translation corpora*. Dordrecht : Kluwer Academic Publishers, j. véronis (ed.) édition, 2000. (Cité à la page 42.)

Pascale Fung et Kenneth Ward Church : K-vec : a new approach for aligning parallel texts. *In Proceedings of the 15th conference on Computational linguistics - Volume 2*, pages 1096–1102, Kyoto, Japan, 1994. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=991328. (Cité aux pages 36, 41, 43 et 69.)

Pascale Fung et Kathleen Mckeown : Aligning noisy parallel corpora across language groups : Word pair feature matching by dynamic time warping. *In Proceedings of the First Conference of the Association for Machine Translation in the Americas, 81–88*, pages 81—88, 1994. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.4548. (Cité à la page 43.)

William A. Gale et Kenneth W. Church : Identifying word correspondence in parallel texts. *In Proceedings of the workshop on Speech and Natural Language*, pages 152–157, Pacific Grove, California, 1991. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=112405.112428&coll=Portal&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité aux pages 36 et 41.)

William A. Gale et Kenneth W. Church : A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19 (1):75–102, 1993. URL http://portal.acm.org/citation.cfm?id=972450.972455&coll=GUIDE&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité aux pages 32, 36, 37, 38 et 40.)

Éric Gaussier : Flow network models for word alignment and terminology extraction from bilingual corpora. *In proceedings of the joint 17th international conference on computational linguistics and 26th annual meeting of the Association for Computational Linguistics*, pages 444—450, 1998. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.1725. (Cité à la page 41.)

Kim Gerdes : L'alignement pour les pauvres : Adapter la bonne métrique pour un algorithme dynamique de dilatation temporelle pour l'alignement sans ressources de corpus bilingues. *In 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, France, 2008. (Cité à la page 37.)

Emmanuel Giguet : Multi-grained alignment of parallel texts with endogenous resources. *In In Proceedings of the Recent Advances in*

Natural Language Processing (RANLP) International Workshop "New Trends in Machine Translations", pages 12–17, Borovets, Bulgaria, 2005. (Cité aux pages 9 et 45.)

Emmanuel Giguet et Marianna Apidianaki : Alignement d'unités textuelles de taille variable. *In 4èmes Journées de la Linguistique de Corpus*, Lorient, France, 2005. URL http://hal.archives-ouvertes.fr/index.php?halsid=50le6pgjvcg7ral86p9i2qt010&view_this_doc=halshs-00202140&version=1. (Cité à la page 42.)

Emmanuel Giguet et Pierre-Sylvain Luquet : Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. *In Proceedings of the COLING/ACL on Main conference poster sessions*, pages 271–278, Sydney, Australia, 2006. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1273108. (Cité aux pages 41 et 45.)

Brian Harris : La traductologie, la traduction naturelle, la traduction automatique et la sémantique. *Cahier de linguistique*, 2:133–146, 1973. ISSN 0315-4025. URL http://id.erudit.org/iderudit/800013ar. (Cité à la page 8.)

Brian Harris : Bi-text, a new concept in translation theory. *Language Monthly (UK)*, 54, 1988. URL http://en.wikipedia.org/wiki/Parallel_text. (Cité aux pages 8 et 149.)

Reinhard Rudolf Kard Hartmann : *Contrastive Textology. Comparative Discourse Analysis in Applied Linguistics*. Numéro 5 *in* Studies in Descriptive Linguistics. Groos Verlag, Heidelberg, 1980. (Cité à la page 30.)

Stéphane Huet, Julien Bourdaillet et Philippe Langlais : Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. *In Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France, 2009. (Cité à la page 34.)

Nancy Ide et Jean Véronis : MULTEXT : Multilingual text tools and corpora. *In Proceedings of the 15th conference on Computational linguistics - Volume 1*, pages 588–592, Kyoto, Japan, 1994. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=991990. (Cité à la page 32.)

H Isahara et M Hiruno : Japanese-English aligned bilingual corpora., 2000. (Cité à la page 33.)

Roman Jakobson : *Linguistique et poétique*. Numéro 1 *in* Essais de linguistique générale. Les éditions de minuit, 1963. (Cité aux pages 28 et 29.)

