Isotopies sémantiques pour la vérification de traduction

Ludovic Tanguy, Susan Armstrong, Derek Walker

▶ **To cite this version:**

Ludovic Tanguy, Susan Armstrong, Derek Walker. Isotopies sémantiques pour la vérification de traduction. TALN, 1999, Cargèse, France. ⟨halshs-01322334⟩

HAL Id: halshs-01322334

https://shs.hal.science/halshs-01322334v1

Submitted on 27 May 2016

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.

HAL Authorization

Isotopies sémantiques pour la vérification de traduction

Ludovic Tanguy, Susan Armstrong, Derek Walker

ISSCO - Université de Genève
54 Route des Acacias - 1227 Genève - Suisse
Tél/Fax : (+41) 22 705 71 12 / (+41) 22 300 10 86
Ludovic.Tanguy@issco.unige.ch

Résumé

À des fins d'automatisation de la vérification de traduction, les méthodes traditionnelles se basent généralement sur un fort niveau de littéralité dans le style de la traduction. En faisant appel à des bases terminologiques multilingues et des algorithmes d'alignement de textes parallèles, il est possible de vérifier dans un travail de traduction le respect de normes strictes, sous la forme d'une liste de possibilités de traduction pour un terme donné. Nous proposons ici une méthode alternative basée sur le repérage, dans les deux textes, de structures sémantiques générales, ou isotopies, et la comparaison des schémas qu'elles présentent au niveau du texte et non plus de la phrase ou du paragraphe, permettant ainsi une plus grande tolérance dans le style de traduction à vérifier.

1. Introduction

Dans le domaine des outils d'aide à la traduction, dont le développement a fait oublier les échecs de la traduction purement automatique, de nouvelles problématiques et de nouvelles technologies ont vu le jour. Le projet IDOL [1] (*IRS-based Document Localisation*), financé par l'Union Européenne (fond INCO-DC 1002) et l'Office Fédéral suisse de l'Éducation et de la Science, propose, dans cette mouvance, une plate-forme d'aide à la traduction et à la localisation pour trois langues : anglais, français et arabe. La plate-forme comporte une mémoire de traduction et un module d'alignement pour alimenter celle-ci, un système multilingue de recherche d'informations, une base de données terminologiques et un module de vérification de traduction, TRACER. C'est ce dernier composant que nous présentons ici. Le principe général de TRACER est de multiplier les approches de textes bilingues, en s'appuyant sur des méthodes et des ressources différentes, pour ne pas limiter à un seul point de vue le jugement sur une traduction. Nous utilisons donc pour cette partie quatre sous-modules distincts, qui opèrent de façon indépendante mais dont les résultats peuvent être interprétés conjointement.

Parmi ces quatre modules, les trois premiers sont des utilisations traditionnelles de techniques classiques en analyse de corpus, et le quatrième se veut une innovation utilisant le

1. Le site WWW de ce projet est : `http://issco-www.unige.ch/projects/idol_public/`

concept d'isotopie, issu de la sémantique structurale et simplifié pour son utilisation dans le cadre de la linguistique informatique.

Comme le signale I. Melamed (Melamed, 1996), il est malheureusement très difficile d'obtenir des textes contenant des erreurs de traduction. Nos expériences porteront ici sur des modifications de textes par ailleurs bien traduits, comme l'inversion, l'omission ou la répétition d'un segment du texte cible, ou encore la conservation dans la cible d'un paragraphe non traduit du texte source. Les exemples sont obtenus à partir de la version multilingue (français, anglais et arabe) de la documentation des logiciels Microsoft Office.

2. Méthodes classiques en analyse de traduction

Nous présentons rapidement ici les approches traditionnelles dans le domaine de la vérification de traduction, et plus précisément celles qui sont utilisées dans le module TRACER.

2.1. Utilisation des algorithmes d'alignement

Les outils d'alignement de textes bilingues, initialement vus comme un moyen d'acquérir de façon non supervisée des lexiques bilingues à partir de corpus, se sont rapidement intégrés aux mémoires de traduction, et est devenu un pré-traitement quasi systématique de tout traitement de corpus multilingues.

