Génération de corpus multilingues dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère

Guy Deville[1], Laurence Dumortier[1], Hans Paulussen[2]

[1]Facultés Universitaires N.D. de la Paix – 5000 Namur – Belgique
Guy.Deville@fundp.ac.be, Laurence.Dumortier@fundp.ac.be

[2]K.U. Leuven – Campus Kortrijk (KULAK) – 8500 Kortrijk – Belgique
Hans.Paulussen@kulak.ac.be

Abstract

This paper presents a method for the automatic generation of aligned bilingual corpora in a Web-based reading tool for Dutch texts by French speaking learners (NEDERLEX). The authors first discuss the major functions of NEDERLEX. Then they describe the role of bilingual corpora in the design and construction of the NEDERLEX tool, as well as the approach adopted for the extraction and alignment of such corpora. A demo of the NEDERLEX prototype will be presented during the conference talk.

Résumé

Cet article expose une méthode de génération automatique de corpus bilingues alignés dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes néerlandais à l'usage des apprenants francophones (NEDERLEX). Les auteurs présentent d'abord les principales fonctionnalités de l'outil NEDERLEX. Ils décrivent ensuite le rôle spécifique des corpus bilingues dans la construction et l'utilisation d'un tel outil, ainsi que la méthode d'acquisition et d'alignement de ces corpus. Une démonstration de la version prototype de NEDERLEX est prévue lors de l'exposé oral durant la conférence.

Mots-clés : alignement de corpus multilingues, logiciel en ligne d'apprentissage des langues étrangères.

1. Introduction

D'expérience, on observe que l'opacité du vocabulaire néerlandais et sa structure morpho-lexicale très éloignée du français constituent une des principales pierres d'achoppement dans la maîtrise de cette langue étrangère par les francophones. L'apprentissage du vocabulaire constitue dès lors une dimension importante des méthodes d'enseignement du néerlandais en tant que seconde langue, qui sont proposées sur le marché en général, et dans l'enseignement supérieur en particulier.

Les manuels traditionnels d'enseignement des langues étrangères sont principalement constitués de textes accompagnés de leur vocabulaire. Ce vocabulaire résulte d'un choix de l'auteur, et est souvent présenté en listes bilingues de mots isolés ou repris dans un contexte minimal. Les limites d'une telle présentation statique sont évidentes : (i) pour déchiffrer un texte, l'apprenant doit constamment passer « physiquement » du texte support à la liste de vocabulaire et vice-versa, (ii) le vocabulaire proposé ne couvre qu'un sous-ensemble des mots du texte ; (iii) la liste ne peut varier selon le niveau de connaissance de l'apprenant.

Les versions logicielles des cours de langues (en ligne sur la Toile ou sous forme de CD-ROM) permettent d'intégrer un dictionnaire traductif sous format électronique, consultable à la demande de l'étudiant.

Cette approche supprime les inconvénients liés au support écrit : elle offre l'accès à un glossaire pour une très large couverture lexicale des textes, et la fréquence de consultation du dictionnaire est en fonction du niveau de l'étudiant. Toutefois, le dictionnaire électronique est un outil non intégré : il est toujours consulté « hors contexte » car le choix de la traduction d'un mot dans son contexte original est laissé à l'initiative de l'apprenant, qui doit sélectionner dans le dictionnaire les informations appropriées (catégorie grammaticale, sens, exemples). Une telle lecture contextualisée de la langue est une démarche malaisée qui n'est pas à la portée immédiate de la plupart des apprenants.

Ce constat a amené les auteurs à concevoir NEDERLEX, qui est un outil original de création de supports en ligne d'aide à la lecture de textes néerlandais (Deville et Dumortier, 2003). Dans sa conception, NEDERLEX fait appel à de grands corpus bilingues alignés pour illustrer en contexte et sans ambiguïté le vocabulaire de tels textes.

