# Corpus parallèles, corpus comparables : quels contrastes ?

## Dossier en vue de l'Habilitation à diriger des recherches

## Synthèse

Olivier Kraif

**Jury :**

Nicolas Ballier, Professeur - Université Paris 7
Béatrice Daille, Professeure - Université de Nantes
Jean-Louis Duchet, Professeur émérite - Université de Poitiers - Tuteur
Agnès Tutin, Professeure - Université Grenoble Alpes
Geoffrey Williams, Professeur - Université de Bretagne Sud

Membre supplémentaire
Hélène Chuquet, Professeure émérite (en retraite) - Université de Poitiers

# Sommaire

Remerciements....7
1.Introduction....11
2.Parallélisme et compositionnalité traductionnelle dans les corpus de traductions....14
2.1.Une pratique très ancienne....14
2.2.L'alignement automatique....16
2.3.L'alignement phrastique....17
2.3.1Des corrélations variées : transfuges, cognats, longueurs des phrases....17
2.3.2Cadres algorithmique pour intégrer les corrélations....19
2.3.3Hiérarchiser les corrélations : architecture d'Alinéa....21
2.3.4Évaluation d'Alinéa....23
2.3.5Prolongements dans le domaine de l'alignement phrastique....24
2.4.L'alignement au niveau lexical....27
2.4.1Le repérage de traduction....27
2.4.2Le test de commutation interlingue....29
2.4.3Extraction de correspondances lexicales....33
2.5.L'alignement de corpus multi-parallèles....38
2.5.1Niveau de l'alignement phrastique....39
2.5.2Cadre algorithmique pour un multi-aligneur....40
2.5.3L'aligneur MulItAl....41
2.5.4Cognats et multi-alignement....44
2.5.5L'aligneur JAM....49
2.5.6Tuilage des couples de langue....54
2.5.7Comparaison avec les méthodes binaires....58
2.6.Conclusion....64
3.Quels contrastes ?....66
3.1.Extraction de lexiques bilingues....67
3.2.Une perspective lexicographique ?....69
3.3.De l'aide à la rédaction aux applications didactiques....78
3.4.Vers une cartographie sémantique ?....82
3.4.1Désambiguïsation lexicale....85

3.4.2Construction d'une ressource multilingue de type WordNet pour l'arabe.....90
3.4.3Quelles sont les unités de sens ?.....97
4.Des corpus parallèles aux corpus comparables.....106
4.1.Corpus parallèles vs corpus comparables.....107
4.1.1Hypothèse d'appauvrissement.....109
4.1.2Présence de calques et d'emprunts.....118
4.1.3Complémentarité.....121
4.2.Des corpus aux applications didactiques.....125
4.3.Développement d'outils pour la recherche d'expressions.....128
4.3.1Interface de requête.....129
4.3.2Étude des profils combinatoires : le projet Emolex.....132
4.3.2.1.Visualisation des profils.....134
4.3.2.2.Prise en compte des pivots complexes.....135
4.3.2.3.Extraction automatique d'expressions polylexicales.....137
5.Perspectives.....148
6.Références.....153
Annexe.....165
Annexe - 1.Activités de bi-concordance proposée par Joseph Rézeau.....165
Traduction de ON en anglais .....165
Exercice 1 : Repérage.....165
Exercice 2 : Complétez les traductions de on (en vous aidant de vos constatations de l'exercice 1).....167
FOR + Groupe Nominal + TO-INFINITIF.....168
Exercice 1 : Repérages.....168
Exercice 2 Complétez les citations anglaises.....169
Annexe - 2.Composition des corpus comparables DE-Source et FR-Source.....171
Annexe - 3.Types de noms apparaissant dans diverses constructions.....173
cacher + DetPoss + N sans négation :.....173
Classe Emo - :.....173
Classe Emo + :.....173
Ne pas cacher + DetPoss + N.....173
Classe Emo - :.....173
Classe Emo + :.....174

# Index des figures

Figure 2.1 : Chemin d'alignement pour le corpus BAF/Verne.....19

Figure 2.2 : Réduction de l'espace de recherche à l'étape 3.....22

Figure 2.3 : Résultats d'Alinéa (système P2) pour la tâche d'alignement de corpus pré-segmenté.....24

Figure 2.4 : Occurrences et cooccurrences de deux unités (n1= 5, n2=4, n12=3).....33

Figure 2.5 : F-mesure des extractions de correspondances lexicales. CO : indice basé sur la cognation (mots apparentés), IM : information mutuelle spécifique, TS : T-score, RV : rapport de vraisemblance, P0 : log-probabilité de l'hypothèse nulle, et PC : combinaison de CO et P0. .....35

Figure 2.6 : Corrélation entre la précision des extractions et leur entropie conditionnelle.....38

Figure 2.7 : Algorithme itératif d'appariement des transfuges.....43

Figure 2.8 : Réduction dans un espace à 2 dimensions des points définis dans le tableau 2.5. 47

Figure 2.9 : Classification hiérarchique ascendante - méthode Ward.....49

Figure 2.10 : Algorithme itératif d'appariement des transfuges.....51

Figure 2.11 : vérification de parallélisme à 3 langues.....52

Figure 2.12 : vérification de parallélisme à 2 langues.....52

Figure 2.13 : Représentation des couples de langues les plus fortement associés.....56

Figure 2.14 : Résultats comparés de Vanilla et JAM pour le corpus français dégradé (blocs de taille 1).....62

Figure 2.15 : Evolution de la précision en fonction de la taille des blocs supprimés.....63

Figure 2.16 : Evolution du rappel en fonction de la taille des blocs supprimés.....63

Figure 3.1 : Concordance extraite à partir d'une requête bilingue d'Alinéa, sur la traduction de With a Donkey in the Cevennes, de Stevenson.....69

Figure 3.2: Une sortie HTML d'Alinea permettant l'exploration des équivalents et de leurs contextes.....70

Figure 3.3 : Exemple de requête avec Linguee.....71

Figure 3.4 : Interrogation de la base et des corpus dans un système d'aide à la rédaction (Kraif & Tutin, 2006).....73

Figure 3.5 : Exemple de requête bilingue avec ConcQuest.....74

Figure 3.6 : Résultats de ConcQuest dans l'interrogation du corpus Emergence.....77

Figure 3.7 : Réseau de relations interlingues manifestant les structurations sémantiques de chaque langue.....84

Figure 3.8 : Exemple de synsets de Princeton Wordnet (PWN) pour l'entrée situation.....91

Figure 4.1 : Accroissement du vocabulaire (lemmes) comparé entre textes orignaux et traductions.....113

Figure 4.2 : Accroissement du vocabulaire (lemmes) comparé entre textes orignaux et traductions (lissé par mélange aléatoire).....114

Figure 4.3 : Accroissement du vocabulaire (formes) comparé entre textes orignaux et traductions (lissé par mélange aléatoire).....114

Figure 4.4 : Accroissement comparé du vocabulaire (lemmes) pour des textes de FR-Source .....116

Figure 4.5 : critères de sélection de texte dans l'interface de reFLEx.....126

Figure 4.6 : visualisation comparatives de différentes « facettes » des textes choisis.....127

Figure 4.7 : Assistant graphique pour la construction des requêtes dans ConcQuest.....131

Figure 4.8 : Interface de recherche simple pour le corpus Scientext.....131

Figure 4.9 : Classification hiérarchique et AFC (domaine sémantique de la 'colère').....135

Figure 4.10 : Extraction itérative d'une expression complexe (vouer une admiration sans borne).....138

Figure 4.11 : Généralisation d'une expression polylexicale dans une requête soumise à EmoConc.....143

Figure 5.1 : Extractions d'expression polylexicales et affichage statique des résultats.....151

## Index des tableaux

Tableau 2.1 : Amélioration de l'alignement grâce à la réduction graphique sur un corpus fr-ar (Arcade 2, corpus non pré-segmenté).....26

Tableau 2.2 : Correspondances lexicales correctes vs aléatoires.....36

Tableau 2.3 : Corpus de test de MulItAl.....43

Tableau 2.4 : Résultats de MulItAl sur le corpus Bovary.....44

Tableau 2.5 : Nombre de transfuges et cognats identifiés dans les bi-phrases par couples de langues.....46

Tableau 2.6 : Nombre de transfuges et cognats avec le texte grec translittéré.....46

Tableau 2.7 : Répartition des langues sources dans le corpus Europarl-00-01-17.....48

Tableau 2.8 : Résultats de JAM pour les combinaisons FR-pivot.....53

Tableau 2.9 : Exemple de points obtenus avant complétion finale.....54

Tableau 2.10 : Filtrage des trois langues les plus proches, par ligne.....55

Tableau 2.11 : Résultats comparés pour différents tuilages.....57

Tableau 2.12 : Résultats comparés de Vanilla et des différentes versions de JAM (avec et sans l'application a posteriori de l'algorithme de Gale & Church).....59

Tableau 2.13 : Groupes obtenus par fusion transitive des 11 alignements de référence avec le français....60

Tableau 2.14 : Alignements transitifs simples issus de JAM....60

Tableau 2.15 : Résultats comparés de Vanilla et JAM (CombMax + GC) pour le corpus français dégradé (blocs de taille 1)....62

Tableau 2.16 : Evolution des résultats en fonction de la taille des blocs supprimés....63

Tableau 3.1 : Extrait d'un lexique bilingue tiré d'un alignement anglais-français de with a Donkey in the Cevennes, de Stevenson....67

Tableau 3.2 : Un exemple de bi-concordance centrée sur "pour", extraite du Petit Prince (Antoine de Saint Exupéry) (Lamy & Klarskov Mortensen, 2012)....80

Tableau 3.3 : Unités équivalentes à l'italien carta....83

Tableau 3.4 : Résultat de la désambiguïsation bilingue manuelle....87

Tableau 3.5 : Réduction des sens pour une méthode de désambiguïsation non supervisée....89

Tableau 3.6 : Corrélation entre s et la proportion des sens éliminés....90

Tableau 4.1 : Répartition des occurrences de Schadenfreude en fonction de la langue source ....110

Tableau 4.2 : Composition des corpus parallèles comparables DE-FR et FR-DE....112

Tableau 4.3 : Quelques exemples de variations morphologiques....115

Tableau 4.4 : extrait du lexicogramme pour le nom lemmatisé surprise pris en tant qu'objet direct (f=fréquence de cooccurrence, f1=fréquence de l1, f2=fréquence de l2....134

Tableau 4.5 : extrait de lexicogramme pour le pivot complexe avouer son + N....137

Tableau 4.6 : Liste des expressions polylexicales extraites pour colère pris en tant qu'objet direct (corpus de presse)....139

Tableau 4.7 : Influence de la négation dans la construction cacher + DetPoss + N vis-à-vis de la classe des noms d'affect....142

Tableau 4.8 : Influence de la détermination dans la construction éprouver + N....142

Tableau 4.9 : Répartition des constructions en fonction du genre, dans un échantillon de 16 millions de mots du corpus Emolex (articles de presse vs romans)....145

Tableau 4.10 : Répartition des constructions avec ne pas cacher DetPoss + N en fonction du genre (articles de presse vs romans)....146

Tableau 4.11 : Répartition des constructions avec ne pas cacher DetPoss + N en fonction du genre (articles de presse vs romans)....146

# Remerciements

Mes remerciements vont surtout, et cela va de soi dans un travail de recherche, à tous les collègues avec qui j'ai eu la chance d'avoir des échanges et des collaborations fructueuses, au cours de ces presque vingt ans de recherche. Même si l'écriture d'une synthèse est un travail solitaire, le travail de recherche dont elle rend compte est le fruit de ces collaborations.

Je commencerai par remercier chaleureusement Jean-Louis Duchet, qui a accepté de m'épauler dans ce travail de synthèse et a fait preuve d'une très grande disponibilité pour me conseiller et m'orienter dans ce travail. J'ai eu le plaisir de collaborer avec lui à plusieurs reprises depuis un peu plus d'une dizaine d'années, et j'ai toujours trouvé beaucoup d'intérêt à échanger avec lui autour de langues aussi variées que l'anglais, l'albanais, l'italien, l'espagnol – et même des familles linguistiques spéculatives telles que la macro-famille nostratique ! – sans compter qu'il « pratique » couramment aussi bien Mac OS X, Linux que Windows. Cette synthèse lui doit beaucoup, grâce à sa relecture attentive et minutieuse, et à ses nombreuses remarques qui ont permis de la compléter et d'en améliorer la clarté.

Mes remerciements vont aussi à Henri Zinglé, mon directeur de thèse, malheureusement décédé il y a quelques années : j'ai fait mes premières armes sous sa tutelle, et son parcours éclectique de germaniste, linguiste et informaticien, m'a montré qu'il était possible de travailler au croisement de différents disciplines sans devoir faire le choix de sacrifier l'une au détriment de l'autre. Il m'a également appris, par son exemple, que l'enthousiasme est le plus puissant des moteurs pour avancer dans la recherche.

Je voudrais également rendre hommage à Jean Véronis, lui aussi disparu trop tôt, avec qui j'ai eu le plaisir de collaborer pour le projet Arcade 2. Esprit visionnaire et réaliste à la fois, il m'a fait comprendre que les plus belles réussites du TAL n'étaient pas forcément celles inspirées par l'Hubris – l'esprit de démesure – le progrès arrivant souvent par le biais de modestes outils répondant à un réel besoin.

J'adresse un grand merci à Marc El-Bèze, qui m'a accueilli pendant un an au LIA, à Avignon, et aux autres co-équipiers du projet Carmel, le premier projet d'importance auquel j'ai participé : Claude de Loupy, Grégoire Moreau de Montcheuil, Régis Meyer et Claude Richard ainsi que Boxing Chen, Mériam Haddara et Bettina Schader qui faisaient partie de l'équipe grenobloise.

Parmi mes collègues du Lidilem, je remercie tout particulièrement Agnès Tutin, qui m'a fait l'amitié de participer à ce jury d'habilitation, et avec qui j'ai eu la chance de collaborer sur de nombreux projets en linguistique de corpus, depuis 2003 : Emergence, Scientext, Emolex puis Termith (avec une contribution très modeste de ma part pour ce dernier projet). J'ai beaucoup appris à son contact, notamment en ce qui concerne la phraséologie et ses prolongements dans l'exploration du discours et de la textualité. Concernant le projet Scientext, merci également à Achille Falaise, qui a beaucoup travaillé pour améliorer l'interfaçage avec ConcQuest, tant en termes d'ergonomie que de richesse des requêtes, des grammaires et des résultats.

Je tiens également à remercier Iva Novakova et Peter Blumenthal, coordinateurs du projet Emolex, qui fut une très belle aventure scientifique de 2010 à 2013 – aventure qui n'est d'ailleurs sans doute pas terminée. Merci à tous les collègues de l'équipe franco-allemande, avec qui ce fut un bonheur de travailler : Magdalena Augustyn, Cristelle Cavalla, Vannina Goossens, Francis Grossmann, Sylvain Hatier, Mathieu Loiseau, Elena Melnikova, Joanna Socha et Julie Sorba – pour l'équipe grenobloise – ainsi que Sascha Diwersy, Beate Kern, Anke Grutchus, Dirk Siepmann – pour les équipes allemandes. Un merci tout spécial à Sascha Diwersy, avec qui j'ai entretenu une collaboration étroite tant sur les aspects ingénieriques que scientifiques du développement d'EmoBase, et qui m'a beaucoup appris sur les travaux de Sinclair, et l'école britannique de la linguistique de corpus.

Merci à mes plus proches collègues de travail, les collègues du Département d'informatique pédagogique, et en particulier Thomas Lebarbé, Claude Ponton et Virginie

Zampa, avec qui j'ai longtemps collaboré sur les questions d'apprentissage des langues assistée par ordinateur.

Je remercie enfin, avec chaleur, Nicolas Ballier, Hélène Chuquet, Béatrice Daille et Geoffrey Williams qui m'ont fait l'honneur de s'intéresser à mes travaux en acceptant de participer à ce jury d'HDR.

*La traduction pourrait enfin révéler la linguistique à elle-même. (...) Elle permet en effet de réintroduire pleinement l'activité interprétative dans la communication linguistique, en ouvrant la voie à sa reconception comme une interaction au sein du texte et de l'intertexte.*

François Rastier, La traduction : interprétation et genèse du sens, Marianne Lederer et Fortunato Israël, éds. , *Le sens en traduction*, Paris, Minard, 2006

*D'autre part, les signes dont la langue est faite, les signes n'existent que pour autant qu'ils sont reconnus, c'est-à-dire pour autant qu'ils se répètent ; le signe est suiviste, grégaire ; en chaque signe dort ce monstre : un stéréotype : je ne puis jamais parler qu'en ramassant ce qui traîne dans la langue.*

Roland Barthes, leçon inaugurale de la chaire de sémiologie littéraire du Collège de France, prononcée le 7 janvier 1977

# 1. Introduction

C'est avec une certaine appréhension que j'ai abordé la rédaction de cette synthèse : retracer presque vingt années de recherche me paraissait à la fois inutile et fastidieux. Souvent pris par l'urgence des dates butoirs pour publier ou communiquer, les impératifs des projets en cours, et l'excitation de nouvelles idées, les enseignants-chercheurs n'ont guère l'occasion – ni l'envie – de regarder dans le rétroviseur.

Quand j'ai entrepris de collationner certains de mes articles pour réunir le dossier de publications, et tenter de retrouver un fil conducteur au milieu de recherches souvent guidées par des inspirations ponctuelles, des rencontres et des collaborations occasionnelles, j'ai pourtant été soulagé d'y trouver une certaine cohérence, et j'ai pris un certain plaisir à retrouver dans des questionnements antérieurs l'amorce de mes recherches actuelles. À ma grande surprise, j'ai repris mes recherches sur des terrains que j'avais délaissés pour passer à d'autres thématiques plus actuelles, terrains que je croyais clôts car amplement labourés par la communauté et sans grande perspective de nouveauté.

C'était le cas de l'alignement phrastique : il faut bien avouer que l'état de l'art n'a guère évolué depuis les années 1990, et que beaucoup pensent avoir fait le tour de la question. En reprenant des recherches interrompues depuis 2004, j'ai voulu trouver la réponse à des questions laissées en suspens, notamment sur la question du multi-alignement. Non que je considère en soit l'alignement phrastique comme un sujet de grande portée scientifique – mais parce qu'à travers l'alignement se posent des questions plus générales sur l'activité

traduisante et les phénomènes linguistiques en général : quels sont les réseaux de correspondances que l'on peut observer à travers un multi-texte ? révèlent-ils des propriétés sur le plan génétique ? sur le plan de la synchronie, les versions traduites en plusieurs langues d'un même texte permettent-elles d'éclairer l'original, d'y révéler certaines propriétés, d'en expliciter le contenu ? Les multi-textes constituent-ils un objet linguistique à part, dont on peut extraire des régularités et des équivalences générales ? En somme, constituent-ils un terrain privilégié pour observer des contrastes entre les langues ? Ou bien faut-il les considérer avec méfiance, dans la mesure où ils sont susceptibles de porter toute sorte de biais traductionnels – calques et emprunts –, et les écarter *a priori* des ressources utiles en linguistique de corpus ? des textes parallèles ne devraient-ils pas occuper une place de choix dans le domaine de la lexicographie bilingue ?

Toutes ces questions étaient déjà au cœur de mes recherches de doctorat. Et la perspective qui était la mienne en 1995 n'a au fond pas changé : le développement de techniques et d'outils informatiques n'est pas pour moi une fin en soi, mais un moyen pour explorer et mettre au jour des phénomènes linguistiques. Ma perspective n'est pas, comme souvent en Traitement automatique des langues (TAL), de développer et d'améliorer des méthodes et des modèles, à la fois informatiques et linguistiques, en vue de certaines applications industrielles. À l'inverse, l'objet de mes recherches a toujours été la langue elle-même, avec l'objectif développer de nouveaux outils et instruments susceptibles de faire émerger de nouveaux faits. Ce que j'ai souvent cherché à élaborer, à travers le TAL, ce sont des « dispositifs expérimentaux » tels que Habert les a définis (2005) :

> On utilisera la dénomination *dispositif expérimental*, empruntée à la sociologie de l'innovation développée par Bruno Latour, pour un montage d'instruments, d'outils et de ressources servant à produire des « faits » dont la reproductibilité et le statut (l'interprétation) font l'objet de controverses. Ressortissent partiellement aux dispositifs expérimentaux les aligneurs.

Depuis 2002, la pente naturelle de mes recherches m'a poussé vers l'analyse et le traitement de corpus comparables, voire monolingues, afin de développer des méthodes pour en extraire des unités pertinentes et des propriétés combinatoires. Mais à part certaines recherches adventices dans le domaine de la génération d'activités en ALAO (apprentissage des langues assistées par ordinateur), que je développerai assez peu dans cette synthèse, mes travaux sur les corpus comparables et monolingues visent au fond le même objet : fournir des

instruments pour identifier des unités de sens et en explorer les distributions, en articulant la perspective monolingue et la dimension contrastive, car elles s'enrichissent et se complètent mutuellement.

La présente synthèse se déroule en trois temps, qui correspondent peu ou prou à la chronologie des articles figurant dans mon dossier de publications : dans un premier chapitre, j'aborde le problème de la compositionnalité traductionnelle, sous-jacent au problème de l'alignement multilingue – notion que je remets en cause au niveau lexical. J'en profite pour développer une technique originale de multi-alignement, afin de montrer l'intérêt qu'il y a à s'appuyer sur la convergence des réseaux de correspondances lorsque plus de deux langues sont mises en jeu.

Dans un deuxième chapitre, j'essaie d'approfondir la notion de contrastes observables à travers les multi-textes, essentiellement sur les plans du lexique et de la sémantique. J'y discute une méthode d'extraction automatique de lexique sémantique s'appuyant sur les cliques multilingues, ce qui m'amène à m'interroger sur la consistance sémantique des unités lexicales considérées hors contexte : en recyclant des données dictionnairiques, je montre qu'on parvient difficilement à constituer des cliques multilingues avec des mots simples, l'équivalence traductionnelle étant trop instable à ce niveau.

Dans le troisième et dernier chapitre, j'étends ces questionnements aux corpus comparables, et je tente d'identifier ce que de tels corpus multilingues non-parallèles peuvent apporter à la recherche des contrastes (et des équivalences). Ce parcours m'amène à discuter le parti-pris consistant à éliminer les traductions du champ de la linguistique de corpus, sous prétexte de biais traductionnel. Enfin, j'aborde le développement de quelques instruments, au sens de Habert (2005), mis en œuvre dans le cadre des corpus monolingues et comparables. Nous verrons que ces instruments, dédiés à l'étude de la combinatoire du lexique, permettent d'ouvrir des perspectives intéressantes dans la recherche des unités de sens, problématique qui se situe au cœur des méthodes contrastives que nous avons essayé de mettre en œuvre.