Le principe de l'alignement est de mettre en correspondance, entre les deux textes, les sous-parties (paragraphes, phrases, ou autres) qui sont une traduction l'une de l'autre. Cette méthode suppose donc un pré-découpage préalable des textes, et l'utilisation de certains critères de choix. Les meilleurs résultats semblent être atteints en ne prenant en compte que la taille des sous-parties, que ce soit en nombre de caractères ou en nombre de mots : le principe sous-jacent est qu'un long segment se traduira par un segment proportionnellement aussi long. Dans le calcul de ces associations, différents schémas sont envisagés, comme par exemple la correspondance bi-univoque (1 unité du texte-source correspond à 1 unité du texte-cible), l'extension (1 pour 2), la compression (2 pour 1) ou la suppression (1 pour 0). L'algorithme proposé par Gale et Church (Gale & Church, 1993), qui est sans doute le plus utilisé, attribue des pénalités aux schémas d'association autres que la correspondance 1-1, et essaiera donc de conserver une identité structurelle linéaire entre les deux textes. À la sortie de cet algorithme d'alignement, outre les schémas d'association, sont donc indiqués les coûts des associations des différents segments. Le coût sera d'autant plus élevé que les tailles des éléments diffèrent, et que le schéma est "non-standard".

Dans une optique de vérification de traduction, ces dernières informations sont ainsi utilisables pour le repérage d'erreurs telles que l'omission ou la répétition de segments de texte, comme présenté dans (Melamed, 1996). Il est clair, toutefois, que ce genre de pratique ne s'applique que difficilement à des traductions non littérales, ou de tels cas de dilatation ou de compression ne sont pas à proprement parler considérés comme des erreurs. Même dans ce cas, comme présenté plus bas, le repérage de l'erreur est difficile à effectuer avec précision.

Le module d'alignement de la plate-forme IDOL reprend l'algorithme de Gale et Church (*op. cit.*) en y ajoutant la prise en compte des informations de mise en forme (niveaux de titre, format des caractères) lorsque celles-ci sont disponibles.

Les problèmes de cette approche directe sont les suivants. Tout d'abord, les techniques ci-dessus mentionnées ne donnent de résultats corrects que dans le cadre d'un texte bien structuré,

disposant d'une hiérarchie de paragraphes claire, et de parties courtes. L'alignement d'un texte plus linéaire, comme un roman, réduit considérablement la pertinence des résultats. Ensuite, ces techniques sont surtout fiables pour des couples de langues proches, comme l'anglais et le français. Nos expériences avec les couples français/arabe et anglais/arabe nous semblent moins satisfaisantes, même sur des textes structurés. Enfin, le principe de programmation dynamique tend à rendre très délicate l'interprétation des distances d'alignement dans le repérage d'éventuelles erreurs de traduction (voir plus bas). Les points d'ancrage absolus que sont le début et la fin des textes à aligner font qu'une perturbation très localisée aura comme conséquence l'amplification de la zone de distances élevées, l'algorithme d'alignement «corrigeant» l'erreur par paliers.

2.2. Utilisation de correspondances terminologiques

Une approche simple dans la vérification de traduction est bien entendu l'utilisation de lexiques bilingues afin de vérifier la bonne traduction de certains termes, ou bien la cohérence entre plusieurs textes, comme présenté dans (Macklovitch, 1996).

Cette méthode requiert donc un lexique multilingue, un alignement des textes au niveau de la phrase ou du paragraphe, et un analyseur capable au minimum d'une recherche d'occurrences de termes. Ainsi, pour chaque couple de phrases ou de paragraphes, la présence d'un terme du lexique dans le texte source est systématiquement évaluée en fonction de l'absence ou de la présence des termes correspondants dans le texte-cible. Le résultat est donc, pour chaque unité du texte cible, la liste des termes sources qui n'ont pas été correctement traduits. Il est de même possible d'effectuer le calcul inverse, en se basant sur le texte cible afin de repérer d'éventuelles répétitions.