Cet article est centré sur les aspects de méthodologie et d'ingénierie linguistique qui sous-tendent l'élaboration de tels corpus bilingues. Les auteurs présentent d'abord les principales fonctionnalitésropriéoutil NEDERLte. Ils décrivent ensuite le rôle spécifique des corpus bilingues dans la construction et l'utilisation d'un tel outil, ainsi que la méthode d'acquisition, de génération et d'alignement de ces corpus. Ils concluent en exposant les principaux bénéfices de la génération et de l'utilisation de corpus multilingues dans NEDERLEX tant du point de vue des concepteurs de l'outil (enseignants), que de leurs utilisateurs finaux (apprenants).

2. Fonctionnalitésropriéoutil NEDERLEX

NEDERLEX est conçu comme un outil générique interactif (à l'usage des enseignants) pour l'édition d'un cours de textes néerlandais multimédia de tous niveaux. Cet outil génère un produit fini (à l'usage des apprenants) sous la forme d'un site Web qui offre un ensemble de documents écrits authentiquesroprtous types et de niveaux de difficulté différents, qui sont entièrement glosés (avec explication du vocabulaire), et assortis d'exercices interactifs. Pour chacun des textes, l'outil reprend la traduction en contexte de chaque mot, qui fait l'objet d'une série d'illustrations sous la forme de citations bilingues (concordances). Ces illustrations sont extraites d'un grand volumeroprtextes néerlandais-français alignés, appelés ici « corpus bilingues alignés ». D'un point de vue technique, le site Web est généré par des pages dynamiquesrconstruitesren PHP via une interaction avec une baseroprdonnéesrMySql reprenant sous formeroprtablerles ressources linguistiquesrassociéesrà chaque cours (textes des leçons, lexique, homonymes, concordances et corpus bilingues alignés).

NEDERLEX serprésente sous la forme d'une interface reprenant les fonctionnalitésrnécessaires à la création d'un cours multimédia, à savoir :

(i) créerret éditerrune leçon à partir d'un fichierrtexte. Cette leçon est ensuite balisée : pour chaque motropria leçon, on déterminerоprmanière automatique sa cléropria table lexique (associant des informations telles que lemmes et formes fléchies néerlandaise et française, catégorie syntaxique, genre, etc.) et on associerà ce motrune fonction javascript qui permet d'afficher, au clicroprsouris, toutesrles concordances trouvées dans les corpus alignés (stockées dans ia table des concordances) qui correspondentrà ce lemme. Une fonction sous forme de menu déroulant permet à l'utilisateur de lever les ambiguïtésrqui perturbentria traduction de certains mots, en raison de phénomènes de polysémie et d'homonymie ; Ce menu déroulant présente à cet effet les informations issuesropria table des homonymes.

(ii) importer, mettre à jour et contrôler le lexique associé au texte d'une leçon (le lexique du prototype compte actuellement environ 4.000 entrées validées et exploitées dans le cadre de cours existants, sur un total d'environ 16.000) ;

(iii) importer et mettre à jour la base de corpus bilingues alignés qui produiront les extraits illustrant chaque mot des leçons du cours et sa traduction en contexte (concordances). L'élaboration de ces corpus (extraction et alignement par phrase) est réalisée de manière semi-automatique. Le jeu de corpus alignés a un volume total de 1.500.000 mots ;

(iv) générer l'ensemble des concordances, c'est-à-dire les extraits bilingues des corpus alignés dans lesquels les mots du lexique et leur traduction apparaissent ; A cet effet, chaque entrée de la table des concordances contient la clé du lemme, ses formes fléchies néerlandaise et française apparaissant dans le corpus aligné, ainsi que les numéros de référence du corpus et du paragraphe concernés.

Pour gérer toutes ces fonctionnalités, nous disposons de cinq tables reprises dans la base de donnée décrite ci-dessus : la table des textes, la table du lexique, la table des homonymes, la table des corpus bilingues alignés et la table des concordances.

La méthodologie retenue favorise un développement et une mise à jour modulaires — et en grande partie automatisés — des textes de leçons, des lexiques associés et du corpus de textes bilingues alignés, selon le schéma de la figure 1.