# 2. Parallélisme et compositionnalité traductionnelle dans les corpus de traductions

## 2.1. Une pratique très ancienne

Un corpus parallèle est un ensemble de textes accompagnés de leurs traductions dans une autre langue. Comme le note Véronis (2000 : 2), bien que la systématisation de l'exploitation de ce type de corpus en TAL ne date que de la fin des années 1980, l'existence de textes parallèles remonte à la plus haute Antiquité : en attestent les inscriptions bilingues des tombes des princes d'Elephantine en Égypte, qui datent du troisième millénaire avant J.-C., bien avant la pierre de Rosette (196 av. J.-C.)[1]. L'usage des textes parallèles est peut-être aussi ancien que la pratique de la traduction écrite, et durant l'antiquité certains textes sacrés, déjà, étaient présentés dans des versions bilingues parallèles, afin d'en faciliter l'accès et l'exégèse : c'est par exemple le cas d'une des plus anciennes versions des Évangiles, le *Codex Bezae Cantabrigiensis*[2], que l'on date vers la fin du IV$^{e}$ siècle.

La fascination liée à ce type de texte provient peut-être de ce qu'elle met en présence deux langues, source et cible, et de ce fait révèle leur nature même de *code* – au sens saussurien et cryptographique à la fois – la version traduite permettant en quelque sorte de

1 Cf. *Encylopedia Universalis*, http://www.universalis.fr/encyclopedie/traduction/, consulté le 5/2/2014.

2 D.C. Parker, *Codex Bezae. An early Christian manuscript and its text*, Cambridge 1992.

*décoder* un message initialement *chiffré*. Dans le cas de la pierre de Rosette, la traduction a même joué le rôle de *clé* de déchiffrement, l'intégralité d'un code encore inconnu – l'écriture hiéroglyphique – ayant pu être déchiffrée par Champollion grâce à la clé fournie par sa transcription en démotique et à sa traduction en grec ancien.

Ce que montre l'exemple de la Pierre de Rosette, c'est que d'une part le texte traduit permet de révéler – dans son sens étymologique de « dé-voiler » – le sens du texte source, mais d'autre part, il permet de mettre en lumière des propriétés de la langue source. La traduction nous parle à la fois du texte d'origine, mais aussi de son idiome, et par ricochet, dans le jeu des écarts et des différences, de l'idiome d'arrivée. C'est cette propriété de *révélateur* qui a motivé mon intérêt tout au long de mes recherches sur les corpus parallèles, qui ont commencé avec l'idée de réutiliser des traductions déjà faites pour en quelque sorte les recycler.

L'idée de rassembler des corpus de textes traduits dans une perspective de recyclage des traductions est apparue à la fin des années 1970, entre le *Xerox Parc* et la *Brigham Young University*. Nagao (1984) propose une méthode de traduction basée sur l'exemple. Le début des années 1980 verra également la constitution du premier corpus parallèle bilingue de grande envergure, le corpus Hansard, qui regroupe des textes issus du Sénat canadien. Le terme de corpus *parallèle* s'est peu à peu imposé dans les années 1990, la propriété géométrique du parallélisme désignant par analogie une propriété caractéristique de la traduction : sa *compositionnalité* – que Pierre Isabelle (1992 : 724), définit ainsi « (…) les traductions obéissent à un principe dit de compositionnalité : la traduction d'un segment complexe est généralement une fonction de la traduction de ses parties, et ce, jusqu'au niveau d'un ensemble d'unités élémentaires ». Ajoutons que le parallélisme implique que les segments issus de cette décomposition se succèdent dans un ordre identique. Ainsi, deux idées sous-tendent en général la notion de parallélisme :

- *Compositionnalité* : la relation d'équivalence traductionnelle globalement mise en jeu entre deux textes, peut se décomposer au niveau de segments plus petits (p.ex. des chapitres, des paragraphes, des phrases,…), également équivalents sur le plan de la traduction.

– *Séquentialité* : les segments équivalents apparaissent dans le même ordre dans la cible et dans la source.

Ces deux propriétés se manifestent visuellement dans les éditions bilingues, le texte cible en page de droite étant mis en regard, page après page, du texte cible en page de gauche.

Notons qu'un corpus parallèle ne contient pas nécessairement le texte original en langue source. Teubert (1996 : 245) donne une définition générale indiquant différents cas de figure que l'on peut rencontrer :

> Un corpus parallèle est un corpus bilingue ou multilingue qui contient un ensemble de textes en deux langues ou plus. Il y a plusieurs cas de figure, parmi lesquels :
> - un corpus parallèle contient des textes originaux dans une langue A et leurs traductions dans d'autres langues B, C, etc.
> - un corpus parallèle contient une quantité égale de textes originaux dans les langues A et B, et leurs traductions respectives
> - un corpus parallèle contient seulement des traductions de textes dans des langues A, B et C, originellement écrits dans une langue Z. (*Nous traduisons*)[3]

## 2.2. L'alignement automatique

L'opération consistant à mettre en correspondance les segments équivalents s'appelle, de façon naturelle, *l'alignement*. Pour dénommer simplement un texte parallèle bilingue aligné, B. Harris (1988) propose le terme de *bi-texte*, repris ensuite par Isabelle (1992), et généralisé à *multi-texte* dans le cas de plus de deux langues (comme pour le corpus JRC-ACQUIS, qui implique 21 langues dans sa version 2.2, cf. Steinberger *et al.*, 2006). La tâche d'alignement s'articule donc en deux phases :

– *Segmentation*, au niveau choisi (section, sous-section, paragraphe, phrase, syntagme…).
– *Appariement* des segments équivalents.

[3] " A 'parallel corpus' is a bilingual or multilingual corpus that contains one set of texts in two or more languages. There are several options, among them: - a parallel corpus containing only texts originally written in language A and their translations into languages B (and C.. .) - a parallel corpus containing an equal amount of texts originally written in languages A and B and their respective translations - a parallel corpus containing only translations of texts into the languages A, B and C, whereas the texts were originally written in language Z. "

## 2.3. L'alignement phrastique

Aligner des chapitres, voire des paragraphes, peut se révéler trivial, car la traduction conserve, le plus fréquemment, la structuration des unités textuelles étendues. En revanche, si l'on s'intéresse au niveau de la phrase, on constate qu'il y a rarement une correspondance biunivoque au niveau des phrases sources et cibles, en se basant sur une définition typographique simple des phrases (p. ex. découpage au niveau des signes de ponctuation forte). Il est fréquent qu'une phrase soit traduite par 2 phrases ou plus dans le texte cible – une phrase peut également être omise (absente de la traduction).

La tâche d'alignement phrastique va donc consister à apparier non seulement des phrases, mais le cas échéant des groupes de phrases (ou pas de phrase du tout dans le cas des omissions ou des ajouts).

### 2.3.1 Des corrélations variées : transfuges, cognats, longueurs des phrases

La première méthode d'alignement automatique, mise au point par Martin Kay & Martin Röscheisen (1988), ouvrira la voie à de nombreux développements. Ces auteurs implémentent une méthode basée sur la distribution des mots, en n'utilisant aucune source d'information complémentaire en dehors des deux textes à aligner. Ils montrent qu'en observant des cooccurrences de mots à l'intérieur de zones probablement correspondantes (le début et la fin des textes, ainsi que les zones se situant au même niveau, dans chacun des textes) il est possible d'extraire des correspondances lexicales, qui peuvent servir ensuite de « points d'ancrage » pour aligner les phrases. Le grand mérite de ces premières recherches est de montrer qu'il est possible d'aligner sans passer par une connaissance précise des deux langues, en se basant sur des propriétés purement formelles.

C'est ce que constate rapidement un annotateur humain confronté à une tâche d'alignement phrastique manuel : même sans connaître les langues impliquées, il peut s'appuyer sur des indices superficiels pour prendre ses décisions. Considérons l'exemple suivant :

> *Il faudra développer les recherches de charbon à pouvoir calorifique plus élevé et transformable en coke, s'employer partout à substituer le charbon aux carburants liquides et à le consommer avec économie.*

*Të zgjerohen kërkimet për qymyre me fuqi kalorifike më të lartë dhe të koksifikueshme.*
*Të punohet kudo për zëvendësimin e karburanteve të lëngëta me qymyr dhe për kursimin e tij.*[4]

Sans connaître l'albanais, un rapide coup d'œil permet de trouver des mots ressemblant, qui confirment que ces groupes de phrases sont bien alignés :

*calorifique* *kalorifike*
*carburants* *karburanteve*

C'est ce qu'on appelle des paires de *cognats* (de l'anglais *cognate*). Bien que le français et l'albanais ne soient pas génétiquement apparentées, elles partagent, comme beaucoup de langues, un fond lexical commun autour du vocabulaire scientifique et technique. Ces ressemblances entre mots apparentées présentent l'intérêt d'être facilement exploitables par la machine, grâce à une simple comparaison de chaînes de caractères. Simard, Forster et Isabelle (1992) ont été les premiers à développer un système d'alignement basé sur le repérage des cognats.

Considérons un autre exemple, tiré d'un corpus français arabe[5] :

*Nous, qui savons que les territoires ne sont pas Israël, et qu'à la fin toutes les colonies devront être évacuées. (...) Nous n'allons plus combattre au-delà des frontières de 1967 afin de dominer, d'expulser, d'affamer et d'humilier un peuple entier.*

نحن الذين نعرف ان الاراضي ليست اسرائيل وان جميع المستوطنات سوف تخلى في نهاية المطاف(...) لن نذهب بعد اليوم الى الحرب خارج حدود *1967* لكي نفرض سيطرتنا على شعب بأكمله ونقوم بطرده وتجويعه واذلاله.

Dans ce cas, on ne peut plus simplement comparer les chaînes de caractères, mais il reste toutefois des indices de surface assez fiables : les nombres et la ponctuation. Cette fois les chaînes comparées sont identiques : c'est ce qu'on appelle des *transfuges* (Langé & Gaussier, 1996) qui passent d'un texte à l'autre sans altération. Dans cette catégorie, on trouve certaines ponctuations ainsi que de nombreuses entités nommées (noms propres, sigles, dates, etc.).

4 *Tiré de : Enver Hoxha (1981) RAPPORT D'ACTIVITE DU COMITE CENTRAL DU PARTI DU TRAVAIL D'ALBANIE, VIIIe Congrès du PTA*, 1er novembre 1981, Editions « 8 NËNTORI » TIRANA, 1981 [URL: http://ciml.250x.com/archive/hoxha/french/eh_rapport8pta1.html, consulté le 5/02/2014]

5 Algazy J. (2002), Ces soldats israéliens qui disent non, *Le Monde Diplomatique*, mars 2002.

Enfin, il existe un autre indice de surface encore plus trivial : les longueurs de phrases. Un simple comptage permet d'établir que dans l'exemple français-albanais, la phrase en français contient 31 mots[6] alors que les deux phrases équivalentes en comptent 30. Pour l'exemple français-arabe, on a respectivement 40 et 34 mots. Bien évidemment, il n'y a pas identité du nombre de mots, mais on constate, sur un plan statistique, une corrélation significative des longueurs de phrases (en mots ou en caractères), autour d'un rapport moyen qui est propre à chaque couple de langues (certaines étant plus économes en mots ou en caractères).

Brown & Lai (1991) ont proposé une méthode basée sur ce principe, et l'algorithme de Gale & Church (1991), s'appuyant sur une méthode similaire, fait encore référence aujourd'hui.

### 2.3.2 Cadres algorithmique pour intégrer les corrélations

Aligner deux textes consiste à extraire automatiquement un « chemin d'alignement », c'est-à-dire un ensemble de correspondances entre phrases ou groupe de phrases qui peut se représenter par une succession de points dans un espace à deux dimensions (chaque dimension correspondant à un texte). La figure 2.1 ci-dessous représente le chemin pour l'alignement du sous-corpus Verne tiré du corpus BAF (Simard, 1998).

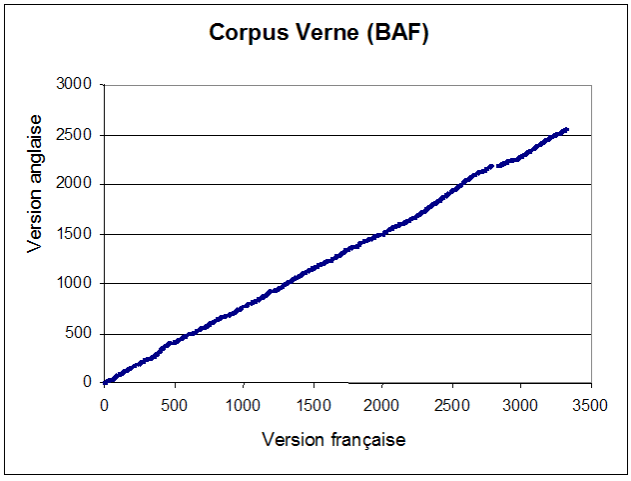


*Figure 2.1 : Chemin d'alignement pour le corpus BAF/Verne*

À l'instar de Simard & Plamondon (1996), on peut distinguer deux types d'algorithmes pour le calcul de l'alignement :

6 On parle de mots "graphiques" tels qu'ils sont dénombrés par les traitements de texte. Ici, en l'occurrence, LibreOffice 3.5

- Les algorithmes *matriciels* (*bi-text mapping*, ibid.) qui découpent cet espace bidimensionnel en un quadrillage, afin d'identifier des points au niveau des couples de formes qui sont potentiellement alignables. Les formes peuvent être appariées sur la base de leurs ressemblances graphiques (Church, 1993) ou bien sur la similarité de leurs distributions au sein de ce quadrillage (Fung & Church, 1994). Les points sont ensuite filtrés en fonction de leur compatibilité avec les points voisins, et de leur proximité à la diagonale. De cette manière, on extrait des *points d'ancrage*, qui permettent éventuellement de redéfinir un quadrillage plus fin, afin de réitérer l'opération.

- Les algorithmes *linéaires* qui recherchent la suite optimale de regroupements – ou *transitions* – de type 1-1, 1-2, 2-1, etc. du début à la fin du bi-texte. Le nombre de chemins possibles à évaluer étant exponentiel, en fonction de la taille des textes, on utilise des algorithmes de programmation dynamique[7], tel que l'algorithme de Viterbi, pour calculer récursivement le meilleur chemin. Les méthodes basées sur les longueurs de phrases (Gale & Church, 1991) ainsi que celles combinant longueurs et ressemblances lexicales (Simard, Foster & Isabelle, 1992 ; Mc Enery & Oakes, 1995) s'appuient sur ce principe. Chez Gale & Church (1991) la probabilité d'un appariement de phrases est calculée comme le produit de la probabilité du rapport des longueurs[8] et de la probabilité empirique de la transition considérée[9]. La probabilité d'un chemin complet est alors le produit des probabilités de tous les appariements successifs.

---

7 La programmation dynamique vise à ramener la résolution globale d'un problème à la résolution de sous-problèmes plus simples. Elle s'applique lorsque la solution optimale du problème pris globalement peut être conçue comme la combinaison de solutions optimales obtenues pour une série de sous-problèmes. Ici, on considère que le chemin optimal permettant d'arriver à l'alignement des deux dernières phrases du bi-texte est fonction des sous-chemins optimaux menant aux phrases précédentes. Le calcul est ensuite réitéré récursivement jusqu'au couple de phrases initiales.

8 Plus exactement, c'est l'écart $\delta$ entre la longueur effective de la phrase cible et sa longueur théorique attendue, normalisée : les auteurs font l'hypothèse que cet écart suit une loi normale centrée réduite. Pour deux phrases de longueurs *l* et *l'*, un rapport moyen de *c*, et une variance *s2* de ce rapport, on a :

$$\delta = \frac{(l' - l \cdot c)}{\sqrt{s^2 l}}$$

9 Les probabilités empiriques pour leur corpus étant les suivantes, 6 transitions seulement étant considérées : $p$(1-1)=0,89, $p$(1-0) = $p$(0-1) = 0,0099, $p$(2-1) = $p$(1-2) = 0,089, $p$(2-2)=0,011

Ces derniers algorithmes étant relativement fragiles lorsqu'ils s'appliquent à de longues portions de textes, on les combine généralement avec les premiers, afin d'extraire au préalable des points d'ancrages fiables permettant de pré-découper les textes. Dans cette optique, à la suite de Kay & Röscheisen (1987,1993), Débili & Sammouda (1992) montrent qu'il n'y a pas de cercle vicieux dans le fait d'utiliser successivement l'alignement des mots pour aligner les phrases, et l'alignement des phrases pour aligner les mots : le processus converge vers un alignement de plus en plus précis, chaque étape apportant de nouvelles informations.

Enfin, Davis, Dunning & Ogden (1995), dans le souci de tenir compte des ruptures de parallélisme fréquente dans les traductions réelles, montrent comment combiner différents types d'indices pour les intégrer dans un même cadre algorithmique. Avec une approche similaire, des résultats très satisfaisants sont obtenus par Langlais & El-Beze (1997) : divers indices, basés sur les longueurs de phrases, les chaînes identiques (transfuges), les cognats, les probabilités de transitions, sont pondérés de façon à optimiser les performances.

### 2.3.3 Hiérarchiser les corrélations : architecture d'Alinéa

Dans nos propres recherches sur l'alignement phrastique (Kraif, 2001a), nous nous sommes basé sur les principes suivants :

– D'abord, privilégier les méthodes génériques, c'est-à-dire ne s'appuyant pas sur des connaissances linguistiques sur le couple de langues. En effet, comme nous l'avons vu précédemment, ces méthodes sont en général suffisantes pour le niveau phrastique. D'autre part pour un système concernant un grand nombre de langues, le nombre de couples pour lesquels il faudrait développer des ressources (p.ex. des lexiques de transfert) explose littéralement : p.ex. pour les 24 langues de l'UE (en 2014), il faut prendre en compte 24*23=552 paires de langues (en tenant compte de la direction source → cible) et 276 couples (sans tenir compte de la direction).

– Pour tirer le meilleur parti de tous les indices superficiels (transfuges, cognats, longueur de phrases, transitions…), il faut travailler dans un cadre itératif permettant de s'appuyer sur les indices les plus fiables d'abord (ce que fait un humain, lorsqu'il effectue un pré-découpage des textes…). Nous avons nommé ce principe l'heuristique de précision d'abord.

– Enfin, pour mieux identifier les mots ressemblants apparentés (les cognats) on peut mettre en œuvre des techniques plus sophistiquées que la semble recherche de n-grammes (n-caractère consécutifs identiques en début de chaîne) habituellement utilisée.

Sur la base de ces principes, nous avons bâti une architecture en 3 étapes :

1. extraction de points d'ancrage à partir des transfuges ;
2. extraction de points d'ancrage à partir de la densité de cognats ;
3. calcul de l'alignement entre les points d'ancrage avec la méthode des longueurs de phrase.

Cette dernière étape, moins robuste, s'effectue donc dans un espace de recherche réduit guidé par les points d'ancrage préalablement extraits, comme le montre la figure 2.2.

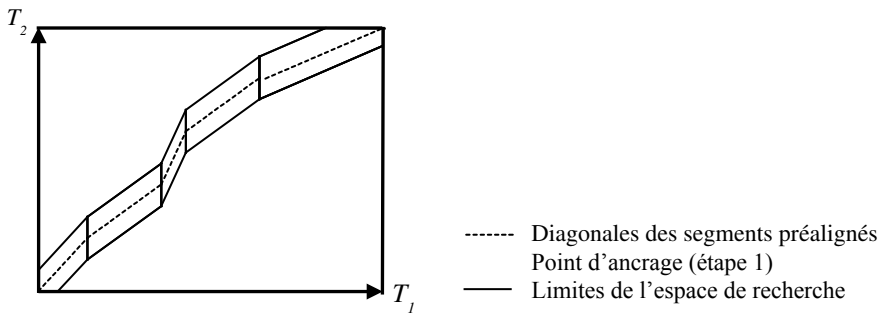


*Figure 2.2 : Réduction de l'espace de recherche à l'étape 3*

En suivant l'heuristique de précision d'abord, nous avons ainsi démontré (Kraif, 2001b), qu'il était possible d'obtenir un préalignement de grande précision (>99%) et de rappel important (plus de 50 % des phrases), sur le corpus BAF, uniquement avec des transfuges, en les priorisant de cette manière : 1/ chaînes alphanumériques, 2/ chaînes commençant par une majuscule, 3/ transfuges quelconques. En outre, on ne retient à chaque itération que les points qui satisfont des critères géométriques stricts : proche de la diagonale, peu déviants par rapport aux points précédents, monotones (c'est-à-dire formant un chemin toujours croissant).

Concernant l'identification des cognats, nous avons comparé empiriquement la méthode classique d'identification par les 4-grammes (Simard *et al.* 1992, Simard *et al.* 1996, Langlais *et al.* 1997), à une autre méthode basée sur la reconnaissance de sous-chaînes communes maximales (SCM, cf. Kraif, 2001b). Dans ce dernier cas, plutôt que de s'intéresser à l'identité

des caractères initiaux, on recherche la plus longue sous-chaîne commune à deux mots. P. ex. entre *préparatoire* et *preparatory* on trouve une sous-chaîne de longueur 9/11 : p-r-p-a-r-a-t-o-r. Notre étude empirique montre qu'en retenant comme candidat toutes les paires de formes dont la SCM constitue au moins les 2/3 des caractères, pour des formes de longueur supérieure à 4 caractères, le rappel est bien meilleur que dans le cas des 4-grammes (cf. figure 2.3).

### 2.3.4 Évaluation d'Alinéa

Pour valider cette architecture, nous avons participé en 2004 à la campagne d'évaluation Arcade 2 (Chiao *et al.* 2006). L'originalité de cette évaluation était de porter sur l'alignement du français avec, d'une part, des langues apparentées (anglais, allemand, espagnol, italien), et d'autre part des langues plus lointaines ou utilisant des alphabets différents (comme l'arabe, le chinois, le farsi, le grec, le japonais et le russe).

Pour le premier groupe de langues, Alinéa obtient une F-mesure (la moyenne harmonique de la précision et du rappel) d'environ 98 %, à 3 dixièmes du meilleur système. Notons que les résultats sont meilleurs pour les couples français-italien, français-espagnol ou français-anglais que pour le couple français-allemand, ce qui montre l'importance de la proximité génétique. Pour le second groupe, Alinéa obtient les meilleurs résultats (mais seul un autre système était en compétition), avec une moyenne de 87,1 % : la dégradation des performances est avérée, mais pas catastrophique. Il existe tout un continuum entre les couples les plus propices (comme le français et le grec, avec 97,6 %) et les plus problématiques (comme le français et le japonais, avec seulement 78,9 % de correction). Ces résultats sont détaillés figure 2.3. Notons que lors de cette tâche, nous n'avons utilisé ni lexique bilingue, ni outil de translittération : le seul prétraitement effectué en amont était la segmentation en phrases, qui garantissait une certaine homogénéité dans la comparaison des alignements (des segmentations différentes conduisant à des alignements corrects différents).