Michèle JARDINO : Identification des auteurs de textes courts avec des n-grammes de caractères. *In Actes des 8èmes Journées internationales d'Analyse statistique des Données Textuelles*, Besançon, France, 2006. (Cité à la page 56.)

Denis JUHEL : Prolixité et qualité des traductions. *Meta*, 44(2):238–249, 1999. ISSN 0026-0452. URL http://id.erudit.org/iderudit/003275ar. (Cité à la page 8.)

Martin KAY et Martin RÖSCHEISEN : Text-translation alignment. *Comput. Linguist.*, 19(1):121–142, 1993. URL http://portal.acm.org/citation.cfm?id=972450.972457&coll=GUIDE&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité aux pages 36 et 39.)

Judith KLAVANS et Evelyne TZOUKCRMANN : The BICORD system : combining lexical information from bilingual corpora and machine readable dictionaries. *In Proceedings of the 13th Annual Meeting of the Association of Computational Linguistics*, page 174–179, 1990. (Cité à la page 33.)

Olivier KRAIF : Architecture d'un système d'alignement : étude pour une intégration optimale des indices d'alignement. *In Actes des Journées internationales de linguistique appliquée*, pages 161–164, faculté des Lettres Arts et Sciences humaines, Université de Nice Sophia Antipolis, 1999. (Cité aux pages 36, 38 et 39.)

Olivier KRAIF : *Constitution et exploitation de bi-textes pour l'aide à la traduction*. Thèse de doctorat, Université de Nice Sophia- Antipolis, 2001. (Cité à la page 33.)

Juha KÄRKKÄINEN et Peter SANDERS : Simple linear work suffix array construction. *In* Jos C. M. BAETEN, Jan Karel LENSTRA, Joachim PARROW et Gerhard J. WOEGINGER, éditeurs : *Automata, Languages and Programming*, volume 2719, pages 943–955. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-40493-4. URL http://www.springerlink.com/content/0nyb22e5amj4rac4/. (Cité à la page 78.)

Philippe LANGLAIS : Alignement de corpus bilingues : intérêts, algorithmes et évaluations. *Bulletin de Linguistique Appliquée et Générale*, numéro Hors Série:245–254, 1997. URL http://www.iro.umontreal.ca/~felipe/Papers/fractal97.ps. (Cité aux pages 34, 36, 37, 40 et 43.)

Philippe LANGLAIS et Marc EL-BÈZE : Alignement de corpus bilingues : algorithmes et évaluation. *In 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la langue de l'AUPELF-UREF (JST)*, Avignon, France, avril 1997. (Cité à la page 40.)

Lucie Langlois : *Bitexte, bi-concordance et collocation*. Thèse de doctorat, Université d'Ottawa, Canada, 1996. URL http://www.dico.uottawa.ca/theses/langlois/introduction.htm. (Cité à la page 33.)

J-M Langé et Eric Gaussier : Alignement de corpus multilingues au niveau des phrases = multilingual corpora alignment at sentence level. *TAL, Traitement Automatique des Langues*, 36(1-2):67–80, 1995. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=3282436. (Cité à la page 34.)

Adrien Lardilleux : L'alignement sous-phrastique multilingue pour les nuls. *In 7ème Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication, 16-18 novembre*, Avignon, France, 2009. (Cité à la page 45.)

Adrien Lardilleux : *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. Thèse de doctorat, Université de Caen/Basse-Normandie, 2010. URL http://hal.archives-ouvertes.fr/index.php?halsid=rsgsimesspm32r8ug106nbpr03&view_this_doc=tel-00520787&version=1. (Cité à la page 42.)

Dekang Lin, Shaojun Zhao, Benjamin Van Durme et Marius Paşca : Mining parenthetical translations from the web by word alignment. *In Proceedings of ACL-08 : HLT*, page 994–1002, Columbus, Ohio, juin 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P08/P08-1113. (Cité à la page 34.)