Les problèmes pratiques de cette méthode sont bien entendu les reprises anaphoriques, et les pronominalisations, qui peuvent exister dans un texte et non dans sa traduction, ou vice-versa, et qui ajoutent du "bruit" dans la vérification de ces correspondances. Une solution, comme on le verra, est de prendre du recul dans ce type d'approche, et de ne pas, dans un premier temps du moins, effectuer une analyse à un niveau aussi précis que la phrase. De plus, cette méthode dépend fortement de la qualité de l'alignement parallèle établi automatiquement entre les deux textes.

Pour illustrer ces limites, on peut voir dans la figure 1, deux courbes correspondant à l'évolution dans le bi-texte de deux valeurs. La première (*alignment score*) correspond à la distance entre les deux unités de texte alignées. Un score élevé correspond généralement à l'appariement d'unités de tailles très différentes, ou à des associations non-bi-univoques. La seconde valeur (*terminology*) est une mesure des correspondances terminologiques sur la base de cet alignement[2]. Les textes ayant servi à calculer ces valeurs contenait une erreur, en ce sens que dans le texte français un paragraphe est répété (traduction anglais vers français). Ce paragraphe excédentaire a une taille de 850 caractères, et pourtant la zone indiquant l'erreur, aisément repérable dans ce graphique, occupe près de 4000 caractères. Elle correspond à un faible (voire nul) taux de correspondance terminologique, accompagné d'une distance élevée de l'algorithme d'alignement ; de plus, elle est située plus à droite que le lieu véritable de l'erreur (les deux parties identiques sont indiquées par les flèches). Ceci est dû aux "remous" créés par l'algorithme d'alignement. Notons qu'il y a toutefois cohérence entre le score d'alignement et le taux de correspondance terminologique, et qu'il est possible de localiser l'erreur comme étant la partie

2. Il s'agit d'un simple coefficient de Dice entre les unités de textes alignées, sur la base d'une liste de couples bilingues de termes, comme définie dans (Salton, 1989).

initiale de la zone perturbée, mais sans garantie aucune que le reste de cette zone soit exempt d'autres erreurs.

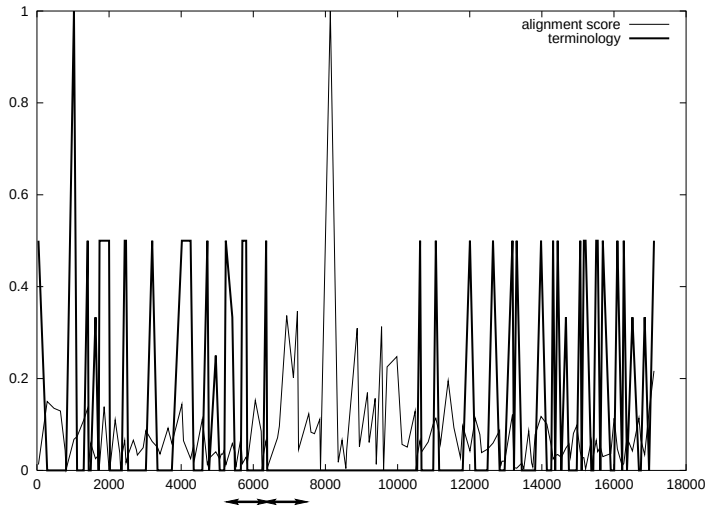

FIG. 1 – *Distance d'alignement et correspondance terminologique dans une traduction erronée (répétition) – Les flèches indiquent les paragraphes répétés dans le texte-cible*

Nous présentons ensuite une troisième méthode classique, plus centrée que les précédentes sur l'analyse monolingue.

2.3. Comparaison des caractéristiques générales des textes

Par *caractéristiques de surface*, nous entendons un ensemble de mesures statistiques réalisées indépendamment pour le texte original et sa traduction. Ces données peuvent être de pertinence et de précision diverses et provenir de différentes méthodes d'analyse de corpus. Nous recensons ici celles qui sont utilisées dans le module d'analyse de TRACER.