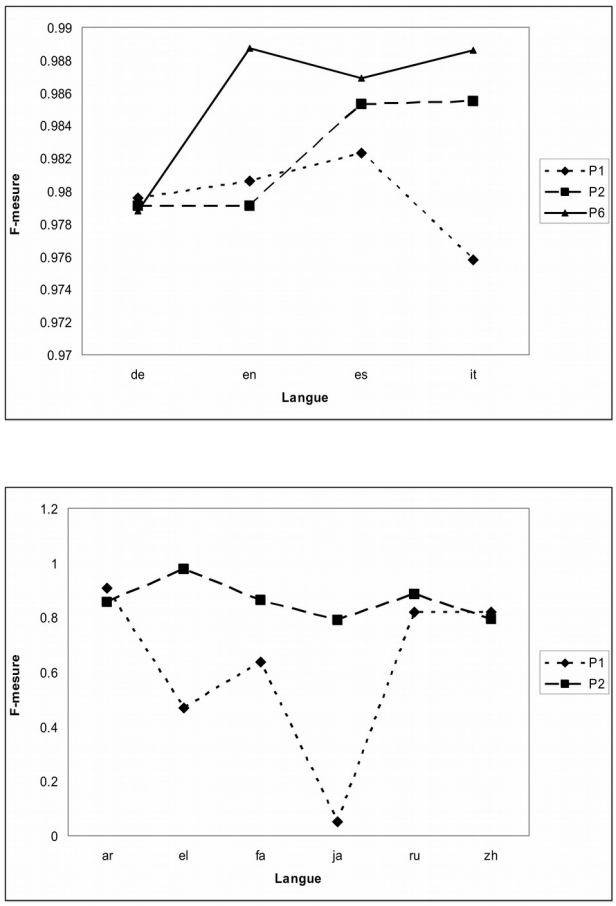


*Figure 2.3 : Résultats d'Alinéa (système P2) pour la tâche d'alignement de corpus pré-segmenté.*

### 2.3.5 Prolongements dans le domaine de l'alignement phrastique

Étant donné l'importance de la première étape d'alignement des transfuges, ceux-ci étant constitués pour une bonne part par des entités nommées – souvent des anthroponymes – il peut être utile d'appliquer un traitement particulier pour des langues à alphabets différents. Dans l'exercice de la traduction, l'usage consiste à utiliser des formes translittérées conventionnelles, résultant d'un ensemble d'équivalences phonologiques/graphémiques plus ou moins régulières. Or il se trouve que les conventions de translittérations sont complexes et pas toujours facilement systématisables. De fait, elles dépendent étroitement du couple de langues impliquées et de l'histoire des échanges linguistiques, débouchant sur une stratification, au fil du temps, de normes concurrentes (ce qu'on retrouve avec des toponymes tels que *Mumbay* ou *Beijing*, en concurrence avec les formes plus anciennes mais toujours en usage *Bombay* et *Pékin*). Pouliquen *et al.* (2007) notent que d'une langue cible à l'autre, des variations dans la translittération sont multifactorielles : variations morphologiques (p.ex. en

slovène, *Tonyem Blairem,* ibid.), variations dans les systèmes graphémiques de la langue cible (p.ex. *Schröder* qui devient *Schroder* en anglais, ibid.), conventions de translittération différentes (p.ex. *Владимир Устинов* qui devient *Wladimir Ustinow* en allemand et *Vladimir Ustinov* en anglais, ibid.) sans compter les variations orthographiques aléatoires (p.ex. *Condoleza Rice, Condaleezza Rice, Condollezza Rice, Condeleeza Rice,* ibid*.)*. Dans le cas de la langue arabe, la situation est encore plus complexe, car les variations de prononciations liées à ses nombreuses variantes dialectales se traduisent naturellement par des translittérations différentes, ces variations étant démultipliées par le fait que l'écriture arabe n'étant pas voyellée, il ne peut y avoir de translittération systématique pour les voyelles brèves. On aboutit, pour certains noms propres, a une véritable prolifération des variantes, comme l'illustre l'exemple donné par Saadane & Semmar (2012), du nom معمر القذافي (*Mouammar Kadhafi* ) « qui est transcrit en latin par plus de 60 formes, parmi lesquelles : *Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi, Muammar Gadafi, Moamer El Kazzafi, Mu'ammar al-Qadhdhafi, Mu'amar Qadafi, Muammar Gheddafi, Mu'ammar Al Qathafi, Muʿammar Al-Qaḏâfî,* etc. ».

Étant donné la complexité de ces systèmes de correspondances, nous avons fait le choix, avec Authoul Abdulhay, de mettre en œuvre, entre le français et l'arabe, un système de translittération *ad hoc*, que nous avons appelé « réduction graphique » (Abdulhay & Kraif, 2008). L'idée est de partir d'un système de transcodage biunivoque vers ASCII, tel que le système Buckwalter[10], et d'appliquer des règles de transformations à la fois sur les formes utilisant l'aphabet latin et sur les formes arabes translittérées avec Buckwalter. Ces règles de transformation visent à rapprocher au maximum les graphies, par réduction des différences. Considérons l'exemple suivant :

*fr : Ignacio Ramonet, ar :* انياسيو رامونه *(translit. AnyAsyw rAmwnh)*

Si on extrait les sous-chaînes communes maximales, on trouve une similarité assez faible :

*SCM (Ignacio ; AnyAsyw) = n-a*
*SCM (Ramonet ; rAmwnh) = r-a-m-n*

Mais en appliquant des règles de transformation ad hoc suivantes :

[10] cf. http://en.wikipedia.org/wiki/Buckwalter_transliteration (consulté le 03/04/2014) - nous avons utilisé la version adaptée à XML de Buckwalter.

*Pour la translittération : w → o, y → i, A → a*
*Pour le français : I → i, ci → si, R → r*

On trouve alors :

*SCM(ignasio, aniasio) = n-a-s-i-o*
*SCM(ramonet, ramonh) = r-a-m-o-n*

Dans son mémoire de master, Authoul Abdoulhay (2006 : 45-46), décrit quelques dizaines de règles de réduction simples, qu'elle a développées empiriquement à partir d'un corpus de 244 couples d'entités nommées (corpus issu du *Monde diplomatique*, utilisé lors de la campagne Arcade 2). L'esprit de ces règles est assez voisin des règles de « normalisation » décrites dans Pouliquen *et al.* (2005), pour le développement d'outils de fouille de texte dans un contexte multilingue (analyse de 25 000 articles par jour dans une trentaine de langues différentes) :

- *accented character → non-accented equivalent*
- *double consonant → single consonant*
- *ou → u*
- *wl (beginning of name)→ vl*
- *ow, ew (end of name) → ov, ev*
- *ck → k*
- *ph → f*
- *ž → j*
- *š → sh* *(Pouliquen et al. 2005)*

Bien que très simples (et linguistiquement pauvres), ces règles permettent néanmoins d'améliorer les résultats des algorithmes d'alignement. Dans nos expérimentations (Abdulhay & Kraif, 2008), nous observons un gain de presque 10 % dans les résultats de l'alignement (cf. tableau 2.1) :

| | Précision | Rappel |
|---|---|---|
| Avec réduction graphique | 85,8 % | 81,7 % |
| Sans réduction graphique | 74,2% | 71,0% |

*Tableau 2.1 : Amélioration de l'alignement grâce à la réduction graphique sur un corpus fr-ar (Arcade 2, corpus non pré-segmenté)*

## 2.4. L'alignement au niveau lexical

Après l'alignement phrastique, on procède souvent à une opération d' « alignement lexical » : ce que font des outils comme Giza++ (Och & Ney, 2003), livré avec la suite MOSES dédiée à la traduction automatique statistique.

Or, si l'hypothèse de compositionnalité traductionnelle est en général valide au niveau des phrases – la relation d'équivalence traductionnelle pouvant être décomposée au niveau de la succession des phrases, moyennant certain regroupements (une seule phrase dans la source ou dans la cible pouvant correspondre à 2 voire 3 phrases…) – cette compositionnalité devient problématique si on considère le niveau des unités lexicales. D'une part, la notion de *monotonie* inhérente à l'alignement (le fait que les unités sources et cibles apparaissent dans le même ordre) est battue en brèche par le fait que chaque langue, du fait de sa syntaxe, impose un ordre spécifique des unités au sein de la phrase. D'autre part, la possibilité de mettre en correspondance un à un les mots sources et cibles serait plutôt le signe d'une mauvaise qualité de traduction – ce qu'on appelle couramment le mot-à-mot – que la norme en vigueur dans la pratique de la traduction.

Pour clarifier ce concept d'alignement au niveau lexical, nous avons étudié de près une des tâches proposées lors de la première campagne d'évaluation Arcade (Langlais *et al.*, 1998), le *repérage de traduction* ou *lexical spotting*. Cette tâche consiste simplement à déterminer, étant donné une unité du texte source, quelle est l'unité ou l'expression équivalente dans la cible.

### 2.4.1 Le repérage de traduction

Considérons l'exemple suivant (tiré du corpus JOC[11]) :

> *Fr. : Eu égard à l'intention de la Commission de présenter un Livre vert sur le secteur des postes dans la Communauté*
> *Angl. : Having regard to the Commission's intention to issue a Green Paper on the postal sector in the Community;*

D'une façon intuitive, on peut en tirer un certains nombres de correspondances :

11 Le corpus JOC, utilisé dans le projet ARCADE, est constitué de questions écrites soumises à la Commission européenne en 1993, dans les Séries C du Journal officiel de la Communauté européenne, et collectées dans le cadre du projet MLCC-MULTEXT [URL : http://www.lpl.univ-aix.fr/projects/multext/CORP/JOC.html].

*(Eu égard à ; Having regard to), (Commission ; Commission), (intention ; intention), (présenter ; to issue), (Livre vert ; Green Paper), (secteur des postes ; postal sector), (Communauté ; Community).*

Ce type de repérage de traduction (pour lequel nous fournirons plus loin des critères), fait apparaître différentes sortes d'informations :

– Le long de l'axe syntagmatique, d'une part, il aboutit à une segmentation spécifique des unités. Certaines de ces unités sont directement issues de ce qu'on pourrait appeler la non-compositionnalité traductionnelle. Par exemple, *Livre vert* et *Green Paper* doivent être appariées en bloc, car la relation d'équivalence ne se décompose pas complètement au niveau des formes qui les constituent. Ainsi, le repérage de traduction peut fournir un critère pour l'extraction de certaines expressions polylexicales figées. Ce critère peut être intégré à des méthodes quantitatives, comme l'a montré Melamed (1997), qui note que pour des unités non compositionnelles, la mesure d'information mutuelle entre unités source et cible est supérieure quand on considère ces unités d'un bloc. Notons que la non-compositionnalité connaît des degrés, et qu'elle peut se manifester avec moins de force au niveau de divergences mineures. Par exemple, le complément nominal *des postes* est traduit par l'adjectif relationnel *postal*. Cette divergence nous a conduit à traiter ces syntagmes en bloc, afin de faire correspondre des unités homogènes. Or, l'examen des occurrences sur l'ensemble du corpus JOC indique que *postal sector* est en relation avec *secteurs des postes* dans 5 cas sur 7 et avec *secteur postal* dans les 2 cas restants, confirmant l'hypothèse d'un emploi préférentiel de la première combinaison. Ainsi, les divergences observées lors du repérage de traduction peuvent indiquer des usages qui auraient peut-être échappé à l'examen monolingue. C'est le cas pour de nombreuses collocations : l'impossibilité de les traduire mot à mot est révélatrice de leur degré de cohésion. Elles sont certes observables dans un corpus monolingue, de par leur récurrence, mais elles sont plus facilement repérables dans un corpus aligné. Ainsi, ce que le repérage de traduction fait apparaître, c'est un niveau de segmentation propre au plan contrastif, qui définit des « unités de traduction » au sens de Vinay et Darbelnet (1958 : 37). Comme le note Véronis (2000) dans la perspective de l'alignement, le repérage monolingue des unités n'est pas indépendant de leur mise en correspondance : « la détermination des unités dans la langue source est dépendante de langue cible (par exemple, il faut aligner d'un

bloc *demande de brevet* et *Patentanmeldung* [BLANK 2000] alors que l'alignement peut se fractionner avec *domanda di brevetto*). » Ces unités de traductions sont intéressantes à deux niveaux : d'une part, elles peuvent révéler l'existence d'une unité phraséologique pertinente au niveau de l'idiome ; d'autre part, capitalisées, elles peuvent intervenir dans le processus de traduction afin d'effectuer un transcodage plus modulaire des unités.

– Sur le plan paradigmatique, on observe une liste d'unités qui portent, dans ce contexte précis, des sens équivalents. Ainsi *intention* est traduit par son cognat *intention*, *Livre vert* est traduit par *Green Paper*, *Eu égard à* par *Having regard to* : ce type de relation, observable en de nombreux points du corpus, est capital pour le traducteur, le lexicographe ou le terminologue. En ce sens, le repérage de traduction constitue une étape préalable à la constitution d'un dictionnaire (général ou terminologique) bilingue. Notons que la correspondance des unités peut dépasser le strict niveau lexical : rien n'empêche de s'intéresser au repérage de morphèmes, ou de traits grammaticaux. L'étude sur corpus permet alors d'observer des régularités concernant l'équivalence d'unités à valeur grammaticale.

### 2.4.2 Le test de commutation interlingue

Comme le note Mahimon (1999), le répérage de traduction peut s'appuyer sur le test de commutation interlingue, suggéré par Catford (1965 : 28), afin de dégager des équivalences entre les unités d'un texte et de sa traduction : « Plutôt que de *se demander où* sont les équivalents, on peut adopter une procédure plus formelle, à savoir la commutation et l'observation de variations concomitantes. En d'autres termes, on peut introduire de manière systématique un changement dans le texte source et observer quels changements éventuels en découlent dans le texte cible. Un *équivalent de traduction textuelle* est donc : *cette portion du texte cible qui change si et seulement si une portion donnée du texte source a été modifiée* » (nous traduisons). La même idée est à l'œuvre dans certaines méthodes d'alignement de textes parallèles, basées sur la reconnaissance des parties variables et les parties constantes dans un corpus d'exemples de traduction, afin d'établir des corrélations d'une langue à l'autre. Malavazos *et al.* (2000) en ont tiré une méthode d'extraction de « modèles de traduction » (*translation templates*) : « L'idée principale est basée sur le constat qu'étant donné une paire de phrases source et cible, toute modification de la phrase source aboutira probablement à un

ou plusieurs changements dans la phrase cible, et qu'il est en outre probable que les unités constantes et variables de la phrase source correspondent respectivement aux unités constantes et variables de la phrase cible. » (nous traduisons). Des deux couples de phrases suivants, les auteurs tirent des correspondances entre les parties constantes et les parties variables :

| | |
|---|---|
| *angl. : Style Manager help menu* | *angl. : Style Manager file menu* |
| *grec : Κατάλογος βοήθειας διαχειριτή ύφους* | *grec : Κατάλογος αρχείων διαχειριτή ύφους* |

D'où les correspondances :

| | |
|---|---|
| *angl. : Style Manager X menu* | *(X,X') = (help, βοήθειας)* |
| *grec : Κατάλογος X' διαχειριτή ύφους* | *(X,X') = (file, αρχείων)* |

Dans ce dernier cas, les commutations ne sont pas produites, mais observées. De ce fait, elles peuvent être extraites automatiquement, par simple comparaison des phrases du corpus. Mais notons qu'il est peu probable qu'un corpus contienne en masse de tels cas de figure, où une seule unité est affectée par la commutation.

Le test manuel suit un parcours redoublé par rapport au test classique de commutation : en introduisant une variation sur le plan de l'expression on produit une variation sémantique ; cette variation sémantique impose ensuite une variation des signifiants cibles afin de rétablir l'équivalence sémantique entre les deux textes. Suivant ce principe, Mahimon (1999 : 37) propose une méthode dédiée à l'alignement manuel des unités lexicales, en reliant les unités qui commutent simultanément dans la source et la cible. Elle donne l'exemple suivant :

> *Fr. : Ce projet de loi prévoira un système de déclaration des maladies infectieuses*
>
> *Angl. : This bill will provide for an infectious disease notification system*

Si on fait commuter *Ce* avec *Chaque* l'équivalence peut-être rétablie en commutant *This* avec *Each* :

> *Fr. : **Chaque** projet de loi prévoira un système de déclaration des maladies infectieuses*
>
> *Ang. : **Each** bill will provide for an infectious disease notification system*

On peut en tirer des correspondances bilingues, que nous noterons de la manière suivante :

*Ce || This, Chaque || Each*

La commutation d'unités polylexicales s'effectue en plusieurs temps, par transitivité (si A et B commutent ensemble, et B et C commutent ensemble, alors A, B et C forment une unité).

> *Fr. : Ce projet de loi **prévoira / entérinera** un système de déclaration des maladies infectieuses*
>
> *Angl. : This bill will **provide for / confirm** an infectious disease notification system*

d'où la commutation : *prévoira || provide for* (1)

> *Fr. : Ce projet de loi **prévoira / prévoit** un système de déclaration des maladies infectieuses*
>
> *Angl. : This bill **will provide / provides** for an infectious disease notification system*

par conséquent, on a : *prévoira || will provide* (2)

Par transitivité *will provide for* est repéré comme une seule unité :

(1) + (2) ⇒ *prévoira || will provide for*

Mais notons que ce test connaît des limites, surtout dans les cas de traductions moins « littérales », où l'impossibilité d'établir des correspondances d'unité à unité le rend inapplicable. Implicitement, pour que le test soit envisageable, les phrases source et cible doivent avoir le même *contenu sémantique*. En effet, la commutation des unités est censée suivre les deux phases déjà décrites : création d'une différence dans la source et rétablissement de l'identité sémantique par création de la même différence dans la cible. Mais lorsqu'il n'y a pas exactement identité sémantique au départ la possibilité de la double commutation devient caduque. En effet, même si l'on rétablit l'identité sémantique en commutant, les unités de départ resteront réfractaires aux correspondances déduites. Prenons l'exemple suivant, et cherchons à appliquer la commutation, en prenant soin, à chaque fois, de rétablir au mieux l'identité sémantique.

> *Fr. :* (…) *l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite*
> *Angl. :* (…) *the marking of banknotes for the benefit of the blind and partially sighted*

Le procédé de traduction (qui va ici de l'anglais vers le français) correspond à ce que Vinay et Darbelnet (1958) nomment modulation. La difficulté à trouver un équivalent pour

*marking* conduit le traducteur à une stratégie d'évitement, qu'il réalise grâce à un élargissement métonymique (le « marquage » des billets étant une partie de leur « émission »), contrebalancé ensuite par un rétrécissement sémantique, *for the benefit* étant rendu par *identifiable*. Il en résulte un schéma de commutation plus complexe :

> *Fr. :* (…) ***l'émission*** / ***la destruction*** *de*/***s*** *billets de banque identifiables par les aveugles et par les personnes à vision réduite*
> *Angl.* : (…) *the* ***marking*** / ***destruction*** *of banknotes* ***for the benefit of*** / ***identifiable by*** *the blind and partially sighted*

> *Fr. :* (…) *l'émission de billets de banque* ***identifiables*** / ***inutilisables*** *par les aveugles et par les personnes à vision réduite*
> Angl. : (…) *the* ***marking*** / ***issue*** *of banknotes* ***for the benefit of*** / ***useless for*** *the blind and partially sighted*

On obtient, par l'application de la transitivité :

*l'émission ...identifiables || marking ... for the benefit of*

Que signifie cette correspondance ? en dehors du contexte précis de ces deux phrases, rien. Cette absence de correspondant clair peut être caractérisée par les possibilités importantes de commutation sans contrepartie : *émission* peut commuter avec *création*, *fabrication*, *impression*, *tirage*, *production*, *diffusion*, *introduction* sans que la relation d'équivalence avec la phrase anglaise en soit altérée. De même *identifiables* peut commuter avec *utilisables*, *reconnaissables*, *lisibles*, *manipulables*, *déchiffrables*, etc. Ces possibilités de commutation « à vide » dénotent le lien très lâche de ces unités avec la phrase cible.

Les observations de Seleskovitch sur la prise de note en traduction consécutive révèlent deux types de comportements lexicaux : certains mots fusionnent et perdent leur identité au sein du produit final, d'autres subsistent et gardent leur identité formelle (Seleskovitch compare ces derniers à des raisins dans une brioche, qui résistent à la cuisson) :

> « En étudiant non seulement l'interprétation proposée par ses collègues mais également les notes de consécutive qu'ils avaient prises, Seleskovitch constate que certains mots du discours original sont notés et traduits par les participants. Ce sont les chiffres, les appellations, les énumérations et les termes techniques. Par contre d'autres mots, qui possèdent ce qu'elle avait appelé dans *L'interprète dans les conférences internationales* des équivalents conventionnels dans l'autre langue, n'avaient été ni notés ni traduits tels quels. Fondus dans l'opération de chimie du sens, ils avaient fait l'objet d'une réexpression. » (Laplace, 1994 : 239)

Ce qui est vrai pour l'interprétation consécutive l'est aussi pour la traduction écrite, car toute traduction implique une interprétation globale de l'équivalence des énoncés. Il faut donc admettre que la commutation interlingue, et par voie de conséquence le repérage de traduction, ne peut pas concerner toutes les unités du texte, mais seulement un sous-ensemble : l'hypothèse de compositionnalité traductionnelle ne s'applique pas complètement au niveau lexical. C'est pourquoi nous préférons parler de *correspondances lexicales* plutôt que d'alignement lexical.

### 2.4.3 Extraction de correspondances lexicales

L'extraction de correspondances lexicales est en quelque sorte une extension automatisée du test. De nombreux travaux (Gaussier et Langé, 1995, Chang et Ker, 1996, McEnery et Oakes, 1996, Melamed, 1997a, Fung, 2000, Och & Ney 2003, Kraif 2004,) ont montré qu'il est possible d'extraire des appariements lexicaux – et par suite des lexiques bilingues – à partir de l'observation des occurrences et des cooccurrences au sein d'un bi-texte. Toutes les méthodes ainsi développées se basent sur une idée simple : des unités source et cible qui apparaissent très fréquemment dans des segments équivalents (c'est-à-dire plus souvent que le hasard ne le laisserait escompter), sont vraisemblablement équivalentes.

(…*u*…, … …)
(…*u*…, …*u'*…)
(… …, … …)
(…*u*…, …*u'*…)
(… …,… …)
(… …, …*u'*…)
(…*u*…, … …)
(…*u*…, …*u'*…)

*Figure 2.4 : Occurrences et cooccurrences de deux unités (n1= 5, n2=4, n12=3)*

Dans l'exemple de la figure 2.4, on compte 5 occurrences de l'unité *u*, 4 cooccurrences de l'unité *u'* et 3 cooccurrences. En fonction des occurrences, on peut estimer le nombre de cooccurrences qu'on obtiendrait dans le cas d'une distribution aléatoire (8*(5/8)*(4/8) = 2,5). Si le nombre de cooccurrences observées dépasse de manière significative cette estimation, on peut alors faire l'hypothèse que les unités sont des équivalents traductionnels.

Or, il apparaît que ce type d'observation n'est rien d'autre qu'une extension du test de commutation, mais en négatif : on s'appuie sur le nombre de fois que les contextes des unités commutent, quand les unités apparaissent ensemble, rapporté au nombre de fois où, dans des

contextes équivalents, les unités apparaissent séparément. Comme dans le test de commutation classique, ce sont les variations concomitantes qui permettent de dessiner l'organisation des unités à travers le jeu des identités et des différences.