P. Majumder, M. Mitra et B. B Chaudhuri : N-gram : a language independent approach to IR and NLP. *In Proceedings of the international Conference on Universal Knowledge and Language, 25-29 novembre*, 2002. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.8275. (Cité à la page 56.)

Paul Mcnamee et James Mayfield : Character N-Gram tokenization for european language text retrieval. *Information Retrieval*, 7:73–97, 2004. ISSN 1386-4564. URL http://portal.acm.org/citation.cfm?id=961294.961313. ACM ID : 961313. (Cité aux pages 56 et 73.)

I. Dan Melamed : Automatic evaluation and uniform filter cascades for inducing N-Best translation lexicons. *In proceedings of the third workshop on very large corpora*, pages 184—198, 1995. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.7877. (Cité à la page 41.)

I. Dan Melamed : Bitext maps and alignment via pattern recognition. *Comput. Linguist.*, 25(1):107–130, 1999. URL http://portal.acm.org/citation.cfm?id=973215.973218&coll=

Portal&dl=GUIDE&CFID=78818668&CFTOKEN=17474915. (Cité à la page 43.)

I. Dan Melamed : Models of translational equivalence among words. *Computational linguistics*, 26:221—249, 2000. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.9615. (Cité à la page 40.)

Robert C. Moore : Fast and accurate sentence alignment of bilingual corpora. *In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation : From Research to Real Users*, pages 135–144. Springer-Verlag, 2002. ISBN 3-540-44282-0. URL http://portal.acm.org/citation.cfm?id=749407. (Cité à la page 36.)

Yayoi Nakamura-Delloye : Méthodes d'alignement des propositions : un défi aux traductions croisées. *In Actes de la 14ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, 12-15 juin*, Toulouse, France, 2007. (Cité à la page 36.)

Franck Neveu : *Dictionnaire des sciences du langage*. Armand Colin, 2004. ISBN 2200263783. (Cité aux pages 10, 17, 19 et 56.)

E. A Nida : *Toward a science of translation*. Brill, Leiden, 1964. (Cité à la page 8.)

Britta Nord : *Hilfsmittel beim Übersetzen : Eine empirische Studie zum Rechercheverhalten professioneller Übersetzer*. Peter Lang, Frankfurt am Main, 2002. ISBN 3631393318. (Cité à la page 8.)

Christiane Nord : TRACI : The trainee translator's card index a self-made tool for acquiring and enhancing translation competence. *Les Cahiers du GEPE*, Outils de traduction - outils du traducteur ?(2), 2010. URL http://www.cahiersdugepe.fr/index.php?id=1318. (Cité aux pages 7 et 30.)

Franz Josef Och et Hermann Ney : A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29 (1):19–51, 2003. URL http://portal.acm.org/citation.cfm?id=778822.778824&coll=GUIDE&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité à la page 41.)

François Ost : *Traduire : Défense et illustration du multilinguisme*. Fayard, 2009. ISBN 2213643660. (Cité aux pages 5 et 8.)

Alexandre Patry et Philippe Langlais : Automatic identification of parallel documents with light or without linguistic resources. *In Canadian Conference on Artificial Intelligence*, pages 354–365, 2005. URL http://www-etud.iro.umontreal.ca/~patryale/papers/patry_langlais_2005_ai.pdf. (Cité à la page 33.)

Emmanuel Planas : Extending translation memories. *Proceedings of the 5th European Association for Machine*, 2000. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.9756. (Cité à la page 34.)

Philip Resnik et I. Dan Melamed : Semi-automatic acquisition of domain-specific translation lexicons. *In Proceedings of the fifth conference on Applied natural language processing*, ANLC '97, page 340–347, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/974557.974607. ACM ID : 974607. (Cité à la page 41.)

Philip Resnik et Noah A. Smith : The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, septembre 2003. ISSN 0891-2017. URL http://dx.doi.org/10.1162/089120103322711578. (Cité à la page 48.)

Calliopi Sachtouri : Etude comparative des chaînes anaphoriques dans vingt langues européennes. Mémoire de master conjoint franco-hellénique mention sciences du langage, spécialité sciences de la traduction : traductologie et sciences cognitives, université de Caen Basse-Normandie et Université ionienne de Corfou (Grèce), Caen, France, 2006. (Cité à la page 19.)