- Nombre de paragraphes, de phrases, de signes de ponctuation, répartition des caractéristique de mise en forme.
- Taille des paragraphes et des phrases.
- Fréquence et richesse du vocabulaire (rapport type / occurrence, indices stylométriques). Nous reprenons ici les principes généraux en stylométrie, présentés dans (Lebart & Salem, 1994)
- Unités lexicales les plus fréquentes (types et occurrences, avec et sans lemmatisation).
- Fréquence et répartition des catégories grammaticales principales.

Ces différentes valeurs sont obtenues soit par un parcours des textes, en s'appuyant sur des critères simples de découpage des unités, soit, pour les dernières, par analyse morpho-syntaxique.

Les couples de valeurs (correspondant au texte et à sa traduction) sont ensuite comparés, en tenant compte des deux langues utilisées, et éventuellement de normes de traduction telles qu'elles peuvent être définies dans le cadre d'un projet.

Cette approche possède l'avantage de considérer le texte dans son ensemble, sans se concentrer, comme dans les deux précédentes, sur des zones locales dont la mise en correspondance peut être erronée, ou trop stricte.

Elle ne donne cependant pas d'indication quant à la localisation des éventuelles erreurs repérées, dont la nature est généralement l'omission de parties de textes (nombre d'unités trop disparates), ou le non-respect de normes de traduction (emploi abusif de pronoms, etc.) Enfin, ces caractéristiques permettent d'évaluer rapidement l'homogénéité d'un travail de traduction réalisé par plusieurs acteurs.

3. Utilisation des classes sémantiques

3.1. Principe de la méthode

Le principe général de notre approche est d'étendre la vérification des correspondances terminologiques, et d'aboutir à une caractérisation au niveau du texte et non plus de zones locales, arbitraires ou non.

Alors que les correspondances terminologiques simples comparent la présence de couples de termes, nous envisagerons ici les corrélations de *classes* de termes. Ces classes seront définies de façon à permettre une caractérisation des principaux thèmes abordés dans le texte traduit, et peuvent donc, en leur sein, contenir des entités hétérogènes sur le plan grammatical ou morphologique, la relation ici étant une forme d'équivalence sémantique.

Comme pour la méthode basée sur les correspondances terme-à-terme, à chaque classe définie dans la langue source, on fait correspondre un classe dans la langue cible, en se basant sur la définition en intension de la classe, et non en extension.

Une fois les classes définies, et ce pour chaque langue, on opère un simple repérage des positions des éléments de chaque classe dans le texte. La notion de position ici est simplement le nombre de caractères séparant le début d'une occurrence du début du texte. Le résultat de cette recherche est donc, pour chaque classe et pour chaque langue, une liste de coordonnées, représentant toutes les positions des éléments de la classe, sans tenir compte plus avant de l'identité des mots.

À partir de ces classes, il est assez aisé d'obtenir une représentation graphique permettant de repérer les différentes «zones thématiques» des textes, et par là-même, les lieux de transition entre les thèmes abordés. Ce seront donc ces derniers critères qui seront utilisés pour estimer les erreurs de traduction, en conservant pour celles-ci un point de vue global sur le texte.

Un exemple minimal de ces représentations graphiques est présenté dans la figure 2.

3.2. Définition des classes sémantiques

Le concept d'*isotopie*, en sémantique structurale, fut introduit par A.-J. Greimas (Greimas, 1986), puis repris par F. Rastier (Rastier, 1987) qui en fait un concept clé de la Sémantique Interprétative. Une isotopie est l'effet de la récurrence d'une unité sémantique le long d'un texte. Cette unité sémantique, un *sème* en sémantique structurale, traduit l'appartenance de plusieurs signifiés à une même zone sémantique, cette zone pouvant être caractérisée par ce même sème ou marqueur sémantique.