Plusieurs indices statistiques permettent de chiffrer la vraisemblance de cette hypothèse : l'information mutuelle spécifique (Church 1990), le *t-score* (Fung et Church, 1994), le rapport de vraisemblance (Dunning, 1993) et la log-probabilité de l'hypothèse nulle (Kraif, 2004). Dans ces derniers travaux, nous avons mis en œuvre ces différents indices sur des unités lexicales manuellement tokenisées et lemmatisées. Les valeurs d'occurrences et de cooccurrences ont été calculées à partir des phrases du JOC (automatiquement alignées par nous). Pour chaque couple de phrases, nous avons appliqué l'algorithme de meilleure affectation biunivoque (noté ABIJ) : 1/ calcul de l'indice d'association pour tous les appariements possibles d'unités ; 2/ sélection et enregistrement du couple d'unités obtenant le meilleur indice ; 3/ élimination, dans l'ensemble des couples candidats, de tous les couples concurrents du couple sélectionné (i. e. qui mettent en jeu une des deux unités sélectionnées) ; 4/ tant qu'il reste des candidats, retour en 2. Les résultats obtenus ont été évalués sur un corpus de référence d'environ 700 couples de phrases alignés manuellement (aléatoirement tirées du corpus JOC). Les jeux d'appariements obtenus automatiquement ont été comparés avec les couples de référence, en calculant la précision P (nombre de couples corrects/nombre de couples extraits), le rappel R (nombre de couples corrects/nombre de couples de référence) et la F-mesure[12] pour synthétiser P et R.

Les valeurs de F-mesure des extractions réalisées avec ces différents indices sont représentées figure 2.5.

Avec PC, indice basé sur les cooccurrences et l'identification des cognats, nous avons obtenu des résultats très satisfaisants (F = 78,5 %). Notons que la seule observation des cooccurrences permet d'obtenir presque aussi bien (F = 77,2 %).

À la différence du test de commutation interlingue, les variations entre les mots qui commutent et leurs contextes ne sont plus appréhendées dans le cadre d'une série d'observations manuelles, reposant sur l'interprétation. Le filtrage de la masse des occurrences et des cooccurrences révèle des *régularités* et non des *règles*. Les corrélations

[12] Il s'agit de la moyenne harmonique $F = \frac{2 \times P \times R}{(P + R)}$

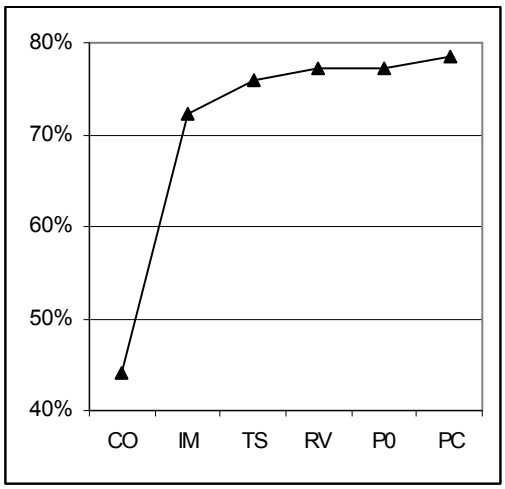


*Figure 2.5 : F-mesure des extractions de correspondances lexicales. CO : indice basé sur la cognation (mots apparentés), IM : information mutuelle spécifique, TS : T-score, RV : rapport de vraisemblance, P0 : log-probabilité de l'hypothèse nulle, et PC : combinaison de CO et P0.*

étudiées entre les deux plans parallèles, c'est-à-dire les deux idiomes confrontés dans la relation de traduction, ne relève pas d'une loi du tout ou rien, comme dans la commutation phonologique de *vache* avec *tache*. Ce qui nous intéresse dans cette masse de commutation, c'est qu'elle prend une forme à mesure qu'elle croît, qu'elle exhibe des régularités qui ne peuvent être imputables au hasard, ni aux choix individuels du traducteur, ni aux contingences de la situation de communication. Au-dessus des aléas traductionnels, ces régularités révèlent les points de contact entre les codes : elles filtrent finalement ce qui dans la traduction ressortit au *transcodage*, que Seleskovitch oppose à la traduction interprétative[13].

Il existe une manière objective de quantifier cette propriété des multi-textes. Si l'on compare un jeu de correspondances lexicales manuellement extraites avec un jeu d'appariements tirés au hasard à l'intérieur de phrases alignées, une différence formelle apparaît immédiatement : les couples corrects présentent beaucoup plus de répétitions, d'« ordre », que les couples pris aléatoirement. Par exemple, si l'on examine les 10 occurrences de *against* dans notre corpus de référence, on dénombre seulement 3 paires différentes, tandis qu'avec un tirage aléatoire des appariements on en a obtenu 10, comme le montre le tableau 2.2.

13 Pour Seleskovitch, le transcodage est une opération de transfert mécanique de code à code, qui ne requiert pas l'interprétation du texte, à la différence de la traduction interprétative : « Le transcodage, applicable à certains éléments des textes, est important en traduction, il n'est pas *la* traduction. » (Seleskovitch, citée par Laplace, 1994 : 240).

| *Correspondances extraites manuellement* | *Correspondances extraites aléatoirement* |
|---|---|
| (against, à l'encontre de) | (against, par) |
| (against, à l'encontre de) | (against, procédure) |
| (against, à l'encontre de) | (against, moratoire) |
| (against, au détriment de) | (against, à l'encontre de) |
| (against, contre) | (against, dont) |
| (against, contre) | (against, contre) |
| (against, contre) | (against, effectivement) |
| (against, contre) | (against, charges) |
| (against, contre) | (against, Etat membre) |
| (against, contre) | (against, qui) |

*Tableau 2.2 : Correspondances lexicales correctes vs aléatoires*

Pour quantifier ce type de dispersion, nous proposons de calculer l'entropie conditionnelle, qui mesure le « désordre » des cooccurrences de deux unités source et cible, par rapport aux occurrences de l'une ou de l'autre. Les équations [1] et [2] donnent l'expression de l'entropie conditionnelle dans les deux sens de la traduction :

$$H(T'/T) = -\sum_{u} p(u) \sum_{u'} p(u'/u) \log p(u'/u) = -\sum_{u} \sum_{u'} p(u,u') \log \frac{p(u,u')}{p(u)} \quad [1]$$

$$H(T/T') = -\sum_{u'} p(u') \sum_{u} p(u/u') \log p(u/u') = -\sum_{u'} \sum_{u} p(u,u') \log \frac{p(u,u')}{p(u')} \quad [2]$$

où $T$ et $T'$ sont respectivement les textes source et cible, $u$ et $u'$ des unités de $T$ et $T'$, $p(u)$ représente la probabilité d'apparition de $u$ à gauche d'un couple d'unités appariées, $p(u')$ la probabilité d'apparition de $u'$ à droite d'un couple d'unités appariées, et $p(u,u')$ la probabilité de l'appariement $(u,u')$. Afin d'étudier la corrélation entre cette quantité et la correction des résultats, nous avons évalué les valeurs d'entropie pour différentes séries d'extractions de correspondances comportant différentes proportions d'erreurs. On a ainsi obtenu 6 séries d'extractions (pour plus de détail, cf. Kraif, 2003b) :

- Les appariements de référence extraits manuellement ;
- 6 extractions (pour les indices CO, TS, IM, RV, P0, PC), avec l'algorithme d'association maximale AMAX[14] ;
- 6 extractions (pour les indices CO, TS, IM, RV, P0, PC), avec l'algorithme ABIJ (analogue à celui décrit par Melamed, 1997) ;

[14] A la différence de ABIJ, avec AMAX, pour chaque unité source, on sélectionne l'appariement qui a obtenu la meilleure valeur de l'indice. Une même unité cible peut donc apparaître dans plusieurs couples. Ainsi AMAX est dissymétrique vis-à-vis des deux textes.

– 7 extractions obtenues avec 7 pondérations différentes d'un indice combinant P0 à une valeur aléatoire[15].

– 6 extractions filtrées[16] (pour les indices CO, TS, IM, RV, P0, PC), avec AMAX ;

– 6 extractions filtrées (pour les indices CO, TS, IM, RV, P0, PC), avec ABIJ.

La figure 2.6 montre une étroite corrélation entre l'entropie conditionnelle[17] et la précision de chaque jeu de correspondances. Le coefficient de corrélation linéaire est en effet de -0,96. Malgré les choix de traduction particuliers, il existe bien des régularités quantifiables. La variabilité traductionnelle constitue un « bruit » au-dessus duquel émergent les structurations des codes. Au-delà des effets de *sens*, se dessinent les *significations*.

Le repérage de traduction comporte donc deux faces : une face subjective, en tant qu'il nécessite l'interprétation d'un sujet pour relier des unités équivalentes d'un point de vue traductionnel dans un contexte particulier ; et une face objective, en tant que certaines correspondances manifestent des régularités (correspondant à un minimum d'entropie) que l'on peut extraire automatiquement avec des méthodes fiables.

Nous explorerons dans la partie 3 de cette synthèse quelques implications de l'extraction de correspondances lexicales, notamment sur les plans de la linguistique contrastive et de la lexicographie.

---

15 Soit l'indice $AL=(1-Coeff)\times P0+Coeff\times Random$ , où *Random* est une valeur aléatoire comprise entre 0 et 10, et Coeff prend les valeurs respectives : °0,25 ; 0,5 ; 0,75 ; 0,95 ; 0,97 ; 0,99 ; 1.

16 Les extractions filtrées ne retiennent que les couples ayant obtenu un indice au moins deux fois supérieur à tous leurs concurrents. Elles présentent en général une précision supérieure pour un rappel dégradé.

17 Pour chaque extraction, nous avons pris *min(H(T/T'),H(T'/T))*, c'est-à-dire qu'à chaque fois, nous avons favorisé le sens de traduction où les régularités apparaissaient avec le plus de force.

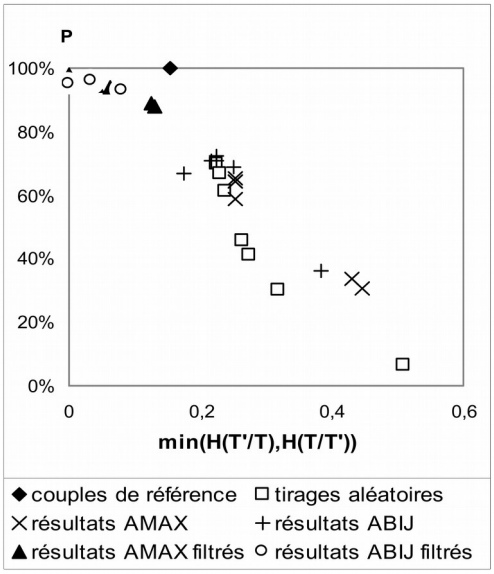


*Figure 2.6 : Corrélation entre la précision des extractions et leur entropie conditionnelle*

## 2.5. L'alignement de corpus multi-parallèles

Concernant la problématique générale de l'alignement, il existe une piste qui a encore été assez peu explorée : celle du *multi-alignement*, à savoir l'alignement de plus de deux langues. Dans le cadre du projet Carmel (Chen *et al.*, 2005, Kraif *et al.* 2006), nous avons travaillé sur des corpus non pas seulement parallèles *bilingues*, mais parallèles *multilingues*, c'est-à-dire impliquant plus de 2 langues. Pour éviter l'ambiguïté, nous parlerons désormais dans ce cas de corpus *multi-parallèles*. Dans une étude pionnière dans le domaine de la désambiguïsation lexicale, Dagan *et al.* (1991) avaient intitulé leur article « *Two Languages Are More Informative Than One* ». Généralisant cette intuition, nous voulions démontrer, à travers le projet Carmel, que mettre en jeu plus de deux langues, dans le cadre de textes parallèles, pouvaient apporter encore plus d'information, chaque version étant susceptible d'apporter un éclairage spécifique sur les autres versions alignées, notamment dans le domaine de la désambiguïsation lexicale. Il s'agit là d'une idée qui a guidé une grande partie de nos recherches, qui s'appuie sur un principe analogue à la triangulation en sciences humaines : une hypothèse formulée dans le cadre de la comparaison de deux langues pourra être corroborée (ou infirmée) par la comparaison avec une troisième langue, voire une quatrième, etc. On peut ainsi formuler le *principe de triangulation multilingue* qui était à l'origine des hypothèses formulées dans le cadre du projet Carmel : du fait de leur similarité et de leur différence, la mise en correspondance d'énoncés équivalents dans plusieurs langues

permet de mettre en lumière des traits qui restent implicites si on considère chaque langue isolément. En outre, plus grand est le nombre de langues impliquées, plus riche est l'effet de la triangulation. Nous développerons cette idée plus loin, notamment dans le domaine de la désambiguïsation lexicale (cf. 3.4.1, p. 85).

Notons que les corpus multi-parallèles ne sont pas rares : l'*Acquis communautaire*, qui constitue le socle législatif et réglementaire de l'Union Européenne en est un des exemples les plus représentatifs : il est actuellement distribué en version 3.0, sous le nom de JRC-Acquis Corpus, et concerne 22 langues européennes[18] – pour un total d'environ 636 millions de mots toutes langues confondues. On trouve par ailleurs sur le site de l'OPUS - Open parallel corpus[19] (Tiedemann, 2012) de très nombreux corpus multi-parallèles dans des domaines variés (juridique, réglementaire, diplomatique, technique, sous-titre de films, etc.). Certains de ces corpus sont massivement multilingues, comme le corpus *OpenSubtitles2013* qui compte 59 langues et intègre la plupart des paires de langues impliquées (1 211 paires de langues sur 1 711). Pour le projet Carmel, nous avons réuni des œuvres littéraires (des récits de voyages) en 4 langues : anglais, espagnol, français, italien.

Notons enfin que pour certains corpus multi-parallèles, comme les textes réglementaires, il n'est pas toujours aisé de connaître la langue source : on considère alors la relation d'équivalence traductionnelle prise globalement, sans faire de distinction entre texte original et traductions.

### 2.5.1 Niveau de l'alignement phrastique

Sur le plan de l'alignement phrastique on peut se poser la question suivante : est-il possible d'aligner simultanément plus de deux textes ? et si c'est le cas, quel intérêt peut-on y trouver ?

Peu de réponses probantes ont été apportées à la première question, à en juger par les méthodes mises en œuvre dans la constitution des principaux corpus multi-parallèles : pour le JRC-Acquis, tous les alignements ont été effectués 2 à 2, en utilisant l'aligneur Vanilla (Danielsson & Riding, 1997), qui implémente la méthode de Gale & Church ; de même les textes du corpus OPUS ont été alignés 2 à 2 grâce à cette même méthode.

18 cf. http://ipsc.jrc.ec.europa.eu/index.php?id=198 (consulté en mai 2014)

19 cf. http://opus.lingfil.uu.se/ (consulté en mai 2014)

Concernant le deuxième point, le principal inconvénient de l'alignement 2 à 2 d'un corpus de textes multi-parallèles réside dans le grand nombre de couples à considérer. Par exemple, pour les 22 langues du JRC Acquis, il faut considérer $22\times21/2=231$ couples différents. D'un point de vue général, pour *n* langues, le nombre de couples impliqués est quadratique : $n\times(n-1)/2$ . Cette complexité peut être pénalisante à la fois du point de vue du temps de calcul et de l'espace de stockage des résultats. Quand on a 22 textes parallèles, pourquoi ne pas aligner les 22 langues en même temps, et représenter l'alignement résultant dans une seule structure de données, par exemple un seul fichier au format TMX contenant tous les groupes de phrases équivalents, plutôt que 231 fichiers différents ?

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

### 2.5.2 Cadre algorithmique pour un multi-aligneur

Les méthodes bilingues telles que celles de Gale & Church sont difficilement généralisables au cas de *n* langues, la complexité des algorithmes de programmation dynamique mis en œuvre étant exponentielle en $O(t^n)$, pour des textes de taille *t*.

Le système *trial*, dans la mesure où il implique de pré-calculer tous les alignements 2 par 2, nous semble également assez lourd sur le plan algorithmique lorsqu'un grand nombre de langues est mis en jeu. Il n'a d'ailleurs jamais été étendu au-delà de trois langues, à notre connaissance.

D'autres méthodes peu coûteuses sont envisageables, comme l'alignement par transitivité : si A est aligné avec B et B est aligné avec C, alors on peut calculer rapidement, par transitivité, un alignement entre A et C. Mais cette méthode présente des défauts importants :

– Lorsque l'on prend la clôture transitive des alignements, on obtient en général des alignements plus grossiers, ce qui aboutit à une baisse de la précision. Par exemple, supposons qu'on ait les alignements suivant : (a1 ;b1)(a2 a3 ;b2) et (b1 b2; c1 c2). On obtient alors par transitivité : (a1 a2 a3 ; c1 c2), même si l'alignement de référence est en fait (a1 ; c1) (a2 a3 ; c2). Notons que ce défaut est intrinsèque à toute méthode d'alignement multilingue produisant des alignements complets satisfaisant la propriété de clôture transitive.

– Cette méthode ne tire pas parti du principe de triangulation : tout repose sur une seule langue pivot, et si l'alignement au niveau d'un couple est faible, cette faiblesse sera propagée vers la troisième langue par le jeu de la transitivité, au lieu d'être éventuellement compensée par la prise en considération d'un autre couple plus solide.

### 2.5.3 L'aligneur MulItAl

Dans le cadre du projet Carmel, nous avons commencé à étudier cette piste, et j'ai co-développé avec Bettina Schrader, une ingénieure contractuelle engagée pour le projet, un script d'alignement vraiment multilingue, nommé *MulItAl*, pour (*Multilingual Iterative Aligner*).

L'idée de MulItAl est de se baser sur l'appariement des transfuges qui forment un réseau de points d'ancrage entre tous les textes parallèles. Lorsqu'un transfuge apparaît le même nombre de fois dans chaque texte, alors ses occurrences peuvent servir à construire des points d'ancrage. Par exemple, dans le corpus étudié (en l'occurrence les trois premiers

chapitre de *Madame Bovary* de Flaubert, en anglais, espagnol, français et italien) on observe que la chaîne *Emma* a les occurrences suivantes :

*Anglais : phrases 279 et 545*
*Espagnol : phrases 250 et 501*
*Français : phrases 273 et 539*
*Italien : phrases 268 et 524*

On peut en tirer deux points d'ancrage dans l'espace du quadri-texte EN-ES-FR-IT :

*(279, 250, 273, 268) et (545, 501, 539, 524)*

À partir de ces deux points d'ancrage, on réalise un découpage de l'espace en sections plus petites dans lesquelles on peut réitérer l'appariement des transfuges. Certains transfuges qui n'étaient pas appariables dans l'ensemble du texte, du fait d'un nombre d'occurrences différents, deviennent appariables dans des sections plus petites où leurs occurrences sont parallèles – et peuvent donc donner de nouveaux points d'ancrage. Lorsqu'on arrive à stabilité, on considère alors des sous-groupes de langues, qui peuvent apporter de nouveaux transfuges, par exemple :

*EN,ES, FR : piano*
*FR,IT : difficile*
*ES, IT : primavera*

On peut alors réitérer sur ces sous-groupes, qui apporteront des points d'ancrage partiels qui densifient le réseau de correspondance, et peuvent compléter, par transitivité, d'autres points partiels. Dans l'algorithme on considère tous les sous-groupes de langues, en traitant d'abord les plus grands, qui donnent des points d'ancrage a priori plus fiables. Par exemple, pour $N$=4 et $L$={EN,ES,FR,IT}, on considère la suite de sous-groupes suivants {EN,ES,FR,IT}, {ES,FR,IT}, {EN,ES,IT}, {EN,FR,IT}, {ES,IT}, {ES,FR}, {FR,IT} {EN,ES} {EN,IT} {EN,FR}. L'algorithme est schématisé figure 2.7 :

```
A={};
Pour (K = N ; K >=2 ; K--)
   Pour chaque S={L1,L2,…LK} sous-ensemble de L de taille K
     Pour chaque transfuge T de ID(S)
       Pour chaque couple de points (Pi,Pi+1) résultant de la suite
ordonnée des points de A définis pour les langues de S
          Si T a n occurrences occL,1, occL,2, … occL,n dans l'intervalle
[Pi,Pi+1] pour chaque langue L de S
            Pour j=1…n
              A←A U (occL1,j, occL2,j, … ,occLK,j)
            Fin Pour
          Fin Si
        Fin Pour
      Fin Pour
   Fin Pour
Fin Pour
```

*Figure 2.7 : Algorithme itératif d'appariement des transfuges*

*ID(S)* est l'ensemble des transfuges appartenant aux langues du sous-ensemble S. Par ailleurs, il faut noter que A contient des points partiels, qui ne concernent pas toutes les langues. Ainsi, l'ajout d'un nouveau point dans A implique le respect de deux conditions :

– *Transitivité* : Lorsque deux points partiels se recouvrent en partie, ils sont fusionnés. Par exemple si on a :(EN-545, ES-501, FR-539) et (EN-545, IT-524) le point résultant est (EN-545, ES-501, FR-539, IT-524)

– *Cohérence* : Si le nouveau point croise un point déjà existant, alors il n'est pas retenu. Deux points $(x_{i1},x_{i2},\ldots,x_{in})$ et $(y_{j1},y_{j2},\ldots,y_{jm})$ se croisent s'ils partagent des langues communes et s'il existe un couple de langues $L_i$, $L_j$ tel que : $x_i \geq y_i$ et $x_j < y_j$ ou bien $x_i \leq y_i$ et $x_j > y_j$.

| Langue | Occurrences | Types | Phrases |
|---|---|---|---|
| | Texte complet | | |
| Anglais | 143 004 | 10 946 | 9 565 |
| Espagnol | 137 567 | 15 567 | 8 777 |
| Français | 153 938 | 13 734 | 9 213 |
| Italien | 137 599 | 16 687 | 9 154 |
| | Chapitres 1-3 | | |
| Anglais | 10 321 | 2 285 | 546 |
| Espagnol | 9 741 | 2 675 | 503 |
| Français | 11 077 | 2 583 | 541 |
| Italien | 9 671 | 2 955 | 526 |

*Tableau 2.3 : Corpus de test de MulItAl*

MulItAl a été testé sur les trois premiers chapitres de *Madame Bovary*, pour lesquels des alignements de référence deux à deux avait été constitués manuellement. La constitution de ce corpus est décrite dans le tableau 2.3.

Les résultats de cette première expérimentation ont été plutôt décevants, sous l'angle du rappel, comme le montre le tableau 2.4 (résultats avant filtrage).

| | Avant filtrage | | Après filtrage | |
|---|---|---|---|---|
| Couple de langues | *P* | *R* | *P* | *R* |
| EN-FR | 0,92 | 0,38 | 0,99 | 0,29 |
| FR-IT | 0,86 | 0,40 | 0,97 | 0,26 |
| IT-EN | 0,86 | 0,29 | 0,97 | 0,21 |
| EN-ES | 0,91 | 0,22 | 1 | 0,17 |
| FR-ES | 0,84 | 0,32 | 0,97 | 0,21 |
| IT-ES | 0,92 | 0,26 | 0,99 | 0,25 |

*Tableau 2.4 : Résultats de MulItAl sur le corpus Bovary*

En analysant les erreurs produites, nous avons noté que la plupart des mauvais points sont dus à des « faux-amis » tel que *fine* (EN) vs *fine* (IT) ou *habit* (EN) vs *habit* (FR) ainsi que des mots fonctionnels tels que *con* (IT,ES) ou *del* (IT,ES). Dans une deuxième version, nous avons écarté ces transfuges peu fiables par l'ajout de deux critères :

- Les transfuges courts (de 3 caractères ou moins) et fréquents (f≥50).
- Les transfuges dont les fréquences sont éloignées (dont le rapport est inférieur à 1/2) dans les différentes versions.