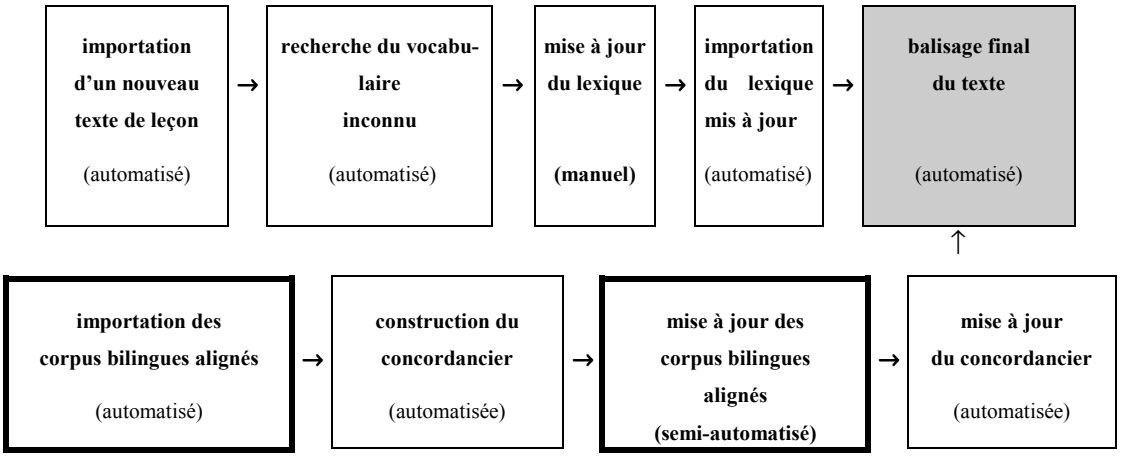

Figure 1. Schéma de développement d'un cours avec l'outil NEDERLEX

L'outil NEDERLEX génère donc un texte de lecture qui a été préalablement balisé, ce qui permet à l'apprenant d'« interroger » chaque mot d'un clic de souris. Lorsque l'apprenant clique sur un mot du texte pour en solliciter la traduction, un tableau apparaît dans une fenêtre en bas de l'écran, avec les informations suivantes : (i) la ligne supérieure affiche les informations grammaticales du mot : forme lemmatisée, catégorie et sous-catégorie syntaxique et formes fléchies pertinentes d'un point de vue didactique (pluriel des noms, forme comparative et superlative des adjectifs, formes prétérit et participe passé des verbes) ; (ii) les cellules inférieures du tableau reprennent sous leur différentes formes fléchies, plusieurs occurrences du mot néerlandais (cellules gauches) avec leurs différentes traductions françaises (cellules droites) dans des contextes ou « concordances » issus de corpus bilingues alignés.

Par un clic sur le mot néerlandais dans son contexte (cellule de gauche), on ouvre une nouvelle fenêtre qui affiche le contexte complet du mot retenu (phrase en néerlandais avec sa traduction en français), avec l'indication de la référence de cette source. Ces fenêtres sont de

couleurs différentes en fonction du type de corpus d'où proviennent les concordances (voir section 3.). Les entités balisées peuvent être des mots simples ou composés ainsi que des unités lexicales constituées de plusieurs mots (formes disjointes de verbes séparables, locutions, syntagmes verbaux ou nominaux, etc.).

On notera que les exemples de traduction de chaque mot sont toujours donnés en fonction du contexte de ce mot dans la leçon, les ambiguïtés ayant été levées lors du balisage préalable du texte ; il y a donc absence de polysémie et d'homonymie. Ce choix désambiguïsé et contextualisé constitue une aide précieuse pour l'apprenant, qui est guidé dans sa recherche de la traduction exacte des mots inconnus. La figure 2. reprend un exemple de texte de lecture développé avec l'outil NEDERLEX.

La structure optimisée des tables permet un affichage rapide des informations lexicales sollicitées par l'utilisateur. En effet, chaque clic de souris ne requiert l'accès qu'à une seule table : soit la table des concordances (en cas de clic sur un mot de la leçon pour faire apparaître les concordances), soit la table des corpus bilingues alignés (en cas de clic sur un mot du tableau des concordances pour faire apparaître le contexte complet du mot retenu).