Ainsi, la définition d'une isotopie commence par celle d'une classe sémantique, en lui attri-

buant un marqueur unique et un ensemble d'éléments. Une même unité lexicale peut fort bien appartenir à plusieurs classes, que ceci soit dû à des effets de polysémie ou à une intersection sémantique non vide des classes elles-mêmes. Par la suite, le traitement se fait isotopie par isotopie, et non par unité lexicale.

Pour notre exemple lié au logiciel de traitement de texte, nous avons défini un certain nombre de classes, présentées ici. Ces classes sont donc identifiées par leur sème (en gras), et contiennent un certain nombre d'unités lexicales appartenant au champ sémantique.

"Format" (Français) : page, champ, police, marge, en-tête, pieds de page, interligne, gras, italique...

"Document" (Français) : symbole, texte, ligne, paragraphe, caractère, phrase, document, lettre, mémo, rapport, graphique...

"Matériel" (Français) : écran, clavier, souris, ordinateur, touche...

"Formatting" (anglais) : edit, format, page, field, case, typing, heading, margin, bold, italic, font...

"Document" (anglais) : symbol, text, typos, document, letter, memo, character, text, page, paragraph...

"Hardware" (anglais) : screen, keyboard, mouse, key, computer...

Ces classes ont été établies sur la base d'une connaissance générale, de la part de l'utilisateur, des principales notions des textes à analyser. La nature et le contenu de ces classes est bien entendu un critère important de la qualité du résultat de notre approche. Toutefois, le travail exigé pour leur établissement est parfaitement justifié dans le cas d'un projet de traduction dont la spécificité du domaine permet une réutilisation de ces ressources. Dans le cadre de la station IDOL, nous faisons appel entre autres, pour l'établissement de ces données à un thésaurus utilisé par ailleurs pour la recherche d'informations, et qui met donc en place une telle classification terminologique minimale. D'autres sources, notamment les réseaux sémantiques comme Wordnet peuvent également fournir de telles classes.

Il convient cependant de bien séparer les classes d'équivalence définies au sein d'un thésaurus de la notion de classe sémantique telle que nous la définissons ici. Une isotopie peut (et doit) être basée sur des notions plus larges que la synonymie ou la para-synonymie. Les classes d'équivalence sémantique peuvent traverser les frontières des catégories grammaticales, et même les champs sémantiques canoniques. Il existe enfin des approches semi-automatiques de traitement de corpus permettant d'établir de telles relations d'équivalence sémantique, comme par exemple (Assadi, 1996) et (Tanguy, 1997).

3.3. Construction des isotopies

À partir de ces classes, il est opéré une projection syntagmatique par l'algorithme suivant :

- Pour chaque classe, et pour chaque mot ou expression de cette classe, on note la position de ses occurrences dans le texte (en nombre absolu de caractères depuis le début du texte).
- À partir de cet ensemble ordonné de points, on opère un passage à une représentation continue par lissage. Un seuil doit être fixé, correspondant à l'intervalle maximal (en nombre de caractères) séparant deux occurrences d'une même classe sémantique. On repère ainsi les zones de texte sémantiquement riches pour une classe donnée, en éliminant

les occurrences isolées. Le résultat peut être représenté par un graphique en créneaux, figure 2, dans lequel les créneaux proprement dits indiquent la présence de termes de la classe sémantique dans la zone correspondante du texte.