Fatiha Sadat, George Foster et Roland Kuhn : Système de traduction automatique statistique combinant différentes ressources. *In Actes de la 16ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, 10-13 avril*, Leuven, Belgique, 2006. URL http://www.iro.umontreal.ca/~foster/papers/taln06.pdf. (Cité à la page 32.)

Michel Simard : Text-Translation alignment : Three languages are better than two. *IN PROC. OF EMNLP/VLC*, pages 2—11, 1999. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.6716. (Cité à la page 45.)

Michel Simard, George F. Foster et Pierre Isabelle : Using cognates to align sentences in bilingual corpora. *In Proceedings of the 4th conference of the Centre for Advanced Studies on Collaborative research : distributed computing - Volume 2*, pages 1071–1082, Toronto, Ontario, Canada, 1992. IBM Press. URL http://portal.acm.org/citation.cfm?id=962367.962411&coll=GUIDE&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité aux pages 36 et 38.)

Frank Smadja, Kathleen R McKeown et Vasileios Hatzivassiloglou : Translating collocations for bilingual lexicons : a statistical approach. *Computational Linguistics*, 22:1–38, mars 1996. ISSN 0891-2017. URL http://portal.acm.org/citation.cfm?id=234285.234287. ACM ID : 234287. (Cité à la page 41.)

Bernd Spillner : Textsorten im sprachvergleich. ansätze zu einer kontrastiven textologie. *In Kontrastive Linguistik und Übersetzungswissenschaft*, pages 239–250. KÜHLWEIN Wolfgang, THOME Gisela, WILSS Wolfram, München, Fink, 1981. (Cité à la page 30.)

Dan Tufiş et Ana-Maria Barbu : Lexical token alignment : Experiments, results and application. *In Proceedings of LREC-2002*, pages 458—465, 2002. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.469. (Cité à la page 41.)

J Vergne et E Giguet : Regards théoriques sur le tagging. *In Proceedings of the conference Le Traitement Automatique des Langues Naturelles*, 1998. (Cité à la page 100.)

Jean Véronis : Evaluation of parallel text alignment systems : the ARCADE project. *In Parallel text processing : Alignment and use of translation corpora*, pages 369–388. J. Véronis, Dordrecht, kluwer academic publishers édition, 2000. (Cité à la page 32.)

Jean Véronis et Philippe Langlais : ARCADE : Evaluation de systèmes d'alignement de textes multilingues. *Lettre de l'ELRA*, 4(1), 1999. (Cité aux pages 32 et 40.)

Dekai Wu : Aligning a parallel English-Chinese corpus statistically with lexical criteria. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 80–87, Las Cruces, New Mexico, 1994. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=981732.981744&coll=GUIDE&dl=GUIDE&CFID=76577594&CFTOKEN=73477001. (Cité à la page 34.)

Dekai Wu et Xuanyin Xia : Learning an English-Chinese lexicon from a parallel corpus. *In Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206—213, 1994. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.710. (Cité à la page 41.)

Yu Zhou, Chengqing Zhong et Bo Xu : Bilingual chunk alignment in statistical machine translation. *In Proceedings of the 2004 IEEE international conference on systems, man and cybernetics, 10-13 october*, The Hague, Netherlands, 2004. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=17523633. (Cité à la page 36.)

Maria Zimina : Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles. *In Actes des 7èmes Journées scientifiques du Réseau de chercheurs "Lexicologie, Terminologie, Traduction"*, pages 175–186, institut supérieur de traducteurs et interprètes (ISTI), Bruxelles (Belgique), 2006. (Cité aux pages 8 et 79.)

Maria ZIMINA-POIROT : *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Sciences du langage, Université Paris 3 - Sorbonne Nouvelle, 2004. URL http://hal.archives-ouvertes.fr/index.php?halsid=7hubfdttvo7pmuoousu7ulelg7&view_this_doc=tel-00008311&version=1. (Cité à la page 41.)