On obtient alors une excellente précision, mais le rappel est assez faible. En l'état, la méthode ne permettait pas d'obtenir une amélioration quelconque par rapport aux alignements deux à deux : cette piste avait donc été abandonnée dans le Projet Carmel.

### 2.5.4 Cognats et multi-alignement

En rédigeant cette synthèse, j'ai néanmoins tenu à poursuivre cette idée en réalisant des expériences complémentaires. Pour des corpus multi-parallèles tels que ceux de l'Union Européenne, il apparaît que la parenté linguistique entre les différents groupes de langues impliqués (langues romanes, langues germaniques, langues slaves, langues baltes, langues finno-ougriennes, pour ne citer que les principaux groupes) doit pouvoir jouer un rôle prépondérant dans le multi-alignement : se contenter d'identifier les transfuges (souvent des nombres ou des noms propres) ne peut donc suffire à tirer profit de cette richesse.

Afin d'explorer cette hypothèse, j'ai téléchargé la transcription de la session du 17 janvier 2000 du parlement européen, tiré du corpus Europarl3[20], qui contient 11 versions alignées dans les langues suivantes : allemand, anglais, danois, espagnol, français, finnois, grec, italien, portugais, néerlandais, suédois (on utilisera désormais les codes ISO, par ordre alphabétique : DA, DE, EL, EN, ES, FI, FR, IT, NL, PT, SV). J'ai manuellement révisé les alignements fournis pour tous les couples impliquant le français afin d'avoir une référence fiable (la plupart des alignements fournis étaient de bonne qualité à part pour le couple fr-nl qui a nécessité un peu plus de révisions).

Notre première tâche a consisté à mesurer le degré de proximité graphique des formes alignées entre toutes les langues prises deux à deux, afin d'évaluer jusqu'à quel point la parenté génétique peut se traduire en un critère automatiquement exploitable (l'identification des candidats cognats).

Pour chaque couple de phrases, nous avons compté les candidats cognats en retenant toutes les paires de mots d'au moins 7 caractères pour laquelle la SCM (cf p. 22) correspond à au moins 80 % des caractères de la chaîne la plus courte des deux chaînes comparées. Avec de tels critères, plutôt sélectifs, on trouve de très nombreux cognats avec un minimum de bruit. Par exemple, pour les langues da, de, en, on trouve les paires suivantes :

*Integration↔integration, explizit↔explicitly,*
*periodiske↔Periodischen, Schroedter↔Schroedterin,*
*diskussion↔Diskussion, programmer↔Programme,*
*Transport↔transport, transport↔Transport,*
*Parlaments↔Parliament, Regionalpolitik↔Regional,*
*regionaler↔regional, Europa-Parlamentets↔Europaparlamentets,*
*Europæiske↔Europäischen, Kommission↔Commission*

Lors de cette comparaison, nous avons traité le grec séparément, parce qu'il utilise un alphabet différent. Les résultats, cumulant le nombre de transfuges (hormis les nombres et les noms commençant par une majuscule) et le nombre de cognats identifiés avec les critères précédents, sont présentés dans le tableau 2.5 :

20 http://opus.lingfil.uu.se/Europarl3.php. Le débat est accessible directement ici : http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+CRE+20000117+ITEMS+DOC+XML+V0//FR&language=FR#creitem2 (consulté en mai 2014)

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

*Tableau 2.5 : Nombre de transfuges et cognats identifiés dans les bi-phrases par couples de langues*

Pour le grec, nous avons effectué la même comparaison en utilisant une translittération standard[21] (on utilisera désormais le code *GR* pour le grec translittéré, plutôt que *EL,* et on traitera ce texte comme une version à part entière, afin d'évaluer l'impact de la translittération). Celle-ci (cf. tableau 2.6), montre sans surprise un nombre de cognats beaucoup plus réduit :

| | DA | DE | EN | ES | FI | FR | IT | NL | PT | SV | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **GR** | 229 | 116 | 245 | 434 | 125 | 188 | 224 | 184 | 183 | 231 | **2 159** |

*Tableau 2.6 : Nombre de transfuges et cognats avec le texte grec translittéré*

Quand on considère les valeurs marginales, on constate que certaines langues cumulent beaucoup plus de cognats que d'autres : elles occupent en quelque sorte une position plus centrale au sein de ces différentes familles linguistiques, position qui leur confère en moyenne une plus grande ressemblance avec un plus grand nombre de langues – c'est notamment le cas du français et de l'anglais.

Pour mieux s'en rendre compte, il est possible d'adopter des représentations topologiques permettant de synthétiser ces phénomènes de proximité, d'éloignement et de centralité.

21 selon la norme ISO-843, cf. http://en.wikipedia.org/wiki/ISO_843 (consulté en mai 2014)

Dans un premier temps, nous pouvons construire une visualisation par échelonnement multidimensionnel (en anglais MDS, pour *Multi Dimensional Scaling*), une technique d'analyse multivariée permettant d'afficher en deux dimensions un ensemble de points définis dans un espace de dimension *n*, en conservant au mieux les distances entre les points. Pour effectuer le MDS, il faut partir d'une matrice de distance (et non de similarité comme c'est le cas dans le tableau 2.5). Pour ce faire, nous avons utilisé les outils de l'environnement 'R'[22], un projet libre réunissant de très nombreux outils pour le calcul statistique et l'analyse de donnée. Nous avons d'abord calculé, au moyen de la fonction *dist()* de R, les distances euclidiennes entre les vecteurs définis par le tableau 2.5[23]. Nous avons ensuite appliqué la fonction *isoMDS()* sur cette matrice de distance, afin d'avoir une représentation en 2 dimensions (cf. figure 2.8).

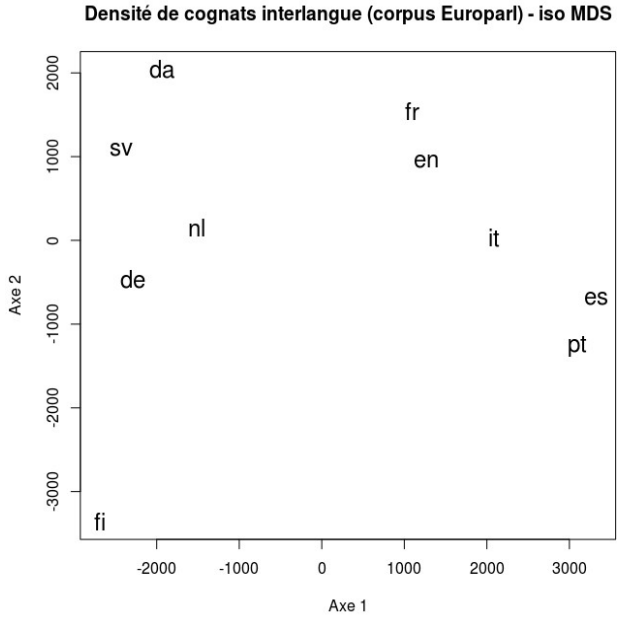


*Figure 2.8 : Réduction dans un espace à 2 dimensions des points définis dans le tableau 2.5*

Le degré d'adéquation entre les distances des points projetés sur un plan et les distances initiales dans l'espace de dimension *n* est ici mesuré par la fonction de *stress* de Kruskal (1964). On obtient ici un stress d'environ 11,656, ce qui est considéré comme bon (ibid., p. 3).

22 cf. http://www.r-project.org/ (consulté en mai 2014).

23 Pour obtenir une matrice de distance significative, il nous a fallu indiquer une valeur de similarité non nulle pour une langue avec elle-même : en d'autres termes, nous avons rempli la diagonale du tableau 2.5 en utilisant une similarité maximale arbitraire de 4 000 (mais en conservant des valeurs nulles, on obtient à peu près les mêmes résultats finaux - cette valeur n'a donc pas d'impact sur l'interprétation).

Les différentes familles linguistiques concernées apparaissent très clairement sur ce graphique : du côté droit les langues romanes avec le portugais, l'italien et l'espagnol, puis le français qui apparaît un peu décalé, peut-être du fait de sa très forte proximité graphique avec l'anglais. Sur la droite, l'allemand et le néerlandais représentent la branche occidentale des langues germaniques, tandis qu'un peu au-dessus le suédois et le danois représentent la branche nordique de cette même famille. L'anglais, du fait de son grand stock lexical emprunté au français, se situe en position assez centrale, tout à côté de ce dernier, dans une position charnière entre langues romanes et langues germaniques. Le finnois, seul représentant de la famille finno-ougrienne et par conséquent seule langue non indo-européenne, apparaît naturellement comme la plus éloignée de toutes les autres, dans le coin inférieur gauche.

En ce qui concerne la position du français, il faut peut-être considérer l'existence d'un léger biais, lié au fait que lors de cette session parlementaire, le français semble avoir été la langue source la plus utilisée. Chaque orateur s'exprimant dans sa langue, nous avons compté le nombre de mots pour chaque langue source de notre petit corpus :

| DE | EL | EN | ES | FI | FR | IT | NL | PT | SV |
|---|---|---|---|---|---|---|---|---|---|
| 4 578 | 843 | 3 381 | 4 502 | 2 108 | **15 253** | 598 | 1 645 | 741 | 446 |

*Tableau 2.7 : Répartition des langues sources dans le corpus Europarl-00-01-17*

La surreprésentation du français est due aux nombreuses interventions de la présidente du parlement d'alors, Nicole Fontaine. Cela explique peut-être que le français obtient une assez bonne similarité avec la plupart des langues, même parmi les langues germaniques.

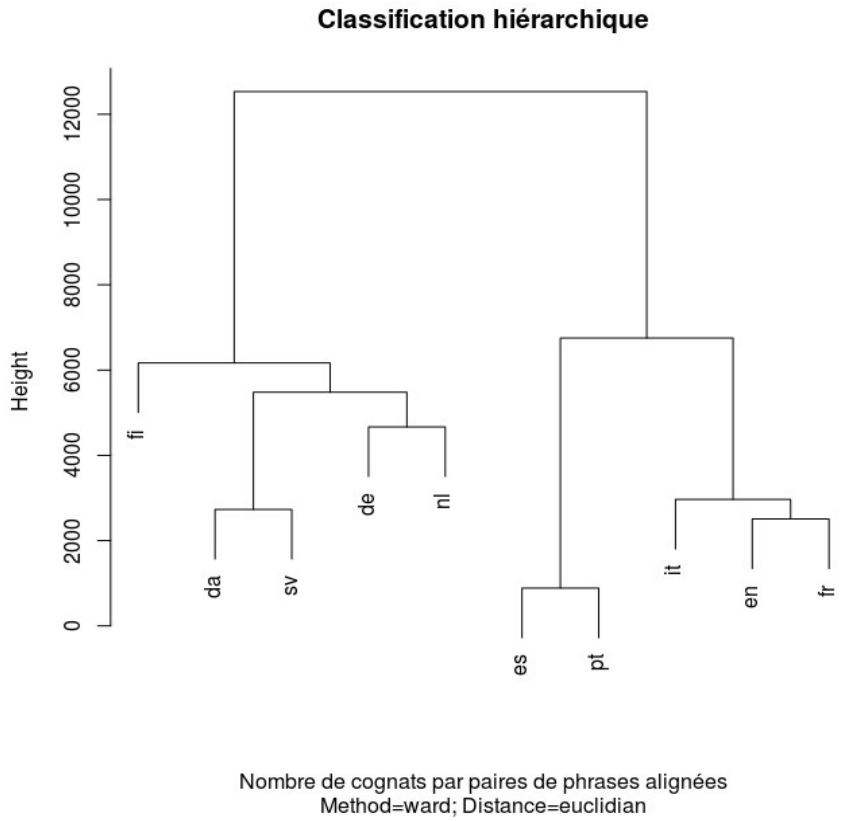


*Figure 2.9 : Classification hiérarchique ascendante - méthode Ward*

Les techniques de classification hiérarchique fournissent un autre type de représentation utile pour synthétiser ces relations de proximités. La figure 2.9 montre les résultats obtenus avec la fonction *hclust()* de 'R' . Dans ce type d'arbre, la hauteur du nœud regroupant une classe est inversement proportionnelle à la proximité des points à l'intérieur d'une classe (ici on voit que les points les plus proches sont *ES* et *PT*).

Bien entendu, pour en tirer des conclusions d'ordre génétique, il faudrait effectuer des comparaisons sur des corpus bien plus vastes – on pourrait alors apporter des données intéressantes pour la méthodologie – par ailleurs très controversée pour ses approximations – de la *mass comparison* défendue par Joseph Greenberg (Ruhlen, 1994). Mais notre objectif est plus pragmatique, et plus modeste : nous cherchons seulement à déterminer s'il est possible de s'appuyer sur le réseau très dense des mots apparentés pour tisser un multi-alignement robuste.

### 2.5.5 L'aligneur JAM

Nous avons ainsi développé une nouvelle version de MulItAl, nommée JAM (Just A Multi-aligner). L'algorithme est identique, à part que toutes les combinaisons de langues ne

sont pas utilisées : on peut se limiter à un sous-ensemble des combinaisons jugé optimal, en fonction des parentés linguistiques (sous-ensemble que nous essaierons de déterminer plus loin). Par ailleurs, à chaque fois qu'un point d'ancrage est créé, on cherche dans le couple de phrases ainsi aligné les éventuelles paires de candidats cognats (selon les mêmes critères que ceux utilisés pour le tableau 2.5). Chaque paire de cognats est alors réinjectée dans le processus, de la même manière que les transfuges. Et tout comme les transfuges, la relation entre cognats est transitive : si on trouve d'abord *explizit*↔*explicitly* puis *explicitly*↔*explicitement* alors ces trois formes seront regroupées sous le même identifiant de cognat.

Enfin, à l'issue de ces itérations, on effectue une étape de complétion : pour tous les points successifs $P_i$ et $P_{i+1}$, on examine la compatibilité des longueurs des intervalles[24] correspondant à chaque couple de langues prises 2 à 2. Si tous les intervalles sont compatibles, on construit les points suivant $P_i$ et précédant $P_{i+1}$, de proche en proche, par simple incrémentation et décrémentation des coordonnées (par exemple, pour le point (EN-545, ES-501, FR-539), on construit (EN-546, ES-502, FR-540), (EN-547, ES-503, FR-541), etc.). On procède ainsi tant que les nouveaux points sont jugés *équilibrés*[25], et *cohérents* (i.e. sans croisement avec des points existants, cf. critères p. 43). L'algorithme complet de JAM est décrit dans la figure ci -dessous :

---

```
A←{};
Comb←{ensemble des combinaisons optimales de langues}

Pour chaque combinaison de langues S={L1,L2,…,LK} de l'ensemble Comb
   # 1 - itérations
   Pour chaque cognat ou transfuge C de CO(S)
      Pour chaque couple de points (Pi,Pi+1) résultant de la suite ordonnée des points
de A définis pour les langues de S
      Si C a n occurrences correspondant aux phrases occL,1, occL,2, … occL,n dans
l'intervalle [Pi,Pi+1] pour chaque langue L de S
         Pour j=1…n
            A←A U (occL1,j, occL2,j, … ,occLK,j)
            Pour toutes les paires de mots (MLx,j MLy,j) des phrases occLx,j, occLy,j
alignées du nouveau point
               Si longueur(MLx,j) > 6 et longueur(MLy,j) > 6 et SCM(MLx,j MLy,j)
>=0.8*min(longueur(MLx,j),longueur(MLy,j))
                  associer le même identifiant de cognat à MLx,j et à MLy,j
               Fin Si
         Fin Pour
```

---

24 La longueur d'un intervalle entre deux phrases est une longueur relative. Elle est calculée en nombre de caractères, et divisée par la taille totale de chaque texte. On évalue la « compatibilité » de deux intervalles $I_1$ et $I_2$ en appliquant la condition suivante : $2\times(I_1-I_2)/(I_1+I_2)<MaxDiffInterval$ avec $MaxDiffInterval$=0,1

25 Par point « équilibré », nous entendons que toutes les longueurs des phrases, prises deux à deux, sont « compatibles », au sens de la note précédente.

```
      Fin Si
   Fin Pour

   # 2 - complétion
   Pour chaque couple de points successifs (Pi,Pi+1) obtenus pour S={L1,L2,…LK}
      Si pour tous les couples de langues (Li,Lj) de S, les longueurs des intervalles
ILi,ILj sont compatibles
         NouveauPoint←incrémentation(Pi)
         Tant que NouveauPoint est équilibré, et sans conflit ni recouvrement avec
un point existant
            A←A U NouveauPoint
            NouveauPoint←incrémentation(NouveauPoint)
         Fin tant que

         NouveauPoint←décrémentation(Pi+1)
         Tant que NouveauPoint est équilibré, et sans conflit ni recouvrement avec
un point existant
            A←A U NouveauPoint
            NouveauPoint←décrémentation(NouveauPoint)
         Fin tant que

      Fin si
   Fin Pour
```

*Figure 2.10 : Algorithme itératif d'appariement des transfuges*

Afin de garantir le maximum de précision, aux deux critères précédemment mis en œuvre dans MulItAl (*transitivité* et *cohérence* des points, cf. p. 43), nous en avons ajouté deux nouveaux :

– *redondance* : dans un premier cycle d'itérations, on ne tient compte que des points contenant au moins *minMatchNumber* appariements de cognats (ou transfuges). Après stabilité, on réitère en décrémentant cette valeur. Dans les résultats qui suivent, on a testé *minMatchNumber*=2 puis 1.

– *parallélisme* : à chaque ajout d'un nouveau point $P$, on considère les deux points existants $P_{inf}$ et $P_{sup}$ qui encadrent ce point (pour les langues considérées dans ce point). On peut alors calculer la longueur des intervalles entre $P_{inf}$ et $P$ (nous notons $Inf_L$) et entre $P$ et $P_{sup}$ (nous notons $Sup_L$) pour chaque langue $L$. La vérification de parallélisme se fait alors langue par langue, en examinant la compatibilité[26] entre les intervalles, en deux temps :

[26] On utilise un autre seuil, noté *MaxDiffInterval2*=0,1, qui peut être relevé jusqu'à 0,5 pour des textes présentant des ruptures de parallélisme, cf. p. 62

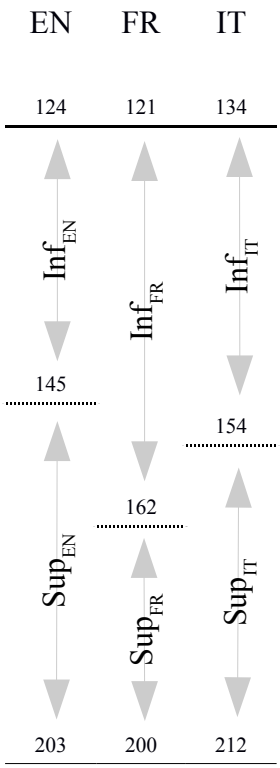


*Figure 2.11 : vérification de parallélisme à 3 langues*

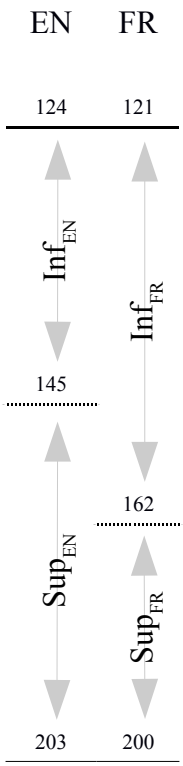


*Figure 2.12 : vérification de parallélisme à 2 langues*

1. Triangulation. D'abord, pour une langue donnée *L*, on examine s'il existe deux autres langues *L1* et *L2* avec des intervalles $Inf_{L1}$, $Inf_{L2}$, $Sup_{L1}$, $Sup_{L2}$ compatibles. Si c'est le cas, alors il faut nécessairement que $Inf_L$ soit compatible avec $Inf_{L1}$ et $Inf_{L2}$, ou que $Sup_L$ soit compatible avec $Sup_{L1}$ et $Sup_{L2}$. À l'issue de ce test, la coordonnée du point *P* pour la langue *L* est soit validée, soit supprimée (ce qui correspond à l'exemple de la figure 2.11). Ce premier test est en quelque sorte une épreuve de triangulation : une coordonnée dans une langue qui est corroborée par deux autres langues est validée – et à l'inverse, si elle est contredite par deux autres langues, elle est rejetée.

2. Parallélisme simple. Ensuite, pour toutes les langues qui ne sont ni validées ni rejetées, on poursuit le test deux à deux, comme dans la figure 2.12.

Nous avons effectué un premier test en utilisant un jeu de combinaisons de langues simple, prenant le français comme pivot (on notera FR-pivot) : $Comb_{FR\text{-}pivot}$={EN-FR, FR-IT, ES-FR, FR-PT, DA-FR, FR-NL, FR-SV, DE-FR, FR-FI, FR-GR, FR-EL}. On obtient une précision excellente et un rappel supérieur à celui obtenu avec MulItAl (mais sur un autre corpus).

| | Sans complétion finale | | Avec complétion finale | |
|---|---|---|---|---|
| Couple de langues | *P* | *R* | *P* | *R* |
| DA-FR | 0,98 | 0,46 | 0,97 | 0,81 |
| DE-FR | 0,99 | 0,59 | 0,96 | 0,82 |
| EL-FR | 0,98 | 0,38 | 0,98 | 0,83 |
| EN-FR | 0,99 | 0,70 | 0,99 | 0,88 |
| ES-FR | 0,99 | 0,68 | 0,98 | 0,86 |
| FI-FR | 1,00 | 0,52 | 0,99 | 0,92 |
| FR-GR | 0,98 | 0,44 | 0,98 | 0,84 |
| FR-IT | 0,99 | 0,65 | 0,98 | 0,84 |
| FR-NL | 0,98 | 0,41 | 0,96 | 0,67 |
| FR-PT | 1,00 | 0,71 | 0,98 | 0,90 |
| FR-SV | 1,00 | 0,51 | 0,98 | 0,80 |

*Tableau 2.8 : Résultats de JAM pour les combinaisons FR-pivot*

En examinant les points obtenus à ce stade, on obtient encore de nombreux « trous », comme en témoigne l'exemple du tableau 2.9 : l'alignement n'est pas complet car certaines langues se trouvent isolées, comme DA, EL, NL ou SV. Pour tenter d'éliminer ces trous, nous appliquons alors un algorithme de complétion finale : pour chaque couple de coordonnées ($P_{LI}$,$P_{LJ}$) non vide de chaque point *P*, on calcule l'espacement – en nombre de phrases – avec les coordonnées ($PS_{LI}$,$PS_{LJ}$) non vides du point suivant *PS*. Si celui est espacé de deux phrases ou plus, on lance l'algorithme d'alignement de Gale & Church (1991) entre *P* et *PS* pour ces deux langues[27]. Notons que cet algorithme livre des appariements groupés de type 1:2, 2:1, 2:2 tandis alors que notre multi-alignement n'enregistre que des correspondances 1:1 sans fusion ni croisement (d'après les critères de *transitivité* et *cohérence* des points, cf. p. 43). Dans ces cas de figure seul la première phrase du groupe est prise en considération.