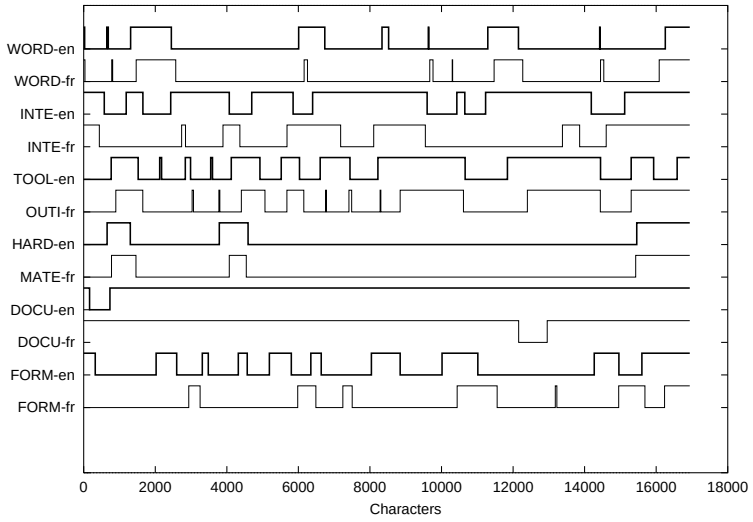

FIG. 2 – *Représentation continue des isotopies - texte sans erreur*

Pour ces résultats, le seuil est fixé à 500 caractères, pour des textes d'une taille d'environ 16000 (français) et 12000 (anglais) caractères. Les classes sont codifiées de la manière suivante (la langue de définition est le français (*fr*) ou l'anglais (*en*)) :

- WORD / WORD : logiciel Word.
- INTE / INTE : interaction avec le logiciel
- TOOL / OUTI : outils d'édition
- HARD / MATE : matériel informatique
- DOCU / DOCU : concepts généraux sur la notion de document
- FORM / FORM : concepts liés au format d'un document

– La même opération est répétée pour le second texte (traduction), en appliquant aux mesures obtenues un coefficient multiplicateur égal au rapport de compression/dilatation en nombre de caractères des deux textes. Ceci rapporte donc les représentations des isotopies dans les deux langues sur une même échelle.

3.4. Interprétation des résultats et introduction d'erreurs

Cet exemple montre bien que les "meilleures" classes sont, sinon les plus courtes dans leur définition, du moins celles dont les occurrences sont localisées. Une preuve en est la classe "Document", omniprésente dans l'ensemble du texte, et donc porteuse de peu d'information. Ces interprétations "naïves" se font aisément sur la base des graphiques en créneaux examinés par paires (anglais/français). Les isotopies les plus "parlantes" sont celles dont la forme est la plus irrégulière ("Word/*Word*" et "Outils/*Tools*"). Les variations entre leurs zones de présence et d'absence se retrouvent dans les deux textes, et permettent une certaine garantie sur la correcte traduction du texte.

Le graphique de la figure 3 est obtenu par remplacement d'une courte partie du document en français (cible) par l'original correspondant en anglais, simulant ainsi le cas de la non-traduction d'une partie du texte-source.

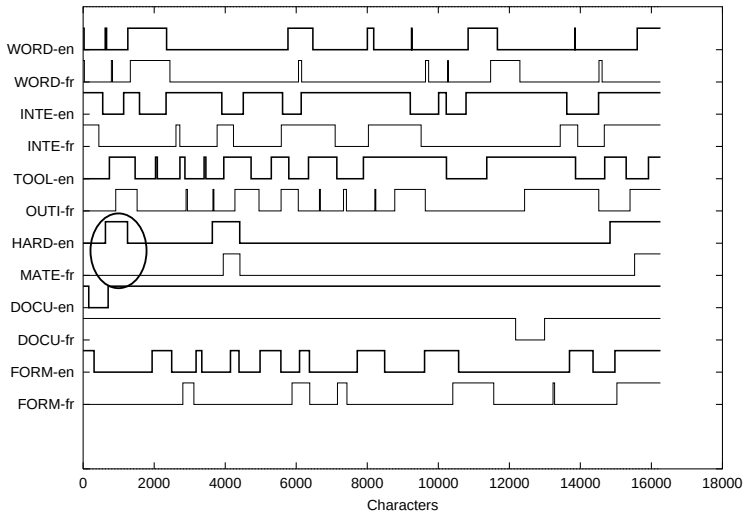

FIG. 3 – *Isotopies de traduction erronée*

Certaines isotopies ne sont pas trop perturbées ("Word" entre autres), tandis que d'autres perdent quelques-unes de leurs caractéristiques principales ("Matériel" perd une de ses rares zones d'activité). Ce dernier couple d'isotopies est en effet le plus atteint par la modification introduite dans le texte, puisqu'un des «pics» de l'isotopie *Matériel* du texte français a disparu au début du texte, comme indiqué par le cercle dans la figure. Comme on le voit, cette méthode permet donc de repérer, sinon précisément la zone où se trouve l'erreur, du moins sa localisation générale et la nature (ici, perte d'information lié à un domaine sémantique particulier).