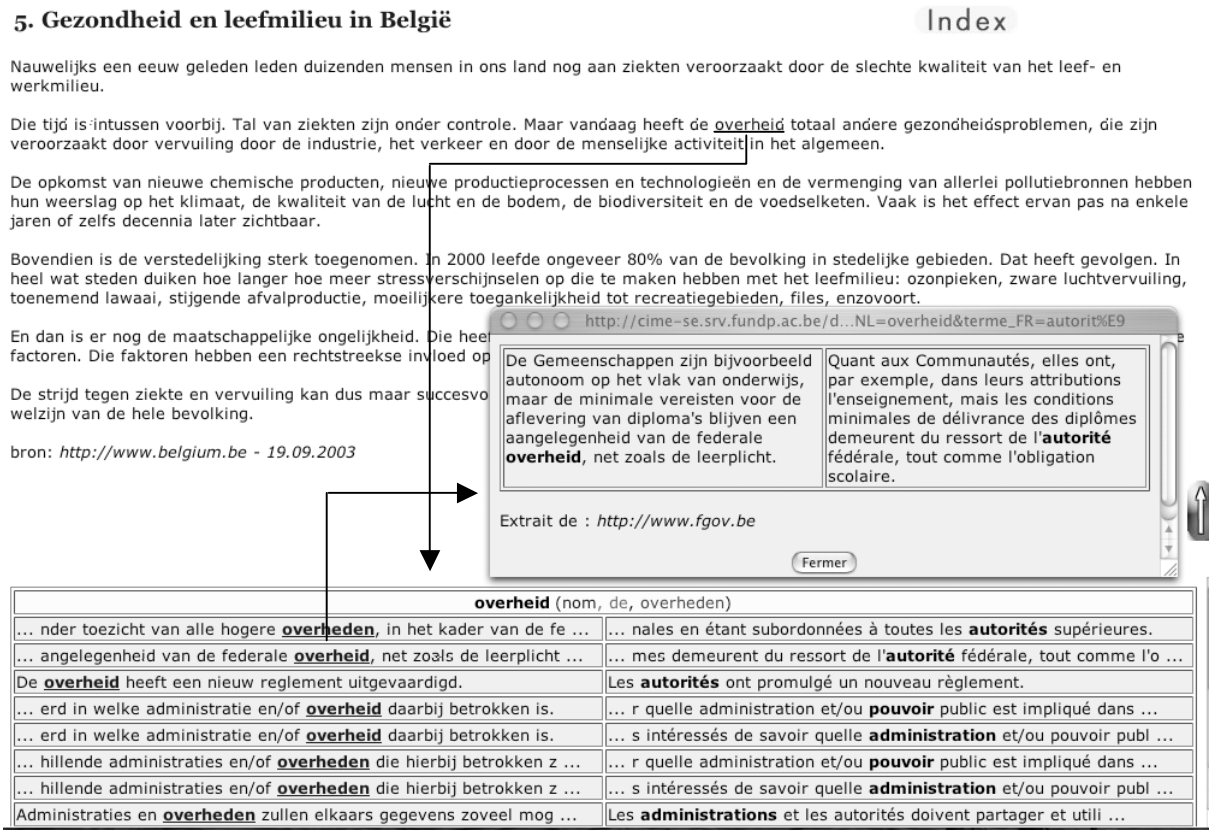

Figure 2. Extrait d'un cours développé avec l'outil NEDERLEX

3. Génération semi-automatique de corpus bilingues alignés

Les corpus bilingues néerlandais-français alignés constituent donc une des ressources linguistiques principales de NEDERLEX. Ces corpus sont de trois types : (i) les supports écrits de cours de néerlandais que nous avons édités jusqu'à présent comportent de précieux glossaires éprouvés sur le plan didactique, qui reprennent plusieurs milliers de phrases avec leur traduc-

tion (néerlandais-français) ; nous avons reformaté ces phrases sous forme de corpus aligné ; (ii) plusieurs sites Web présentent aujourd'hui de nombreux textes de grande qualité dans plusieurs versions linguistiques (dont le français et le néerlandais). Il s'agit de sites fédéraux officiels (gouvernement belge, ministères, cours et tribunaux), de sites d'organismes internationaux (par ex. l'Union européenne) ou encore de sites commerciaux (par ex. dans le secteur de la distribution alimentaire) ; (iii) enfin plusieurs sites présentent des textes de nature technique (par ex. des textes de loi, jugements et jurisprudence) qui nous servent à illustrer un cours de terminologie juridique néerlandaise (Deville et Dumortier, 2002).