[27] On utilise les mêmes paramètres que Gale & Church (1991), avec une légère adaptation car on ajoute les transitions 1:3 et 3:1. On a : P(1:1)=0,89, P(1:0)=P(0:1)=0,0099, P(1:2)=P(2:1)=0,089, P(1:3)=P(3:1)=0.005, et enfin P(2:2)=0,005.

| DA | DE | EL | EN | ES | FI | FR | GR | IT | NL | PT | SV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 1 | | | 1 | 1 | 1 | | 1 | |
| | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 3 | 3 | | 3 | 3 | 3 | 3 | 3 | | 3 | |
| | 4 | | | 4 | 4 | 4 | | 4 | | 4 | |
| | 5 | | 4 | 5 | 5 | 5 | | 5 | | 5 | |
| | 6 | | 5 | 6 | 6 | 6 | | 6 | | 6 | |
| | 7 | | 6 | | 7 | 7 | | | 9 | 7 | |
| | 8 | | 7 | 8 | 8 | 8 | | 8 | 10 | 8 | 9 |
| | 10 | | 8 | 9 | 10 | 9 | | 9 | 11 | 9 | |
| 11 | 11 | | 9 | | 11 | 10 | | 10 | 12 | | 13 |
| 12 | 12 | | 10 | 11 | 12 | 11 | | 11 | 13 | 11 | 14 |
| … | … | … | … | … | … | … | … | … | … | … | … |

*Tableau 2.9 : Exemple de points obtenus avant complétion finale*

On obtient alors les résultats de la deuxième colonne du tableau 2.8. Notons que le coût de cet algorithme est modéré, étant donné l'étroitesse de l'espace de recherche : pour obtenir les résultats précédents, l'algorithme de Gale & Church a été lancé 6 662 fois sur des intervalles d'une longueur moyenne de 4 phrases environ, l'intervalle le plus grand ayant une longueur de 75 phrases[28].

Pour le grec translittéré (GR) on remarque sans surprise que les performances de l'aligneur rejoignent celle du grec original (EL) à l'issue de l'étape de complétion finale.

### 2.5.6 Tuilage des couples de langue

Il est important de noter que les résultats précédents, s'appuyant sur les combinaisons de $Comb_{\text{FR-pivot}}$, peuvent être légèrement biaisés, vu que nous n'évaluons que les couples avec le français : il est vraisemblable que le rappel soit artificiellement majoré pour les couples de $Comb_{\text{FR-pivot}}$.

Cherchons maintenant une combinaison de langues qui soit optimale, sans s'appuyer *a priori* sur le français. Une piste consiste à chercher le meilleur *tuilage* des alignements. Par tuilage on entend un ensemble minimal de couples de langues tel que :

- chaque langue apparaît dans au moins un couple ;

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

– chaque couple possède au moins une langue en commun avec un autre couple.

Parmi les tuilages possibles, on cherchera le tuilage qui met en jeu les couples les plus fortement associés (d'après les données du tableau 46). Les couples ainsi formés peuvent en quelque sorte s'appuyer les uns sur les autres, de manière complémentaire, pour former un tout plus solide. En prenant pour chaque langue les trois meilleurs couples (en ligne) on obtient la matrice ci-dessous :

| | **da** | **de** | **en** | **es** | **fi** | **fr** | **it** | **nl** | **pt** | **sv** |
|---|---|---|---|---|---|---|---|---|---|---|
| **da** | | 1 114 | 1 202 | | | | | | | 2 325 |
| **de** | 1 114 | | 863 | | | | | | | 925 |
| **en** | | | | 1 968 | | 2 367 | 2 225 | | | |
| **es** | | | 1 968 | | | | 2 234 | | 3 750 | |
| **fi** | | | 527 | | | | 481 | | | 617 |
| **fr** | | | 2 367 | 1 829 | | | 2 120 | | | |
| **it** | | | 2 225 | 2 234 | | 2 120 | | | | |
| **nl** | 1 019 | | 1 174 | | | | 978 | | | |
| **pt** | | | 1 493 | 3 750 | | | 1 935 | | | |
| **sv** | 2 325 | 925 | 1 256 | | | | | | | |

*Tableau 2.10 : Filtrage des trois langues les plus proches, par ligne.*

On voit que dans la perspective d'un tuilage des alignements, l'anglais occupe une position centrale. On peut visualiser ce positionnement en utilisant la représentation graphique ci-dessous (figure 2.13) obtenue grâce au logiciel Gephi[29]. Cette figure montre le graphe associé à la précédente matrice, avec les paramètre suivants : l'épaisseur des arcs est proportionnelle à la force du lien d'association, la taille des nœuds est proportionnelle au *degré* pondéré de chaque nœud (c'est-à-dire à la somme de ses liens d'association), et la spatialisation a été obtenue grâce à l'algorithme ForceAtlas[30].

29 cf. https://gephi.org/, consulté en mai 2014.

30 Dans ce type d'algorithme de spatialisation, dit *force-based*, chaque nœud subit une force de répulsion qui diminue avec leur distance, comme des aimants, et les arcs se comportent comme des ressorts dont la raideur est proportionnelle à la pondération du lien. L'algorithme cherche à déterminer une spatialisation stable en fonction de ces contraintes. Les paramètres employés sont les suivants : Inertie=0.1, Force de répulsion=10 000, Force d'attraction=0,005, Gravité=30.

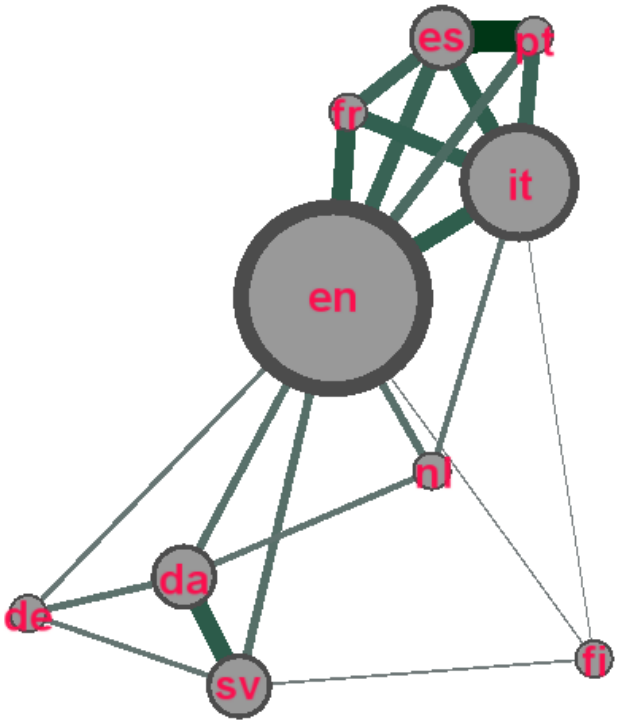


*Figure 2.13 : Représentation des couples de langues les plus fortement associés*

Par cette spatialisation, on cherche à représenter l'intensité des similarités (i.e. la quantité de cognats) entre les langues prises deux à deux. On obtient donc une représentation très différente de celle de la figure 2.8, basée sur un modèle vectoriel où la distance est une fonction de l'angle des vecteurs comparés, et non de leur norme[31].

On peut alors supposer qu'un tuilage basé sur l'anglais comme pivot est susceptible de donner de bons résultats : on notera $Comb_{EN\text{-pivot}}$={EN-FR, EN-IT, EN-ES, EN-PT, EN-SV, DA-EN, EN-NL, DE-EN, EN-FI, EN-GR, EN-EL} (ici les couples sont listés par force d'association décroissante).

Enfin, une autre stratégie consiste à prendre l'ensemble des couples qui maximise la somme des liens d'association, tout en tenant compte des contraintes de tuilage. Pour ce faire, on commence par constituer la liste de tous les couples, triée par force d'association décroissante. En parcourant cette liste, on retient alors les couples qui introduisent une ou

[31] Par exemple le vecteur de l'allemand (1114, _ , 863, 448, 397, 735, 747, 722, 376, 925) est considéré comme proche du vecteur du néerlandais (1019, 722, 1174, 638, 197, 936, 978, _ , 489, 893) parce que l'angle entre ces deux vecteurs est relativement faible, et il s'éloigne du vecteur de l'italien (1041,747, 2225, 2234, 481, 2120, _ , 978, 1935, 354), avec qui l'angle est plus important. En revanche, si on considère la similarité deux à deux en quantité de cognats, l'allemand est plus proche de l'italien (747) que du néerlandais (722). Les arcs de la figure 2.13 montrent la force de ces associations 2 à 2, ainsi que leur cumul (taille des noeuds).

deux nouvelles langues (par rapport aux couples déjà parcourus) ou qui introduisent un arc qui ne peut être déduit des précédents par transitivité. On s'arrête dès que l'on obtient un tuilage complet. Avec cet algorithme, on obtient : $Comb_{Max}$={ES-PT, EN-FR, DA-SV, ES-IT, EN-IT, SV-EN, NL-EN, DE-DA, FI-SV}. Enfin, à titre de *baseline*, nous avons testé également deux autres combinaisons :

– un tuilage « aléatoire » basé sur l'ordre alphabétique des codes de langue : $Comb_A$={DA-DE, DE-EL, EL-EN, EN-ES, ES-FR, FR-FI, FI-GR, GR-IT, IT-NL, NL-PL, PT-SV}

– un tuilage incomplet basé sur la liste des meilleurs couples pour chaque langue : $Comb_{Inc}$={ES-PT, EN-FR, DA-SV, ES-IT, NL-EN, DE-DA, FI-SV}. Pour que cette combinaison fonctionne avec notre algorithme, nous devons néanmoins ajouter la combinaison DA-DE-EL-EN-ES-FR-FI-GR-IT-NL-PT-SV, afin de fournir, dans une première passe, quelques points d'ancrage susceptibles de relier toutes les langues entre elles, afin de servir de point d'appui aux algorithmes de complétion.

Les résultats comparés de ces combinaisons sont donnés dans le tableau ci-dessous :

| | $Comb_{EN\text{-pivot}}$ | | $Comb_{Max}$ | | $Comb_A$ | | $Comb_{Inc}$ | |
|---|---|---|---|---|---|---|---|---|
| Couple de langues | *P* | *R* | *P* | *R* | *P* | *R* | *P* | *R* |
| DA-FR | 0,97 | 0,80 | 0,98 | 0,82 | 0,97 | 0,78 | 0,93 | 0,76 |
| DE-FR | 0,95 | 0,79 | 0,97 | 0,81 | 0,95 | 0,75 | 0,93 | 0,74 |
| EL-FR | 0,96 | 0,79 | 0,96 | 0,80 | 0,96 | 0,77 | 0,93 | 0,75 |
| EN-FR | 0,98 | 0,88 | 0,99 | 0,89 | 0,99 | 0,85 | 0,99 | 0,86 |
| ES-FR | 0,97 | 0,81 | 0,95 | 0,81 | 0,99 | 0,85 | 0,93 | 0,77 |
| FI-FR | 0,95 | 0,85 | 0,97 | 0,88 | 0,99 | 0,90 | 0,93 | 0,80 |
| FR-GR | 0,96 | 0,80 | 0,94 | 0,78 | 0,97 | 0,81 | 0,91 | 0,73 |
| FR-IT | 0,97 | 0,81 | 0,97 | 0,81 | 0,97 | 0,79 | 0,93 | 0,74 |
| FR-NL | 0,92 | 0,64 | 0,93 | 0,65 | 0,94 | 0,63 | 0,92 | 0,63 |
| FR-PT | 0,97 | 0,86 | 0,97 | 0,85 | 0,97 | 0,82 | 0,93 | 0,79 |
| FR-SV | 0,96 | 0,76 | 0,96 | 0,78 | 0,95 | 0,70 | 0,92 | 0,72 |
| Moyenne | 0,96 | 0,80 | 0,96 | 0,81 | 0,97 | 0,79 | 0,93 | 0,76 |

*Tableau 2.11 : Résultats comparés pour différents tuilages*

Bien que $Comb_{Max}$ s'avère la meilleure combinaison, les différences sont modestes, sauf pour la combinaison incomplète, beaucoup moins bonne car elle s'appuie sur les algorithmes de complétion pour établir des liens entre les langues qui ne sont pas reliées, directement, ou indirectement, par les couples initiaux. Le biais lié à l'utilisation du français comme pivot est confirmé, puisque $Comb_{Max}$ obtient une F-mesure de 2 points inférieure à celle de $Comb_{FR\text{-pivot}}$. (89,94 % contre 87,73% pour $Comb_{Max}$).

Ces résultats très proches s'expliquent par le fait que l'algorithme de complétion finale compense les résultats plus pauvres des combinaisons moins appropriées. Avant complétion finale, $Comb_{Max}$ obtient un rappel de 0,38 alors que $Comb_{A}$ seulement 0,29 pour une précision presque identique de 0,98 : la nature du tuilage a donc bien un effet sur les premières phases de préalignement, mais cet effet disparaît presque totalement après l'étape finale.

Notons que le tuilage optimal, ici déduit de l'alignement de référence, peut également être calculé à partir d'un alignement automatiquement obtenu à partir d'un tuilage aléatoire. Une fois déterminé, il peut être réutilisé pour d'autres multi-textes. En effet, étant étroitement lié à des aspects génétiques, il possède un caractère de généralité.

### 2.5.7 Comparaison avec les méthodes binaires

Reste à déterminer, à l'issue de ces différentes observations, si le recours au multi-alignement présente vraiment un intérêt par rapport à l'alignement binaire : c'est la question centrale à laquelle il nous faut maintenant tenter de donner une réponse. Pour ce faire, nous avons téléchargé l'aligneur Vanilla[32] (Danielsson & Riding, 1997), basé sur l'algorithme de Gale & Church (1991), encore très couramment utilisé (notamment pour les corpus OPUS et JRC). Voici les résultats obtenus pour le français sur notre corpus d'évaluation :

[32] Nous avons téléchargé cette implémentation à l'adresse: http://www2.lael.pucsp.br/corpora/alinhador/ (consulté en mai 2014)

| Couple de langues | Vanilla | | JAM $Comb_{Max}$ | | JAM $Comb_{Max}$ + *GC* | | JAM *bi* + *GC* | |
|---|---|---|---|---|---|---|---|---|
| | *P* | *R* | *P* | *R* | *P* | *R* | *P* | *R* |
| DA-FR | 0,94 | 0,93 | 0,98 | 0,82 | 0,97 | 0,96 | 0,90 | 0,91 |
| DE-FR | 0,94 | 0,95 | 0,97 | 0,81 | 0,95 | 0,95 | 0,95 | 0,95 |
| EL-FR | 0,09 | 0,12 | 0,96 | 0,80 | 0,96 | 0,97 | 0,93 | 0,96 |
| EN-FR | 0,98 | 0,98 | 0,99 | 0,89 | 0,98 | 0,99 | 0,97 | 0,98 |
| ES-FR | 0,90 | 0,92 | 0,95 | 0,81 | 0,97 | 0,97 | 0,97 | 0,97 |
| FI-FR | 0,96 | 0,97 | 0,97 | 0,88 | 0,98 | 0,99 | 0,96 | 0,97 |
| FR-GR | 0,94 | 0,95 | 0,94 | 0,78 | 0,96 | 0,97 | 0,93 | 0,96 |
| FR-IT | 0,95 | 0,96 | 0,97 | 0,81 | 0,96 | 0,97 | 0,94 | 0,97 |
| FR-NL | 0,80 | 0,80 | 0,93 | 0,65 | 0,92 | 0,92 | 0,90 | 0,90 |
| FR-PT | 0,96 | 0,97 | 0,97 | 0,85 | 0,96 | 0,97 | 0,95 | 0,97 |
| FR-SV | 0,87 | 0,89 | 0,96 | 0,78 | 0,94 | 0,95 | 0,92 | 0,94 |
| Moyenne (hors EL) | 0,92 | 0,93 | 0,96 | 0,81 | 0,96 | 0,96 | 0,94 | 0,95 |

*Tableau 2.12 : Résultats comparés de Vanilla et des différentes versions de JAM (avec et sans l'application a posteriori de l'algorithme de Gale & Church)*

Les résultats pour le grec (EL) sont mauvais, mais nous pensons qu'il s'agit d'une mauvaise prise en compte du codage UTF-8 par Vanilla, et nous n'avons pas intégré ces résultats dans la moyenne (à priori, l'algorithme de Gale & Church ne s'appuyant que sur les longueurs de phrases, EL et GR devraient être identiques). Nous avons donc calculé les moyennes sans cette ligne (en gris).

À première vue, il semblerait que Vanilla obtienne de meilleurs résultats, avec plus de 10 points d'écart pour le rappel et une précision légèrement inférieure. Mais il se trouve que les alignements de Vanilla et de JAM ne sont pas directement comparables, notamment en ce qui concerne le rappel, car ils sont construits différemment : Vanilla extraie un alignement *complet*, comportant de nombreux regroupements de type 1-2, 2-1 et 2-2. tandis que JAM n'extrait que des alignements 1-1, et ceci afin de conserver son caractère de multi-alignement – ce qui explique un rappel inférieur.

Un multi-alignement complet, construit à partir de regroupements de type 1-2, 2-1 et 2-2, serait *ipso facto* beaucoup moins précis. En effet, si on applique la propriété de transitivité sur des alignements binaires complets, on peut obtenir des regroupements très larges : il suffit qu'un alignement pour un couple de langues chevauche deux groupes de phrases différents pour d'autres couples pour que ceux-ci fusionnent, et ainsi de suite. Nous en avons fait l'essai en prenant la clôture transitive de nos alignements de référence avec le français, et nous obtenons des groupes élargis qui

peuvent compter jusqu'à 13 phrases pour un seul groupe. Le tableau 2.13 en donne un échantillon pour le début du corpus et le tableau 2.14 donne les alignements transitifs simples issus de JAM correspondants, à titre de comparaison :

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

*Tableau 2.13 : Groupes obtenus par fusion transitive des 11 alignements de référence avec le français*

| DA | DE | EL | EN | ES | FI | FR | GR | IT | NL | PT | SV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. |
| 7.3 | 7.3 | 7.3 | 7.3 | 7.3 | 7.3 | 7.3 | 7.3 | 7.3 | | 7.3 | 7.4 |
| 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 |
| 8.2 | 8.2 | 8.2 | 8.2 | 8.2 | 8.2 | 8.2 | 8.2 | | 8.2 | 8.2 | 8.2 |
| 8.3 | 8.3 | 8.3 | 8.3 | 8.3 | 8.3 | 8.3 | 8.3 | | 8.4 | 8.3 | 8.3 |
| 8.4 | 8.4 | 8.4 | 8.4 | 8.4 | 8.4 | 8.4 | 8.4 | 8.2 | 8.5 | 8.4 | 8.4 |
| 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 |
| 10.1 | 10.2 | 10.1 | 10.1 | 10.1 | 10.1 | 10.1 | 10.1 | 10.1 | 10.1 | 10.1 | 10.2 |
| | 11.1 | 11.1 | 11.1 | 11.1 | | 11.1 | 11.1 | | 11.1 | | |
| 11.2 | 11.2 | 11.2 | 11.2 | 11.2 | 11.1 | 11.2 | 11.2 | 11.1 | 11.2 | | 11.1 |
| | | | | 11.3 | | | | | | 11.1 | |
| 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 | 12.1 |
| 13.2 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 | 13.1 |
| 13.3 | 13.2 | 13.2 | 13.2 | 13.2 | 13.2 | 13.2 | 13.2 | 13.2 | | 13.2 | 13.2 |
| | 14.1 | | | | | | | | 13.2 | | 14.1 |
| 14.1 | 14.2 | 14.1 | 14.1 | 14.1 | 14.1 | 14.1 | 14.1 | | | 14.1 | 14.2 |
| 14.2 | 14.3 | 14.2 | 14.2 | 14.2 | 14.2 | 14.2 | 14.2 | 14.1 | 14.1 | 14.2 | 14.3 |
| 14.3 | 14.4 | 14.3 | 14.3 | 14.3 | 14.3 | 14.3 | 14.3 | 14.2 | 14.2 | 14.3 | 14.4 |
| ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. |

*Tableau 2.14 : Alignements transitifs simples issus de JAM*

S'il n'est pas intéressant d'extraire un multi-alignement complet, il est en revanche possible de s'appuyer sur un tel multi-alignement pour en tirer rapidement un alignement bilingue complet. C'est ce que nous avons fait en appliquant l'algorithme de Gale & Church pour chaque couple évalué, en guidant l'espace de recherche par les points obtenus avec JAM (les points considérés ne doivent pas s'éloigner de plus d'une phrase des points issus de JAM, lorsqu'une de leurs coordonnées coïncide avec un de ces points). En moyenne les points issus de JAM sont éloignés de 1,126 phrases, ce qui rend l'espace de recherche très étroit et l'alignement trivial : on peut dire qu'à ce stade le chemin d'alignement est déjà connu et l'algorithme se contente d'effectuer les regroupements nécessaires entre deux points consécutifs. Les résultats sont donnés dans la troisième colonne du tableau 2.12 (colonne $Comb_{Max}$ + *GC*). Cette fois-ci les résultats sont bien meilleurs : Vanilla obtient une F-mesure globale de 92,9 %, tandis que JAM obtient 96,2 %. JAM est meilleur pour presque tous les couples de langues (sauf pour FR-PT, où Vanilla obtient 0,48 % de F-mesure en plus), mais il est surtout plus robuste, et ne connaît pas de forte dégradation pour les couples les plus « difficiles » comme FR-NL, FR-SV ou FR-DA.

Enfin, nous avons voulu comparer les résultats de JAM utilisé en bilingue, afin d'évaluer le gain du multi-alignement par rapport à un simple bi-alignement, en utilisant rigoureusement les mêmes algorithmes. On obtient les résultats de la 4$^{e}$ colonne du tableau 2.12 (JAM *bi+GC*) : on constate qu'ils sont meilleurs que ceux de Vanilla, mais plus coûteux en calcul que ce dernier (du fait de l'algorithme itératif) et légèrement inférieurs à ceux du multi-alignement. La différence est toutefois modeste : les textes étant bien parallèles et plutôt faciles à aligner, la marge de progression liée au multi-alignement est sans doute assez faible. En outre, les alignements bilingues de JAM *bi+GC* sont tous centrés sur le français, et bénéficient donc du biais positif déjà évoqué (avec le tuilage *FR-pivot* on obtenait 2 points de F-mesure en plus). L'intérêt du multi-alignement est qu'il fournit des résultats globalement meilleurs, et sous une forme qui concerne tous les couples de langues à la fois.