En comparaison, voici le résultat de la méthode du score d'alignement et des correspondances terminologiques pour les mêmes textes (figure 4). Dans ce cas, l'identification et le repérage de l'erreur sont bien moins aisés sur la base de ces informations.

Le problème principal ici est bien entendu la lecture des graphiques de ces isotopies. La présence d'erreurs conduit souvent à un décalage des zones correspondantes, dont l'interprétation automatique pose des problèmes (par exemple, à l'aide d'une métrique de corrélation entre les courbes), puisque le taux de compression/dilatation entre les deux textes est basé sur la longueur effective de ceux-ci en nombre de caractères. Une ligne directrice semble être de prendre en considération le meilleur couple d'isotopies, et de comparer sa structure aux autres schémas pour un même texte (notamment les transitions d'une isotopie à l'autre). La lecture doit donc se faire à un niveau aussi proche que possible du texte pris dans son ensemble, vu le manque de correspondance directe entre les éléments des classes entre les langues.

3.5. Utilisation

Ce genre de méthode ne se veut pas autonome, mais propose un nouveau point de vue sur une traduction, qui doit être confronté à des résultats provenant de techniques et de ressources différentes. La spécificité de ces corrélations d'isotopies réside sans doute dans la souplesse de la notion elle-même : en définissant un niveau d'abstraction par rapport à des correspondances

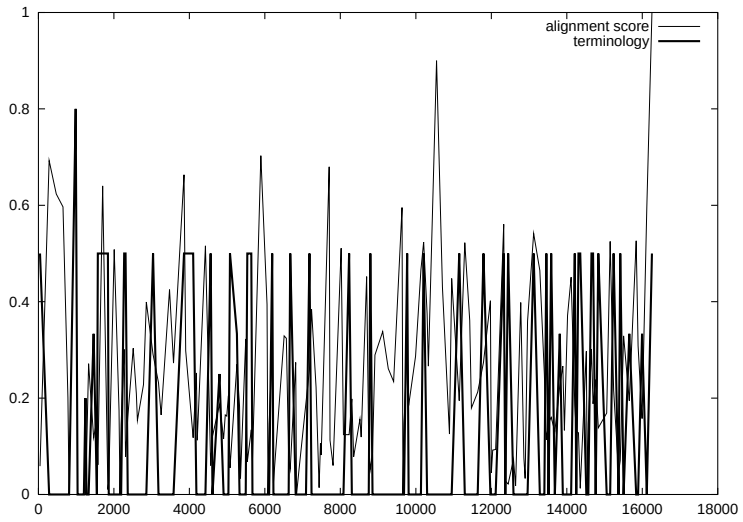

FIG. 4 – *Alignement et terminologie, mêmes textes*

unitaires entre mots ou expressions, elle permet une plus grande tolérance dans le jugement d'une traduction, et peut donner un avis positif sur une traduction moins canonique.

Bien entendu, la plus grande responsabilité des résultats de cette méthode repose sur la sélection et la définition des classes sémantiques. Dans notre exemple, les bons résultats sont en grande partie dus à l'étroitesse thématique des textes traités, et donc à la grande facilité d'établir des classes pertinentes. Le problème est tout autre lorsque l'outil se doit de traiter n'importe quel texte, quel qu'en soit le sujet et la forme. Sans tomber dans cet extrême où de toutes façons bien peu de méthodes sont accessibles (si ce ne sont les méthodes classiques d'alignement basées sur la longueur des unités), dans le cas d'un corpus homogène, et surtout disposant d'une base de données lexicales ou d'un thésaurus, l'établissement des classes devient envisageable.