Les corpus de type (i) sont constitués manuellement, car ils possèdent d'emblée une structure proche du format souhaité (alignement au niveau de la phrase), et constituent un corpus fini, stable et homogène. Dans une première phase, les corpus de type (ii) et (iii) ont été constitués manuellement en identifiant des versions néerlandaise et française de textes pertinents tels que décrits plus haut, qui sont copiées dans un tableur et alignées ensuite au niveau du paragraphe. Cette procédure fastidieuse a un double inconvénient : la mise à jour et l'enrichissement systématique des corpus est malaisée et les corpus ne sont alignés qu'au niveau du paragraphe. Dans le cas de longs paragraphes, cette dernière contrainte rend la lecture du mot-clé et de sa traduction peu lisible, et il arrive que ce mot-clé soit restitué dans une traduction erronée, pour des raisons techniques liées à l'algorithme de construction des concordances. Pour ces motifs, nous avons décidé d'automatiser (i) l'élaboration de corpus à partir de sites Web multilingues tels que décrits plus haut, et (ii) l'alignement de ces corpus au niveau de la phrase au lieu du paragraphe. Détaillons à présent cette démarche.

Le *Corpus Namur* développé dans le cadre d'une thèse de doctorat (Paulussen, 1999) constitue un exemple de corpus multilingues qui a été créé de manière semi-automatique[1]. Il contient des textes en français, anglais et néerlandais d'un volume total de 2.000.000 mots. Une moitié du corpus contient des textes littéraires (fiction), l'autre moitié contient des textes autres que de fiction : les débats du Parlement européen d'une part, et un volume du Courier de l'Unesco de l'autre. L'utilisation d'outils développés en Awk et Perl, nous a permis de nettoyer et d'aligner au niveau du paragraphe les textes multilingues du *Corpus Namur* en provenance de différentes plates-formes informatiques.

Au début des années 90, la standardisation de l'encodage de tels textes était à un stade embryonnaire, et ceux-ci ne comportaient aucune forme d'annotation (tels les balises HTML), de sorte qu'un texte était simplement défini comme un bloc de lignes (d'une longueur d'environ 60 caractères), chaque ligne se terminant par une fin de ligne, et un bloc se terminant par deux lignes. Grâce à une procédure semi-automatique d'alignement, une sélection de phrases échantillons a été automatiquement convertie en une base de données trilingues (TRIPTIC), qui constituait la plate-forme de travail pour une étude en linguistique contrastive que nous ne détaillerons pas ici.

Cette approche réalisée dans un environnement Unix/Linux s'est avérée très puissante, mais limitée à des utilisateurs quelque peu spécialisés de la programmation. Une nouvelle approche consiste à présent à raffiner et intégrer dans un environnement convivial les outils que nous avons préalablement créés, afin de permettre à tout utilisateur d'exécuter les différentes étapes de sélection et d'alignement de corpus multilingues par un simple clic de souris. Cette approche est réalisable aujourd'hui grâce au développement de l'internet et à la standardisation des formats de textes.

[1] Voir également : http://www.fundp.ac.be/~hpaulus/NamurCorpus.html

Depuis l'avènement et la diffusion des navigateurs (browsers) HTML dans les années 90, tout internaute peut aujourd'hui récupérer n'importe quel fichier du monde entier en provenance de tout type de plate-forme informatique, sans devoir maîtriser les commandes énigmatiques d'outils spécialisés de programmation (tels que *ftp*, par exemple). La page Web ou le fichier à récupérer se trouve à la portée d'un simple clic de souris. Chaque clic active une connexion selon le protocole HTTP, dont les détails se déroulent dans les coulisses du navigateur. La consultation de pages HTML s'opère la plupart du temps selon ce mode, c'est-à-dire un simple clic sur un lien hypertexte dans une page HTML affichée dans la fenêtre d'un navigateur. On peut également consulter tout serveur Web (qui contient des documents HTML) à l'aide de n'importe quel langage de programmation ou langage script qui supporte le protocole HTTP. Cette consultation non interactive permet d'automatiser le processus de récupération régulière de fichiers HTML, ce qui peut s'avérer très utile lorsqu'on visite fréquemment des sites « périodiques », qui changent à un rythme quotidien, hebdomadaire ou mensuel.