Nous terminerons cette comparaison par une évaluation de la robustesse comparée de ces approches vis-à-vis des ruptures dans le parallélisme des traductions. Pour ce faire, nous avons créé artificiellement des « trous » dans la version française du corpus, en éliminant de façon aléatoire des blocs de phrases. Dans une première expérimentation, nous avons supprimé aléatoirement un bloc d'une seule phrase, en réitérant respectivement 10 fois, 50 fois et 100 fois. Nous avons alors lancé JAM (avec $Comb_{Max}$ cf. deuxième colonne du tableau 2.12) et Vanilla sur ces textes au parallélisme dégradé. On obtient les résultats suivants :

| Nombre de blocs supprimés | Vanilla P | JAM P | Vanilla R | JAM R | Vanilla F | JAM F |
|---|---|---|---|---|---|---|
| 10 | 0,89 | 0,96 | 0,91 | 0,80 | 0,90 | 0,87 |
| 50 | 0,77 | 0,93 | 0,84 | 0,76 | 0,80 | 0,83 |
| 100 | 0,62 | 0,81 | 0,72 | 0,63 | 0,66 | 0,71 |

*Tableau 2.15 : Résultats comparés de Vanilla et JAM ($Comb_{Max}$ + GC) pour le corpus français dégradé (blocs de taille 1)*

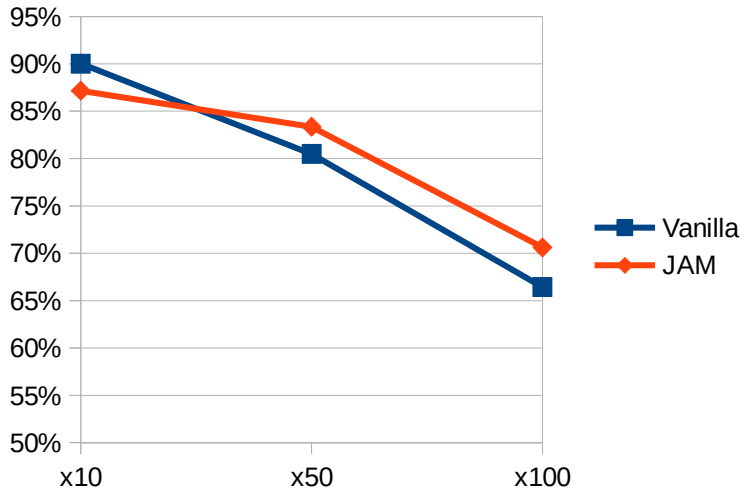


*Figure 2.14 : Résultats comparés de Vanilla et JAM pour le corpus français dégradé (blocs de taille 1)*

Ces résultats montrent que JAM, bien que partant avec un rappel plus faible (puisqu'il produit des multi-alignements et non des alignements binaires complets), résiste mieux à la dégradation du corpus, et atteint rapidement une F-mesure supérieure à Vanilla.

Dans une deuxième expérimentation, nous avons étudié l'effet de la taille des blocs supprimés : cette fois nous ne supprimons qu'un seul bloc, comportant respectivement 10, 20, 50, 100, 200 et 300 phrases. Il s'agit de déterminer comment ces méthodes se comportent vis-à-vis d'une rupture de grande taille (et non par rapport à plusieurs petites ruptures disséminées ça et là). Pour JAM, avec des ruptures de 100 phrases et plus, il est nécessaire de relâcher la contrainte de parallélisme en augmentant le seuil *MaxDiffInterval2* à 0,5 (au lieu de 0,1), sans quoi les points situés autour de la zone supprimée ne peuvent plus être considérés.

| Taille du bloc supprimé | Vanilla P | JAM P | Vanilla R | JAM R | Vanilla F | JAM F |
|---|---|---|---|---|---|---|
| 10 | 0,82 | 0,96 | 0,85 | 0,81 | 0,83 | 0,88 |
| 20 | 0,72 | 0,95 | 0,75 | 0,79 | 0,74 | 0,87 |
| 50 | 0,45 | 0,96 | 0,50 | 0,78 | 0,47 | 0,86 |
| Relâchement de la contrainte de parallélisme pour JAM (*MaxDiffInterval2* = 0,5) | | | | | | |
| 100 | 0,54 | 0,90 | 0,62 | 0,67 | 0,57 | 0,76 |
| 200 | 0,25 | 0,91 | 0,33 | 0,68 | 0,28 | 0,78 |
| 300 | 0,06 | 0,87 | 0,08 | 0,44 | 0,07 | 0,59 |

*Tableau 2.16 : Evolution des résultats en fonction de la taille des blocs supprimés*

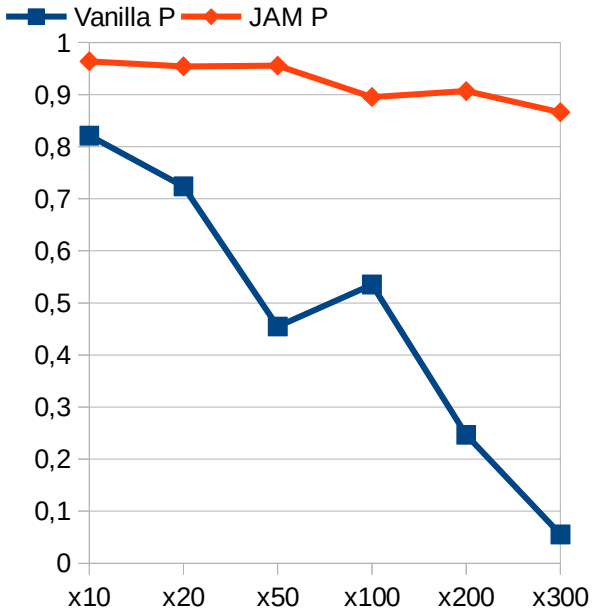


*Figure 2.15 : Evolution de la précision en fonction de la taille des blocs supprimés*

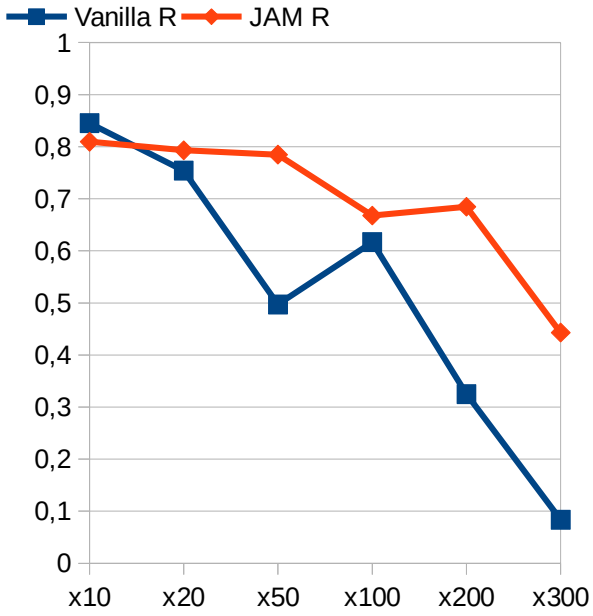


*Figure 2.16 : Evolution du rappel en fonction de la taille des blocs supprimés*

Cette fois la différence de comportement est très nette : alors que Vanilla connaît une rapide dégradation à la fois de la précision et du rappel, la précision de JAM se maintient à un niveau élevé, même si le rappel chute progressivement. Globalement, l'alignement issu de JAM reste exploitable pour toutes les valeurs (par exemple pour une exploitation en traduction statistique), même s'il devient de plus en plus incomplet. JAM est donc plus robuste : pour reprendre un terme de systémique, on peut parler de dégradation gracieuse (*graceful degradation*) des résultats.

Enfin, nous avons voulu faire une ultime vérification, pour confirmer que cette robustesse est bien due au multi-alignement : nous avons lancé JAM en bilingue, comme précédemment. Cette fois nous n'avons testé qu'un seul cas de figure – la suppression d'un bloc de 200 phrases : on trouve une précision de 0,94 avec un rappel de 0,57. Si on effectue un multi-alignement comparable, c'est-à-dire centré sur le français avec la combinaison *FR-pivot*, on obtient la même précision de 0,94 mais avec un rappel de 0,71, soit une différence de 10 points au niveau de la F-mesure. La robustesse spécifique du multi-alignement se manifeste donc plutôt au niveau du rappel, la bonne tenue de la précision étant due à notre architecture itérative plus qu'à l'intervention de plusieurs langues simultanément.

## 2.6. Conclusion

Nous avons retracé nos premiers travaux concernant l'alignement bilingue et l'extraction de correspondances lexicales. Ce faisant, nous avons complété ces travaux par une étude originale concernant le multi-alignement, c'est-à-dire l'alignement de plusieurs langues simultanément. Les résultats expérimentaux confirment nos intuitions initiales, à savoir que les multi-textes renferment un entrelacs de correspondances superficielles, qui manifestent tout un réseau de *corrélations*, situées à plusieurs niveaux :

– corrélations de segmentation et de compositionnalité, observables au niveau des longueurs des paragraphes, des phrases ou de groupes de phrases ;

– corrélations de contenu, observables par la régularité des correspondances lexicales, statistiquement significatives, signes de la construction d'équivalences traductionnelles récurrentes au niveau lexical ;

– enfin, découlant de ces dernières, des corrélations de langue à langue, résultant d'une complexe stratification de phénomènes génétiques et historiques, et manifestes au niveau le plus superficiel des ressemblances graphiques – même pour des langues qui ne partagent pas les mêmes alphabets comme le français et l'arabe.

Par leur caractère superficiel, ces corrélations peuvent être exploitées, au moins dans un premier temps, par des méthodes génériques sans traitement linguistique spécialisé. Dans cette perspective, nous avons montré qu'il était profitable d'intégrer simultanément tous les niveaux de corrélation (segmentation et ressemblances graphiques) et de s'appuyer le plus possible sur l'ensemble de la structure multi-textuelle. Notamment, nous avons montré qu'il

était possible d'appuyer nos méthodes sur une forme de *tuilage*, économique sur le plan calculatoire, et s'appuyant sur les « parentés linguistiques » déduites *a posteriori* des ressemblances de surface.

Nous en avons fourni la preuve en démontrant la supériorité des méthodes de multi-alignement vis-à-vis des algorithmes usuels de bi-alignement, avec une meilleure précision et un meilleur rappel, tant pour des textes strictement parallèles que pour des textes avec une compositionnalité traductionnelle dégradée. En outre, sur le plan de la complexité, le multi-alignement, bien qu'un peu plus coûteux que les classiques algorithmes linéaires de type Gale & Church (1991), présente l'avantage de fournir une structure de données compacte renfermant un grand nombre de couples – avec une complexité en espace[33] pour le stockage des résultats bien meilleure (en O($n$) pour $n$ langues, contre O($n^2$) dans le cas bilingue).

Dans la partie suivante de cette synthèse, nous allons tenter de dépasser le niveau des corrélations superficielles, afin de déterminer comment la multi-textualité peut permettre d'observer des *contrastes*, non plus au niveau des textes, mais au niveau des langues.

[33] Le terme « complexité en espace », en informatique, désigne le coût algorithmique en termes d'occupation de l'espace la mémoire (en mémoire vive ou sur une unité de stockage).

# 3. Quels contrastes ?

> « La traductibilité apparaît comme une des propriétés fondamentales des systèmes sémiotiques et comme le fondement même de la démarche sémantique : entre le jugement existentiel " il y a du sens " et la possibilité d'en dire quelque chose, s'intercale en effet la traduction ; " parler du sens " c'est à la fois traduire et produire de la signification. »
>
> Greimas & Courtès (1993 : 397-398)

Greimas & Courtès (1993) suggèrent ainsi qu'il y a une parenté entre l'acte de traduire et la démarche sémantique, consistant à « parler du sens ». La traduction se présente comme un premier pas vers la glose et l'explicitation. D'où vient ce pouvoir quasi-métalinguistique de la traduction ? Jakobson (1963 : 80) l'a très bien résumé : « l'équivalence dans la différence est le problème cardinal du langage et le principal objet de la linguistique ». Par le jeu des différences et des équivalences, le texte traduit en dit plus que le texte original : il fournit certes une interprétation de ce dernier – mais en sus, il « parle » de l'idiome d'arrivée et de l'idiome de départ – qui se dessine par ses différences, comme en négatif.

Dans ce chapitre nous allons aborder la question des corrélations entre langues, révélées par les corrélations entre textes, et nous verrons que ces corrélations font également apparaître des contrastes.

## 3.1. Extraction de lexiques bilingues

C'est au niveau lexical que les corrélations apparaissent de prime abord de la façon la plus évidente. Que l'on utilise les techniques issues des travaux précurseurs d'IBM sur la traduction statistique (Brown *et al.*, 1991, Och & Ney, 2003) ou des techniques plus simples basées sur un algorithme de type *Competitive linking algorithm* (Melamed, 1998, Kraif & Chen, 2004), l'extraction de correspondances lexicales permet de dériver à peu de frais un lexique bilingue spécifique à un corpus.

Pour filtrer le bruit lié aux correspondances erronées, et éliminer les correspondances trop idiosyncratiques car non séparables de leur co-texte, il suffit de retenir les correspondances observées avec une fréquence statistique significative. Dans l'exemple ci-dessous, tiré d'un alignement anglais-français d'un récit de Stevenson, seules les correspondances observées plus de 3 fois ont été retenues :

| | | | |
|---|---|---|---|
| during-PRE | pendant-PRE (6) | eight-QUA | huit-QUA (4) |
| dust-NOM | poussière-NOM (14) | eighty-QUA | quatre-QUA (5) |
| dusty-ADJ | poussiéreux-ADJ (3) | elegance-NOM | élégance-NOM (3) |
| dwarf-NOM | rabougrir-VER (3) | eloquence-NOM | éloquence-NOM (3) |
| dye-PPS | teinter-PPS (3) | embarrass-PPS | embarras-NOM (5) |
| ear-NOM | oreille-NOM (18) | | embarrasser-PPS (3) |
| earth-NOM | terre-NOM (4) | employé-NOM | libre-ADJ (3) |
| eastern-ADJ | là-ADV (3) | empty-ADJ | vide-ADJ (6) |
| easy-ADJ | facile-ADJ (6) | encampment-NOM | campement-NOM (3) |
| eat-VER | manger-VER (6) | encumber-PPS | estimer-VER (3) |
| edict-NOM | le-DET (3) | engage-PPS | engager-PPS (3) |
| egg whisk-NOM | oeuf-NOM (3) | enough-ADV | assez-ADV (10) |

*Tableau 3.1 : Extrait d'un lexique bilingue tiré d'un alignement anglais-français de* with a Donkey in the Cevennes*, de Stevenson*

Au vu du lexique ainsi obtenu, on constate que les correspondances erronées sont souvent liées à des problèmes d'identification des unités polylexicales (comme *egg whisk* ↔ *œuf*, ou *eighty* ↔ *quatre*). Ce « bruit » peut cependant être aisément écarté pour des corpus de grande dimension : à mesure que les données deviennent statistiquement plus significatives,

les régularités émergent et se distinguent des associations bruitées, plus instables par nature. Par ailleurs, les effets liés à la textualité, aux spécificités d'un thème, aux habitudes de l'auteur, aux choix particuliers d'un traducteur, etc., s'estompent à mesure que le corpus augmente et devient plus représentatif de la langue générale (ou d'une langue de spécialité si l'on vise un corpus spécialisé).

Comme dans toute recherche de linguistique de corpus, on peut alors partir de l'observation du texte pour viser la langue, par un mouvement inductif. De ce point de vue, les multi-textes ne permettent pas seulement d'étudier deux langues, prises du point de vue du code, mais de les confronter et de les éclairer réciproquement, en s'appuyant sur les structures et les régularités originales que font apparaître les contrastes.

Il est par exemple relativement aisé d'établir automatiquement des classes de synonymes, sur la base de la transitivité de la relation d'équivalence (Kraif, 2008a). La figure 3.1 montre les résultats d'une requête élaborée de manière itérative, en recherchant initialement l'expression *de temps en temps*. Les couples de phrases trouvés pointent l'expression équivalente *from time to time*. En recherchant cette dernière expression en anglais, de nouveaux couples de phrases sont identifiés, contenant d'autres équivalents en français *de temps à autre*, *par instants*. En cherchant ces nouvelles expressions, on trouve alors de nouveaux équivalents anglais *now and then*, *ever and again*… On peut réitérer ce processus de l'aller-retour jusqu'à obtenir des classes stables. L'alignement contenant des appariements bruités, un filtrage est parfois nécessaire, afin de ne retenir que les équivalences les plus significatives, et constituer des classes réduites avec un noyau sémantique cohérent.

Les classes ainsi obtenues à partir de l'équivalence traductionnelle représentent plus des communautés de voisinage sémantique que des relations strictes de synonymie, et elles transcendent souvent les catégories morphologiques. Par exemple, en appliquant l'« aller-retour » au mot *âne* on obtient la classe suivante : *âne, ânesse, ânier, bourriquet, bourrique, bourricot, baudet*. On y trouve donc des embryons de paradigmes morphologiques qui pourraient servir d'appui à une étude sur la morphologie dérivationnelle et/ou flexionnelle.

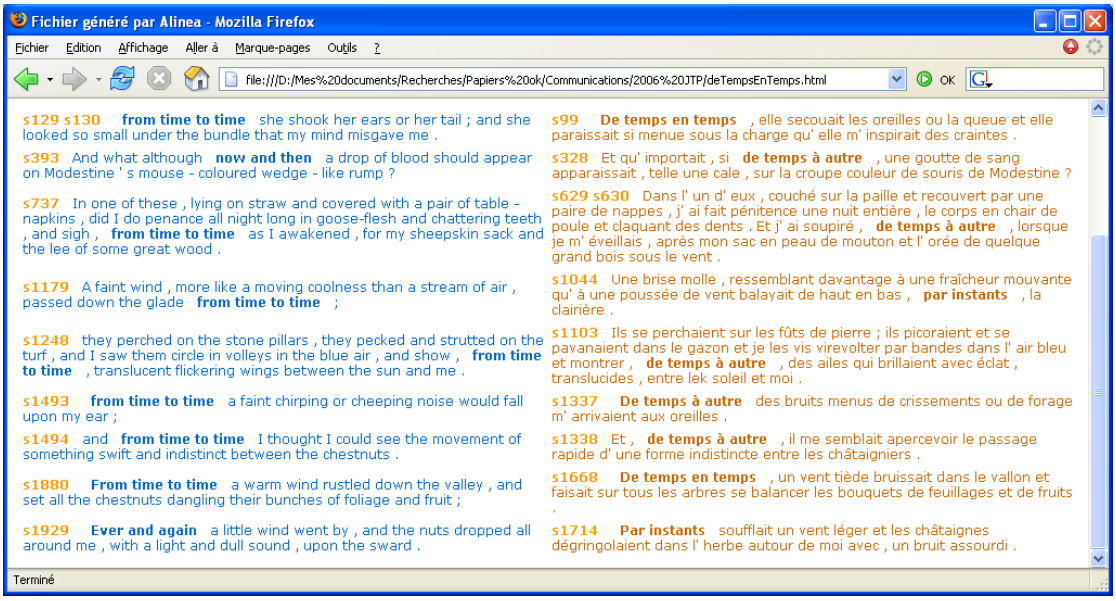


*Figure 3.1 : Concordance extraite à partir d'une requête bilingue d'Alinéa, sur la traduction de* With a Donkey in the Cevennes*, de Stevenson*

## 3.2. Une perspective lexicographique ?

Le même procédé appliqué sur une traduction italienne de *Madame Bovary*, de Flaubert, donne, en partant de *s'imaginer* :

> *FR : s'imaginer, penser, croire, soupçonner, se douter, songer, se persuader, craindre, avoir peur, redouter (+ que/de)*
>
> *IT : pensare, credere, immaginare, supporre, aspettarsi, temere, essere convinto, essere nella speranza, convincersi, sospettare (+ che/di), interpretare come*

Notons qu'ici, l'aller-retour est pratiqué manuellement : la classe obtenue est le résultat d'une sélection, la méthode pouvant rapidement diverger vers des formes sans rapport avec l'entrée initiale. Cette classe fait apparaître plusieurs acceptions de *s'imaginer* : /opinion/, /croyance/, /doute/ /crainte/, … Le bi-texte, en reliant une entrée à différents équivalents traductionnels, et réciproquement, permet donc d'explorer, sans recours à une glose, ses virtualités sémantiques.

En 2006, pour permettre une navigation simplifiée entre les équivalents de traduction et les contextes correspondants, nous avons doté Alinéa d'un format de sortie en HTML, comportant des liens hypertextes entre les différentes traductions repérées et leurs contextes alignés (cf. figure 3.2).

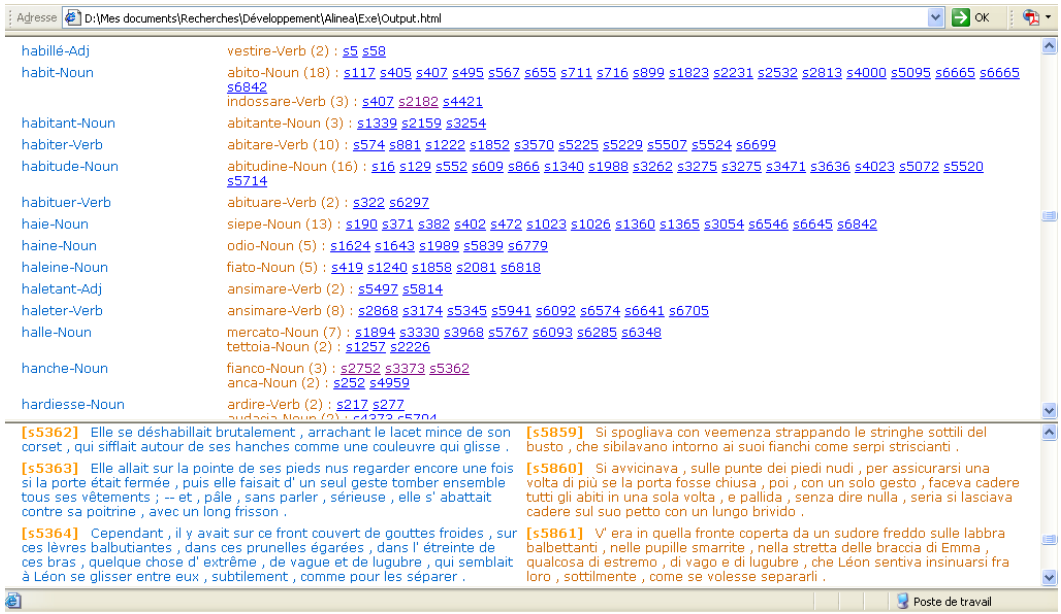


*Figure 3.2: Une sortie HTML d'Alinea permettant l'exploration des équivalents et de leurs contextes*

Une telle présentation peut s'apparenter à un dictionnaire bilingue brut susceptible de donner des traductions et des exemples en contexte. On peut reprocher à un tel dispositif de ne pas faire le tri nécessaire entre les occurrences spécifiques à un contexte et les exemples de portée générale – ce que fait justement un lexicographe. Atkins (citée par Grundy, 1996 :146) note que cette recherche d'exemplarité et de généralité est impérative en lexicographie bilingue :

> Il y a une différence considérable entre l'équivalent qui correspond parfaitement bien au contexte spécifique dont a besoin le traducteur et l'équivalent hors contexte que le dictionnaire bilingue se doit de proposer. Le rôle principal d'un dictionnaire est de ne pas induire l'utilisateur en erreur (…) le devoir du lexicographe est de proposer une traduction générale, dont l'utilisation ne peut pas être totalement fausse plutôt qu'une traduction qui serait parfaite dans certains contextes mais impossibles dans d'autres.