GLOSSAIRE

Aligner ou apparier : « Aligner ou apparier deux textes dont l'un est une traduction de l'autre, consiste à mettre en relation des unités logiques qui se correspondent dans les deux textes. Ces unités logiques peuvent être de diverses sortes : paragraphes et structures logiques du document, phrases, syntagmes, mots… » (Harris, 1988).

Alignement : Un alignement est une correspondance sémantique locale, prise en contexte. Il met en correspondance une occurrence d'une unité donné dans une langue avec une occurrence d'une unité d'une autre langue.

Appariement : Un appariement est une correspondance sémantique fortement généralisée telle qu'on en trouve dans un dictionnaire. Par extension, l'appariement, en tant que méthode, est la mise en correspondance de deux chaînes de caractères répétées entre des multidocuments, i.e des populations, grâce à leur similitude de répartitions, i.e. effectifs et positions.

Bi-texte : Ensemble constitué d'un texte original en langue source et d'une de ses traductions, terme introduit par Harris (1988).

Cognats : Chaîne de caractères qui reste invariante du point de vue graphique d'une langue à une autre : noms propres, chiffres, sigles…

Corpus comparables : Ensemble de documents non traduits présentant une homogénéité d'un point de vue thématique, chronologique et de leur registre. (Déjean et Gaussier, 2002) en donnent la définition : « Deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1. »

Corpus parallèles : Ensemble de documents composé d'originaux et de leurs traductions.

Correspondances traductionnelles : Ensemble d'unités textuelles en relation d'équivalence traductionnelle. On distingue des correspondances traductionnelles bi- ou quasi-univoques et des correspondances multiples.

Correspondances traductionnelles bi- ou quasi-univoques : On parle de correspondances bi-univoques lorsqu'au sein d'un corpus bilingue, par exemple, un mot source est toujours traduit par le même mot cible dans l'autre langue et qu'ils présentent donc des similtudes de fréquence totale.

Correspondances traductionnelles multiples : On parle de correspondances bi-univoques lorsqu'au sein d'un corpus bilingue, par exemple, un mot source possède différents équivalents dans la langue cible.

Dotplot ou matrice : Le dot plot est un outil graphique servant à étudier la similarité entre deux séquences, il est principalement utilisé en bio-informatique.

Foisonnement : « En traduction, le foisonnement est la prolifération de mots en surnombre, c'est l'augmentation du volume du texte d'arrivée par rapport au texte de départ » (Durieux, 1990).

Grain : Taille d'une unité linguistique donné. Les grains s'emboîtent les uns dans les autres selon une hiérarchie de grains : le grain document, le grain phrase, le grain mot…

Hapax : Du grec « ἅπαξ λεγόμενον » */hápax legómenon/* '[dit] une seule fois', le terme hapax signifie un mot qui n'apparaît qu'une fois dans un texte ou un corpus (de façon monolingue dans nos travaux).

Individu : Occurrence d'un n-gramme de caractère répété.

Intertextualité : Ensemble des relations qu'un texte entretient avec un ou plusieurs autres textes (citations, allusions, références). « Tout texte se situe à la jonction de plusieurs textes dont il est à la fois la relecture, l'accentuation, la condensation, le déplacement et la profondeur. » (Philippe Sollers, 1980)

Multi-document : Ensemble constitué d'un document original en langue source et plusieurs de ses traductions.

Multi-zone : Ensemble constitué d'une zone de texte en langue source et plusieurs de ses traductions.

Population : Ensemble constitué de l'ensemble des occurrences d'un n-gramme de caractère.

Précision : (Mesure de) calcul statistique qui reflète la proportion de bi-document correctement diagnostiqués.

Quasi-bijection : Dans le bi-texte $T1$-$T2$, pour un segment de texte $T1$, il existe dans la majorité des cas un seul candidat, issu de $Fs(T2)$, comme équivalent traductionnel (Fs : Fonction de segmentation).

Quasi-synchronisation : Également appelée quasi-monotonie, signifie que dans le bi-texte $T1$-$T2$, l'ordre des segments de $T1$ respecte, à quelques variations locales près, l'ordre des segments de $T2$.