Aussi simple que cette approche puisse paraître, il n'empêche que la démarche exige un travail de programmation défensive de la part des développeurs de tels outils, car la connexion internet n'est pas toujours fiable et la structure du site à consulter peut changer d'un jour à l'autre. Heureusement, la gestion des sites Web se stabilise de manière croissante, puisqu'une diffusion de qualité des données exige une structure plus rigide permettant une mise à jour efficace des informations disponibles. Ce constat est particulièrement vrai dans le cas de sites multilingues. Ainsi, la stabilisation de la structure des sites Web et les performances accrues des fonctionnalités HTTP dans les langages script nous ont permis d'envisager la récolte automatique de textes multilingues en ligne. Cette récolte est la première étape dans la construction d'un module de génération automatique de textes multilingues. Les étapes suivantes comportent une procédure qui divise les textes en unités de phrases et une procédure automatique d'alignement de ces phrases.

L'alignement automatique de textes au niveau de la phrase est une condition préalable à une exploitation efficace d'un corpus parallèle. Un tel alignement implique la mise en correspondance automatique des portions sélectionnées d'un texte source et des portions équivalentes dans un texte cible. Il existe aujourd'hui plusieurs outils d'alignement, mais la plupart d'entre eux requièrent un imposant travail de post-édition, que nous voulions réduire de manière significative. En outre, aucun outil d'alignement n'a été développé spécifiquement pour le néerlandais.

Les algorithmes d'alignement utilisent tantôt une approche statistique tantôt une approche lexicale, bien qu'un mélange des deux approches soit de plus en plus utilisé (Brown *et al.*, 1990 ; Gale et Church, 1991 ; Simard *et al.*, 1992 ; Church, 1993 ; McEnery *et al.*, 1997).

(Simard *et al.*, 1992) sont probablement les premiers à allier une approche statistique avec un support linguistique, en introduisant la notion de « cognates ». Leur approche est une amélioration du modèle probabiliste de (Gale et Church, 1991) qui est strictement basé sur la longueur de phrase. Une approche plus linguistique est utilisée dans le programme d'alignement développé par (Hofland, 1996), dans le cadre du projet ENPC (English-Norwegian Parallel Corpus). Dans cette étude, Hofland utilise des mots appelés « ancres », qui sont stockés dans un lexique bilingue contenant des mots qui sont soit (i) raisonnablement fréquents, ou (ii) qui ont des équivalents transparents dans chacune des deux langues utilisées.

Dans le cadre de cet article, nous présentons une série préliminaire de tests d'alignement de corpus bilingues selon une procédure qui comporte les étapes suivantes :

La première étape consiste à télécharger automatiquement un ensemble de textes parallèles (néerlandais-français) de bonne qualité à partir de sites web sélectionnés sur base de leur structure bilingue, comme indiqué plus haut. Notre choix s'est porté sur (i) des textes relatifs à l'alimentation repris sur le site d'une chaîne de grande distribution (www.delhaize.be), et sur (ii) la transcription des débats parlementaires européens accessibles sur le site du Parlement européen (www.europarl.eu.int/plenary). En outre, (iii) le texte du dernier accord gouvernemental a été repris manuellement du site du gouvernement fédéral belge (www.belgium.be) à partir d'un fichier pdf. Le volume de chaque échantillon de textes est repris au tableau ci-dessous.

Le nettoyage de ces fichiers HTML en simples fichiers textes a été réalisé à l'aide d'un navigateur texte (lynx) qui permet d'extraire automatiquement les balises d'un fichier HTML. Cette approche, qui exige un nettoyage supplémentaire du texte à l'aide de filtres (scripts) écrits en langages Awk et Perl, est perfectible. Ainsi, à l'avenir, cette procédure devrait être optimisée par l'utilisation de parseurs de documents HTML, dont des versions efficaces pour Perl ou PHP sont librement disponibles.