Enfin, cette méthode peu exigeante en terme de traitement linguistique des textes à analyser s'applique particulièrement bien aux couples de langues plus disparates que l'anglais et le français. Le projet IDOL s'attachant au traitement de la langue arabe, pour laquelle les méthodes d'analyse traditionnelle (morphologique et syntaxique) sont moins développées, et les algorithmes d'alignement bilingue moins fiables, ce genre d'approche propose un contournement des problèmes spécifiques de cette langue (cf. (Al-Slahlabi & Evens, 1998)). Les résultats obtenus sur les textes correspondants en langue arabe donnent des résultats en tous points comparables, alors que les méthodes d'alignement (et donc de vérification des correspondances terminologiques) sont bien moins fiables.

3.6. Améliorations envisageables

Ce travail, bien qu'il ait abouti à la réalisation d'un prototype opérationnel, reste cependant une simple ébauche dans l'utilisation de données sémantiques dans la vérification de traduction. D'un point de vue pratique, il serait souhaitable d'approfondir le mode d'interaction du relecteur avec la représentation graphique, notamment en donnant la possibilité à l'utilisateur de «zoomer» sur une zone précise d'un texte, et d'avoir accès pour celle-ci au détail des occurrences de termes dans cette zone.

Dans un même esprit de rapprochement entre les méthodes de vérification, il est possible

d'obtenir un schéma d'alignement des textes en se basant sur les schémas d'isotopies. En ce sens, nous nous rapprochons des travaux de P. Van Der Eijk (Van der Eijk, 1999) sur la notion de cohérence terminologique au sein d'un texte. La différence de notre approche réside dans l'identification des spécificités sémantiques des zones mises en correspondance[3]. Une telle approche permettrait également un traitement plus évolué dans la mise à l'échelle des schémas d'isotopies, à la place de la réduction linéaire présentée ici.

4. Conclusion

Nous avons présenté ici une nouvelle approche dans la vérification de traduction, qui permet une plus grande souplesse. En s'attachant à réduire la trop grande restriction des approches habituelles à des zones de textes réduites le plus souvent à la phrase, ce qui pourrait être initialement pris comme une volonté de réduire la quantité d'information disponible est en fait un atout. Nous avons vu que certaines erreurs ne pouvaient être repérées en se basant sur des méthodes directes d'alignement parallèle et de correspondance terme-à-terme. Toutefois, une analyse conjointe des différentes approches proposées semble la plus prometteuse.

De même, le passage d'une représentation discrète (ici les vecteurs de coordonnées des occurrences) à une représentation continue (les graphiques) permet un meilleur jugement synthétique, qui peut toujours être affiné par un retour aux valeurs précises.

Références

AL-SLAHLABI R. & EVENS M. (1998). A computational morphology system for arabic. *COLING-ACL 98*.

ASSADI H. (1996). Interactive semantic analysis for building conceptual models from corpora. In *Corpus-oriented semantic analysis*. Proceedings of the ECAI-96 Workshop.

GALE W. & CHURCH K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**(1), 75–102.

GREIMAS A. (1986). *Sémantique structurale : recherche de méthode*. PUF.

LEBART L. & SALEM A. (1994). *Statistique Textuelle*. Dunod.

MACKLOVITCH E. (1996). Peut-on vérifier automatiquement la cohérence terminologique ? *META*, **41**(3).

MELAMED I. D. (1996). Automatic detection of omissions in translations. *COLING*, **2**, 764.

RASTIER F. (1987). *Sémantique interprétative*. PUF.

SALTON G. (1989). *Automatic Text Processing*. Addison Wesley.

TANGUY L. (1997). Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative. Thèse de Doctorat, Université de Rennes 1.

VAN DER EIJK P. (1999). Comparative discourse analysis of parallel texts. In *Natural Language Processing using very large corpora*. Armstrong et. al. Eds., Kluwer (à paraître, 1999).

3. P. Van Der Eijk utilise un calcul de corrélation entre les parties successives d'un texte en fonction de la coprésence de termes ou de séquences de caractères. Il repère ainsi les principales transitions thématiques d'un texte, et propose un découpage puis un alignement des textes sur la base de cette information.