Segment : Un segment de volet correspond à une portion de volet définie en pourcentage. Dans notre hiérarchie de grains (voir figure 11, page 57), il se situe entre la zone et le N-gramme de caractères. Ainsi, une zone peut comprendre plusieurs segments et un segment plusieurs N-grammes de caractères.

Volet : Document pris comme version, le plus souvent monolingue, d'un multidocument.

Zone : Grain intermédiaire entre le document et les unités sous-phrastiques, la zone est définie en contexte grâce aux segments. Elle est constituée de caractères pouvant en contexte recouvrir plusieurs réalités : du document à la chaîne de caractères en passant par le paragraphe, la phrase, la proposition, l'expression ou le mot. Ainsi, à la façon du bi-texte, une bi-zone correspond à la mise en correspondance de deux zones de textes de deux langues différentes.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

Cette thèse a été composée avec LaTeX 2ε en utilisant
le style classicthesis, disponible via CTAN. La
police principale est *Minion*® d'Adobe™.

RÉSUMÉ

Alignement de documents multilingues sans présupposé de parallélisme

Aujourd'hui les travaux exploitant des documents multilingues se tournent vers l'étude de textes comparables alors même que tous les aspects des documents parallèles n'ont pas été étudiés ni tous les verrous liés aux méthodes d'alignement levés, notamment leur mise en forme et les cas d'inversions et de suppressions au niveau sur-phrastique. Ainsi, nous ne disposons pas à ce jour d'outils permettant de valoriser cette mine d'informations, d'en extraire aussi massivement qu'envisagé des ressources pourtant utiles tant aux traducteurs qu'aux lexicologues.

Nous présentons ici une méthode sans présupposé de parallélisme entre les différents volets d'un multidocument. L'idée essentielle de ces travaux est la suivante : entre deux volets d'un multidocument, il existe des grains qui maximisent le parallélisme, nous les appelons des *multizones*. Celles-ci peuvent recouvrir plusieurs réalités : documents, série de paragraphes, paragraphes, propositions… Ces multizones ne sont pas délimitables de façon ad hoc, il convient de le faire en contexte et de façon indépendante des langues. Á ces fins, nous combinons plusieurs procédés originaux : étudier chaque multidocument au travers d'une collection de multidocuments, exploiter la mise en forme des documents par traitement direct du source ou encore traiter des chaînes de caractères répétées plutôt que des mots.

Notre objectif est double : appariement et alignement, i.e. création de ressources et analyse de documents. Cette méthode requiert peu de supervision, l'ajout d'une nouvelle langue ou le changement de corpus d'entrée ne représentent pas un coût important.

MOTS-CLÉS : traitement automatique des langues, alignement, multilinguisme, parallélisme, collection de multidocuments, multizones, chaînes de caractères répétées.

ABSTRACT

Multilingual document alignment method without assumption of parallelism:

Today the works using multilingual documents are turning to the study of comparable texts even though all aspects of parallel documents have not been studied nor alignment method locks raised, including their formatting and the cases of inversions and deletions at macro level. Thus, to date there is no tools to take benefit from this wealth of information, to extract resources as massively as envisaged, despite their usefulness both for translators and lexicologists…

We present a method without assumption of parallelism between the different components of a multiple document. The basic idea of this work is: between two components of a multi-document, there are grains that maximize the parallelism, we call them *multizones*. They can cover several realities: document, series of paragraphs, paragraphs, proposals… Their boundaries can not be defined in an ad hoc way, it should be done in context and independently of languages. To this end, we combine several original processes: study each multiple document through a collection of multi-document, use the formatting of documents by direct processing of source or process repeated strings rather than words.

The purpose of this work is twofold: matching and alignment, i.e. resource creation and document analysis. This method requires little supervision. Add a new language or change corpus of entry do not represent a significant cost.

KEYWORDS: natural language processing, alignment, multilingualism, parallelism, set of multidocuments, multizones, repeated character N-grams.

Laboratoire GREYC – Campus Côte de Nacre, Boulevard du Maréchal Juin – BP 5186 – 14032 CAEN CEDEX