Nous avons ensuite écrit une procédure d'identification automatique des phrases en Perl (splitSentence.pl) qui s'applique sur ces textes nettoyés. Pour ce faire, nous avons utilisé le module Sentence.pm, développé par Shlomo Yona (http://search.cpan.org/~shlomoy), et qui est téléchargeable à partir du site *Comprehensive Perl Archive Network* (www.cpan.org).

L'alignement des phrases de chaque version linguistique des textes (néerlandais et français) a été réalisé à l'aide du programme d'alignement développé par (Danielsson et Ridings, 1997), qui est entièrement basé sur l'algorithme de (Gale et Church, 1991). Cet algorithme a l'avantage de fonctionner indépendamment de la langue des textes à aligner.

TEXTE	TAILLE (# mots)	SCORE (%) mode automatique	SCORE (%) mode semi-automatique
DELHAIZE	13 688	83,78	94,17
ACCORD GOUVERNEMENTAL	51 332	78,30	96,77
DEBATS PARLEMENT EUROPEEN	46 635	88,93	92,44
TOTAL	**111 655**	**83,67**	**94,46**

Résultats de l'algorithme d'alignement
de trois échantillons de textes néerlandais-français

Cette série préliminaire de tests nous a amenés à faire les observations suivantes : tout d'abord, nous avons constaté que les sites web multilingues dont nous avons extrait les textes sont structurés de manière très variée. Ainsi, le site www.delhaize.be (textes relatifs à l'alimentation) est une arborescence de niveau variable constituée à la fois de pages HTML intermédiaires (branches) comprenant des hyperliens et de pages HTML finales (feuilles) exemptes d'hyperliens. Cette structure particulière a nécessité l'écriture d'un script sur mesure permettant l'extraction des fichiers textes néerlandais et français du site. Inversement, le site www.europarl.eu.int/plenary (débats du Parlement européen) est très strictement structuré et hiérarchisé, ce qui nous a permis d'écrire un script générique qui extrait les transcriptions des débats parlementaires dans les langues souhaitées (néerlandais-français). L'architecture rigoureuse d'un tel site est motivée par la nécessité d'une mise à jour optimisée car une très grande quantité de fichiers vient l'enrichir à une fréquence hebdomadaire.

Comme indiqué dans le tableau, l'outil d'alignement a été appliqué sur les textes selon deux modes : (i) dans le mode automatique, les textes ont été extraits du site web, nettoyés, découpés en phrases et alignés sans aucune intervention manuelle ; (ii) dans le mode semi-automatique, l'identification (ou découpage) en phrases a été corrigé manuellement avant l'alignement automatique proprement dit. Le score de l'algorithme exprime (en %) le quotient du nombre d'alignements corrects par le nombre total d'alignements générés automatiquement.

Dans le mode automatique, l'algorithme d'alignement obtient des scores plutôt faibles pour les textes moins structurés — qu'il soient sous forme de pages web ou en format pdf — (83,78 % pour les textes *delhaize* et 78,30 % pour les textes *accord gouvernemental*), alors que ce score s'élève à 88,93 % pour les textes *débats du Parlement européen*, à la structure plus rigoureuse.

Dans le mode manuel, l'algorithme d'alignement obtient des scores sensiblement supérieurs au mode automatique pour ces même textes moins structurés (94,17 % pour les textes *delhaize* et 96,77 % pour les textes *accord gouvernemental).* Cette nette amélioration des résultats par rapport au mode automatique s'explique par des différences typographiques importantes des versions néerlandaise et française de ce type de textes, alors qu'on note une moindre amélioration (92,44 %) dans le cas des textes *débats du Parlement européen*, qui observent une structure typographique nettement plus normalisée.

D'une part, ces résultats plaident en faveur d'une amélioration de l'algorithme de segmentation des textes en phrases, notamment lors du traitement de certains signes de ponctuation tels que les suites de caractères «). », « ... mot », « ...) », « - mot » qui sont incorrectement interprétés dans certains cas.

Ensuite, nous devrions procéder à une comparaison des performances de notre programme d'alignement avec d'autres outils existants, ce qui sort du cadre strict de cet article. Enfin, nous comptons tester une approche plus linguistique en vue d'améliorer les performances de notre algorithme d'alignement automatique de corpus bilingues.

4. Conclusion

Nous avons présenté NEDERLEX, un outil original de création de supports en ligne d'aide à la lecture de textes néerlandais, qui exploite de grands corpus bilingues alignés (néerlandais-français) pour illustrer en contexte et sans ambiguïté le vocabulaire de tels textes.

Lors de la mise en œuvre de cet outil, nous avons développé une méthode d'acquisition, de génération et d'alignement de ces corpus bilingues au niveau de la phrase. Les premiers résultats de l'algorithme d'alignement appliqué sur des échantillons représentatifs de notre base de donnée de corpus bilingues — constituée manuellement — sont encore timides mais encourageants : ils génèrent en moyenne plus de 94 % d'alignements corrects (avec une correction manuelle des textes après la phase de découpage en phrases). L'objectif de réduire de manière significative le travail de post-édition manuelle a donc été rencontré. Un des points perfectibles de l'algorithme réside dans l'utilisation de connaissances linguistiques lors de la phase d'alignement.

Sans insister sur les avantages évidents d'une aide lexicale en ligne intégrée dans un outil générique de lecture de textes en langue étrangère sur le Web, nous mentionnerons les principaux bénéfices de l'utilisation de corpus multilingues dans un tel outil.

Du point de vue des concepteurs de l'outil (enseignants), l'acquisition systématique de très grands corpus bilingues a été facilitée par une série d'outils d'extraction de textes multi-

lingues sur le Web et d'alignement de ces textes, qui optimisent le travail de post-édition. Notons que les outils mentionnés n'exigent pas de compétence informatique particulière de la part des utilisateurs.

Du point de vue des utilisateurs finaux (apprenants), les mots des leçons sont systématiquement illustrés à l'aide d'exemples issus de ces corpus bilingues alignés. L'apprenant peut ainsi observer le mot recherché sous plusieurs formes fléchies et sa traduction dans plusieurs contextes minimaux et différenciés, qui sont spécifiques au contexte du mot candidat. Cette aide lexicale désambiguïsée et contextualisée constitue une des plus fortes valeurs ajoutées du système. Notons également que tous les mots des textes font systématiquement l'objet d'une telle traduction, ce qui rend l'outil fortement adaptatif : un même texte peut être déchiffré par des apprenants de niveaux différents.

Références

Brown Peter F., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Lafferty J., Mercer R. et Roossin P. (1990). A statistical approach to machine translation. *Computational Linguistics*, vol. (16/2) : 79-85.

Church K. (1993). Chat_align: a program for aligning parallel texts at the character level. In *Proceedings of the 31st annual meeting of the AC*L : 1-8.

Danielsson P. et Ridings D. (1997). Practical presentation of a vanilla aligner. Paper presented at the telri workshop in alignment and exploitation of texts, llubljana, feb. 1-2 1997. Research reports from the Department of Swedish, Göteborg University GU-ISS-97-2, Språkdata.

Deville G. et Dumortier L. (2002). Tussen de Regels – deel II, Lecture de textes juridiques néerlandais, cours en ligne (www.droit.fundp.ac.be/langues/termino_nl.htm). Facultés universitaires de Namur.

Deville G. et Dumortier L. (2003). Tussen de Regels – deel I, Lecture de textes néerlandais, cours en ligne (www.droit.fundp.ac.be/langues/nl.htm). Facultés universitaires de Namur.

Gale W. et Church K. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* : 177-184.

Hofland K. (1996). A program for aligning English and Norwegian sentences. In G. Perissinotto (Ed.), *Research in Humanities Computing*, vol. (5). Oxford University Press : 165-178.

McEnery A, Oakes M. et Garside R. (1997). CRATER: resource creation for corpus-based machine translation. In Lewandowska-Tomaszczyk B. et Thelen M. (Eds), *Translation and meaning*, Part (4) : 495-500.

Paulussen H. (1999). *A corpus-based contrastive analysis of English 'on/up', Dutch 'op' and French 'sur' within a cognitive framework.* Thèse de doctorat non publiée, Université de Gand.

Simard M., Foster G. et Isabelle P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the fourth international conference on theoretical and methodological issues in machine translation* (TMI92) : 67-81.