Cependant, comme nous le signalions précédemment, le filtrage des équivalents les plus fréquents permet de parvenir à un certain degré de généralité sur le plan de l'équivalence. Quant à l'interprétation fine du sens en contexte, on peut légitimement supposer que la multiplication des exemples peut permettre à l'utilisateur de faire le tri, et de s'orienter. C'est le principe d'un outil comme *Linguee*, lancé en 2009 à Cologne, aujourd'hui devenu très


*Figure 3.3 : Exemple de requête avec Linguee*

populaire avec 32 millions de visites par mois[34], et dont nombre d'utilisateurs sont des traducteurs professionnels ou des rédacteurs en langue étrangère. Tout comme notre prototype, *Linguee* propose un double affichage, avec dans le volet gauche un « dictionnaire rédactionnel » proposant des équivalents de traductions manuellement validés et ordonnés par fréquence décroissante, et à droite une série de bi-phrases illustrant ces correspondants (cf. figure 3.3).

Dans ce même esprit, mais cette fois plus spécifiquement appliqué au domaine de l'aide à la rédaction, nous avions réfléchi à une architecture permettant d'intégrer un dictionnaire de collocations françaises à un corpus bilingue français-anglais (Kraif & Tutin, 2006, 2011). Selon cette architecture, une fiche concernant une collocation contiendrait les informations suivantes :

- Classe sémantique de la base, classe sémantique du collocatif
- Fréquence d'occurrence des différentes alternances syntaxiques observées dans le corpus (voix active, voix passive, tournure impersonnelle, nominalisation, etc.).
- Distribution par type de texte
- Autres collocations liées à la même base (*ibid.*)

[34] IVW-Measurement, septembre 2013 (cf. http://www.linguee.fr/francais-anglais/page/advertising.php, consulté en juin 2014)

À ce dictionnaire serait associé un corpus bilingue dans lequel les collocations en français seraient identifiées et manuellement annotées (en relation avec les entrées du dictionnaire). Nous avions prévu plusieurs modes d'interrogation :

– De la langue source (EN) vers la langue cible (FR), en partant d'une forme simple (par exemple *hypothesis*) ou d'une structure collocationnelle (*to put forward a hypothesis*). Deux types de résultats seraient présentés : les collocations présentes dans le dictionnaire, équivalents potentiels correspondant à une traduction séparée de la base et du collocatif ; une concordance tirée d'un corpus bilingue, dans lequel les collocations en français (langue cible) seraient préalablement annotées (et manuellement validées). Dans ces deux types de résultats, les collocations identifiées en français renverraient vers la fiche lexicographique précédemment décrite.

– En partant de la langue cible (FR), la recherche pourrait être faite à partir d'une base connue (p.ex. *hypothèse*) ou d'une classe sémantique particulière (p. ex. les verbes liés à la démonstration), en appliquant éventuellement une fonction lexicale (verbe support de *hypothèse*). En outre, comme le note Caviglia (2005) les rédacteurs en langue seconde sont souvent conscients de l'inadéquation de certaines formulations calquées sur leur langue maternelle. Il serait donc intéressant d'interroger la base en lui soumettant une collocation jugée douteuse. La méthode de recherche serait alors identique à celle mise en œuvre précédemment, mais plutôt que de recourir à un lexique bilingue pour traduire, on s'appuierait sur un dictionnaire de synonymes afin de tester divers candidats, de manière similaire à Shei & Pain (2000). Comme le proposent ces auteurs, il peut en effet être intéressant d'enregistrer les collocations erronées ou douteuses dans une liste d'erreurs fréquentes (*error library*), qui viendrait s'enrichir avec l'utilisation de la base.

La figure 3.4 montre ces deux parcours d'interrogation du dictionnaire et des corpus.

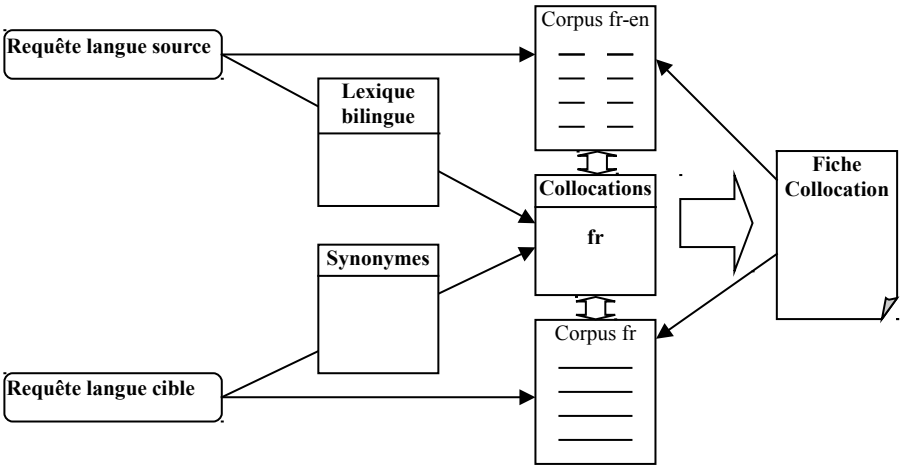


*Figure 3.4 : Interrogation de la base et des corpus dans un système d'aide à la rédaction (Kraif & Tutin, 2006)*

Dans le cadre d'un projet financé par la région Rhône-Alpes, coordonnée pas Agnès Tutin et moi-même, nous avions commencé a réunir, en 2004, un corpus bilingue français-anglais de textes scientifiques et techniques (que nous nommerons *Corpus Emergence*), dans la perspective de l'aide à la rédaction scientifique. Malheureusement, il ne nous a pas été possible de réunir un corpus parallèle d'une dimension suffisante, car il existe peu de traductions d'articles et de thèses – soit que les auteurs rédigent directement en langue étrangère, soit qu'ils se traduisent eux-mêmes en réadaptant / réécrivant leurs publications sur de nouveaux supports.

Le petit corpus annoté réuni dans ce cadre compte environ 750 000 mots (anglais et français réunis) et peut être interrogé grâce à l'interface en ligne de ConcQuest[35]. Il s'agit d'un concordancier bilingue qui permet notamment d'intégrer des corpus personnels pour les utilisateurs inscrits, de les étiqueter avec Treetagger (Schmid, 1994) et de les interroger grâce à un langage de requête permettant de rechercher des expressions complexes, à l'instar de CQP (Evert *et al.* 2010). Ce langage, décrit dans Kraif (2008b), permet d'élaborer des méta-expressions régulières combinant des contraintes sur les formes, les lemmes, les étiquettes morphosyntaxiques, et de définir des contraintes syntaxiques du type relation de dépendance (cf. exemple de la p. 129)[36]. Avec la requête : <cat=ADJ> <lemma=recherche>|<cat=ADJ> <lemma=étude>, on obtient par exemple les résultats de la figure 3.5 :

35 Ce corpus, nommé *Emergence*, comporte 685 578 mots dans les deux langues et est interrogeable ici : http://olivier.kraif.u-grenoble3.fr/ConcQuest/concquest.php, consulté en juin 2014

36 Une documentation est disponible ici : http://olivier.kraif.u-grenoble3.fr/index.php?option=com_content&task=view&id=42&Itemid=61, consulté en juin 2014

| | |
|---|---|
| **c1-[JV2-s248,JV2-s249]** However , in Church 's ( 1993 ) case and in several of the other studies mentioned above , the discrepancies between the two parts of the bitext are in fact genuine noise ( due , for example , to poor detection of punctuation in OCR ) . But in many other cases , the discrepancies are structural in nature , due , for instance , to a different way of presenting the documents ( sometimes specifically marked up in representation languages like SGML or XML ) . | **c1-[JV2-s176]** Toutefois , dans le cas de Church ( 1993 ) et de plusieurs **autres études** que nous avons citées précédemment , les divergences entre les deux parties du bitexte sont effectivement un véritable bruit ( par exemple dû à la mauvaise détection des ponctuations par des techniques d' OCR ) , mais dans de nombreux autres cas , les différences sont de nature structurelle , à cause par exemple d' une présentation différente des documents , qui peuvent être parfois explicitement marquées lorsque les documents sont codés dans des langages de représentations tels que SGML ou XML . |
| **c1-[tao2-s44]** Which of these problems would be amenable to short-term solutions and which would have to await the results of long-term research ? | **c1-[tao2-s39]** Lesquels de ces problèmes pourraient être résolus rapidement , et quels autres devraient attendre les résultats d' une **longue recherche** ? |
| **c1-[WHR95-s199]** Recent studies show that children 's lives can be saved by bednets impregnated with insecticides . | **c1-[WHR95-s180]** De **récentes études** montrent que l'on peut sauver des vies d' enfants grâce à des moustiquaires imprégnées d' insecticide . |
| **c1-[WHR95-s609,WHR95-s610]** A recent British study concluded that smoking can kill or cause harm in 24 different ways . The diseases include cancers of the lung , mouth , pharynx , larynx , oesophagus , stomach , intestine , pancreas and bladder , leukaemia , chronic bronchitis and emphysema , pulmonary heart disease , tuberculosis , pneumonia , raised blood pressure , heart disease , blocking of the arteries , brain clots , two forms of brain haemorrhage , sudden rupture of the aorta , stomach ulcers and duodenal ulcers . | **c1-[WHR95-s567]** Une **récente étude** , effectuée au Royaume-Uni , a permis de conclure que le tabagisme peut tuer ou endommager la santé de 24 façons différentes correspondant aux maladies suivantes : cancers du poumon , de la bouche , du pharynx , du larynx , de l' oesophage , de l' estomac , de l' intestin , du pancréas et de la vessie , leucémie , bronchite chronique et emphysème , coeur pulmonaire , tuberculose , pneumonie , hypertension artérielle , cardiopathies , thrombose d' artère , occlusion et sténose d' une artère cérébrale , deux formes d' hémorragie cérébrale , rupture soudaine de l' aorte , ulcère de l' estomac et ulcère duodénal . |
| **c1-[WHR95-s613]** Many studies , including the above , have highlighted the benefits of stopping smoking , even after many years . | **c1-[WHR95-s570]** De **nombreuses études** , y compris celle qui vient d' être mentionnée , ont mis en évidence les avantages d' un arrêt du tabagisme , même après de nombreuses années . |
| **c1-[WHR95-s922]** Other studies deal with the effectiveness of various antibiotic drugs in the treatment of acute respiratory infections in infants and malnourished children . | **c1-[WHR95-s853]** D' **autres études** portent sur l' efficacité de divers antibiotiques pour le traitement des infections respiratoires aiguës chez le nourrisson et les enfants malnutris . |

**Statistiques d'occurrences (1701 segments analysés.)**

Requête <ctag=ADJ> <base=recherche> |<ctag=ADJ> <base=étude> : 71

**Répartition par textes**

- Emergence.en-fr.dat : 71

**Répartition par formes/lemmes**

- **autre étude** : 25
- **premier étude** : 7
- **récent étude** : 7
- **nombreux étude** : 6
- **dernier étude** : 4
- **multiple étude** : 4
- **futur recherche** : 2
- **même recherche** : 2
- **nouveau recherche** : 2

*Figure 3.5 : Exemple de requête bilingue avec ConcQuest*

Une telle requête permet d'identifier les traductions de collocations suivantes :

*further studies ↔ d'autres études, de nouvelles études*
*recent studies ↔ des études récentes*
*long-term research ↔ longue recherche*
*various investigations ↔ diverses recherches*
*many studies, numerous studies ↔ de nombreuses études*
*Initial studies ↔ Les premières études*
*Recent work ↔ Les récentes recherches*
*… etc.*

ConcQuest permet en outre d'effectuer une requête sur les deux langues en même temps. On peut par exemple chercher les couples de phrases ou *study*, *research*, *étude* et *recherche* apparaissent accompagnés d'un adjectif (immédiatement antéposé ou postposé). La requête s'écrit :

*en :<cat=ADJ><lemma=research>|<cat=ADJ><lemma=study>*
*fr :<cat=ADJ><lemma=recherche>|<cat=ADJ><lemma=étude>|*
*<lemma=recherche><cat=ADJ>|<lemma=étude><cat=ADJ>*

Le résultat donne les concordances de ces expressions, puis un récapitulatif des occurrences et cooccurrences pour toutes les expressions trouvées. On en tire divers types d'observations. D'abord on constate que la grande majorité des adjectifs qualifiant les deux noms, en anglais et en français, sont relatifs à la chronologie des études en questions, qui peuvent être antérieure ou postérieure, anciennes ou récentes, en phase de commencement ou achevée, etc. Sur 55 appariements trouvés par ConcQuest, 27 qualifient la temporalité[37] :

*available study_étude disponible : Cooc = 1*
*early study_premier étude : Cooc = 1*
*existing study_étude disponible : Cooc = 1*
*existing study_étude existant : Cooc = 1*
*final study_autre étude : Cooc = 1*
*further research_recherche complémentaire : Cooc = 1*
*further study_autre étude : Cooc = 1*
*further study_étude complémentaire : Cooc = 4*
*future research_futur recherche : Cooc = 1*
*future study_étude ultérieur : Cooc = 1*
*latter study_dernier étude : Cooc = 1*
*modern study_étude moderne : Cooc = 1*
*new study_dernier étude : Cooc = 1*
*old study_étude ancien : Cooc = 2*
*old study_premier étude : Cooc = 1*
*preliminary study_étude préliminaire : Cooc = 2*
*previous study_étude préalable : Cooc = 1*
*prospective study_étude prospectif : Cooc = 6*
*recent study_étude récent : Cooc = 6*
*recent study_récent étude : Cooc = 2*
*subsequent study_étude ultérieur : Cooc = 4*

Concernant la syntaxe, on constate que seuls les adjectifs français *récent* et *futur* se trouvent à la fois antéposés et postposés. Lorsque plusieurs adjectifs sont combinés, on observe que la position n'est pas neutre :

*en :<cat=ADJ><cat=ADJ><lemma=research>|*
*<cat=ADJ><cat=ADJ><lemma=study>*
*fr :<cat=ADJ><lemma=recherche><cat=ADJ>|*
*<cat=ADJ><lemma=étude><cat=ADJ>|*
*<lemma=recherche><cat=ADJ><cat=ADJ>|*
*<lemma=étude><cat=ADJ><cat=ADJ>*

*recent prospective study_dernier étude prospectif : Cooc = 1*
*long-term prospective study_autre étude prospectif : Cooc = 1*

[37] 6 autres cas ont été laissés de côté, car l'expression de recherche ne permettait pas d'apparier les bons adjectifs ensemble. Notons que ConcQuest fournit des appariements lemmatisés.

*large longitudinal study_vaste étude longitudinal : Cooc = 1*
*recent prospective study_étude prospectif récent : Cooc = 1*
*recent molecular study_étude moléculaire récent : Cooc = 1*
*recent genetic study_étude génétique récent : Cooc = 1*
*formal epidemiologic study_étude épidémiologique rigoureux : Cooc = 1*
*analytic epidemiologic study_étude épidémiologique descriptif : Cooc = 1*

Les adjectifs permettant de catégoriser le nom, et qui correspondent à des collocations ou à des termes (qui sont ici le plus souvent des adjectifs relationnels), sont immédiatement antéposés en anglais, ou immédiatement postposés en français, tandis que le second adjectif, moins essentiel sur le plan sémantique, apparaît en première position en anglais, antéposé ou en dernière position en français. Ces faits sont tout à fait conformes aux systèmes syntaxiques des deux langues, et n'ont rien de surprenant : mais ils illustrent de quelle manière le bi-texte permet de « mettre en évidence » ces contrastes. Barlow (2008 : 104), montre comment un logiciel similaire à ConcQuest, ParaConc, permet d'observer la congruence de certaines formes dans les deux langues, et de décrire les équivalences d'un point de vue quantitatif :

> Grâce à des outils d'analyse de corpus il est possible de dépasser les équivalences générales et de donner une vision quantitative des équivalences, ce qui dans une perspective centrée sur l'usage, apparaît comme potentiellement plus intéressant. Pour un mot, une collocation ou une construction dans la langue A, on peut chercher quels sont les équivalents traductionnels les plus communs et de même pour la langue B. Grâce aux données fréquentielles il est possible de dresser une meilleure cartographie des équivalences, et de décrire les équivalents les plus centraux.[38]

Ainsi, sur le plan lexical, on peut noter des tendances plus ou moins marquées concernant les équivalences traductionnelles : dans notre exemple précédent, la traduction la plus fréquente de *further* semble être *complémentaire,* dans l'idée de compléter des études déjà entreprises… Ces observations touchent aussi à la phraséologie : dans le corpus *étude + ultérieur* semble plus fréquent que *étude + futur*, de même qu'en anglais l'équivalent *subsequent + study* est plus fréquent que *future + study*.

Il va de soi que ces observations sont à prendre avec précaution : le mouvement inductif, qui nous fait passer de l'observation du corpus à la langue, générale ou de spécialité, n'est valide qu'à condition d'avoir des observations suffisantes quantitativement sur un

[38] " Using corpus analysis tools it is possible to go beyond general equivalence and give a quantitative view of equivalence, which from a usage perspective is potentially more important. For a word or collocation or construction in language A we can ask what the most common translation equivalents are, and similarly for language B. Using frequency data it is possible to build up a more detailed equivalence map and describe the central translation equivalents. "

corpus à la fois vaste et équilibré pour ne pas être biaisé par des phénomènes « locaux » : phénomènes idiolectaux, choix de traduction, thématique, domaine, etc. Ce n'est évidemment pas le cas du petit corpus duquel nous avons tiré ces exemples : cette étude n'est à prendre qu'en tant qu'illustration d'une certaine méthode d'observation permise par les bi-textes, mais non comme une étude contrastive en tant que telle.

Pour éviter les biais traductionnels, un critère important est la langue source : on peut supposer que les observations de nature idiomatique (collocations, phraséologie, etc.) doivent porter, exclusivement, sur des textes sources. Nous reviendrons plus loin sur cette question (cf. partie 4.1, p. 107).

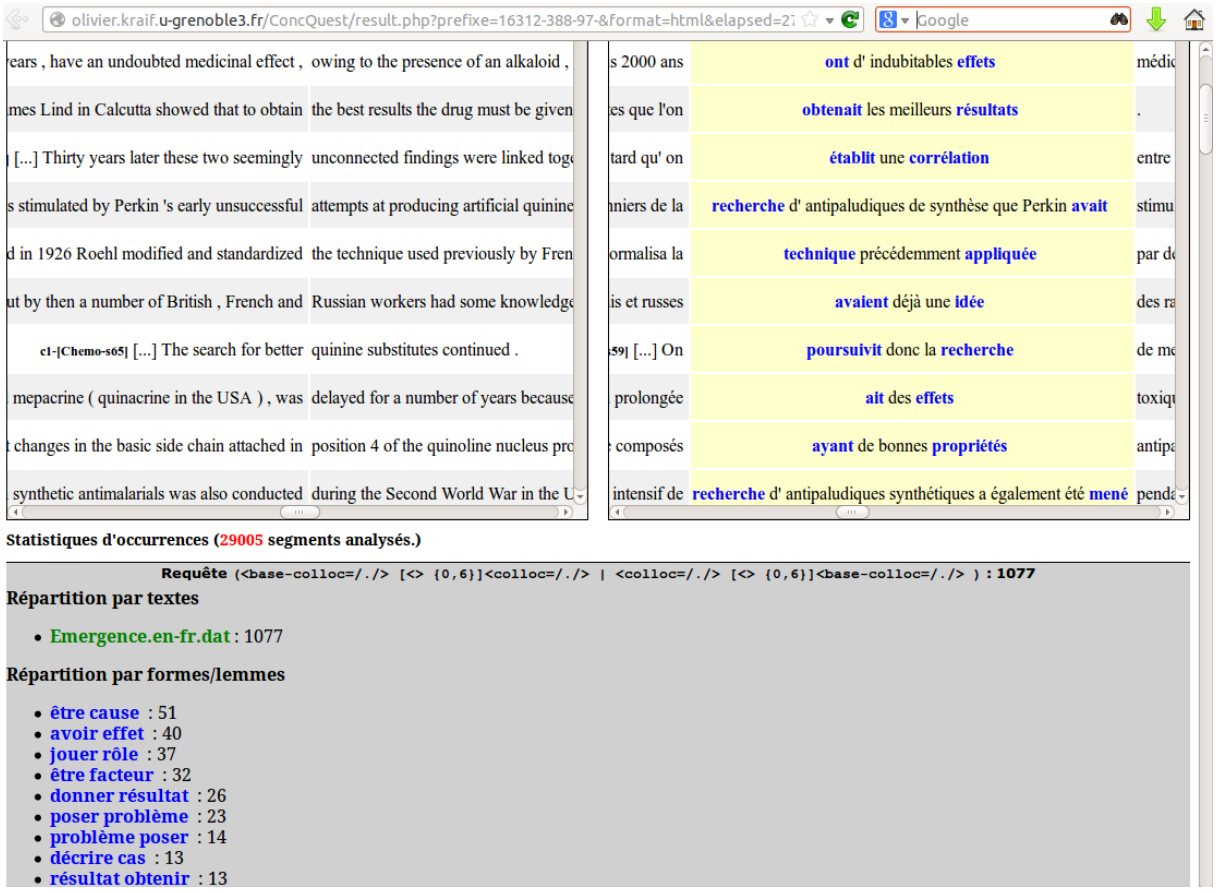


*Figure 3.6 : Résultats de ConcQuest dans l'interrogation du corpus Emergence*

Toujours dans la perspective de l'aide à la rédaction, notons que des collocations transdisciplinaires ont été manuellement annotées dans ce corpus. La figure 3.6 donne un exemple de sortie pour une requête ciblant ces collocations avec un autre format de sortie (KWIC).

## 3.3. De l'aide à la rédaction aux applications didactiques

Entre 2003 et 2007, nous avons concentré nos recherches sur des applications didactiques. Dans Kraif (2004), nous proposions un certain nombre de pistes de recherche pour une application générique des outils du TAL à l'apprentissage des langues assisté par ordinateur (ALAO, ou CALL en anglais). Ces idées ont ensuite été développées avec le projet MIRTO (Antoniadis *et al.*, 2005). En 2006 nous avons collaboré avec des partenaires de Louvain-la-Neuve, notamment Sylviane Granger, pour travailler sur un corpus de productions d'apprenants (FRIDA), et développer un outil d'exploration de ce corpus, baptisé Exxelant (Granger *et al.*, 2007) – ce type de concordancier spécialisé pouvant être utile à l'analyse des erreurs et à la remédiation, comme le montre Rézeau (2007). Par la suite, nous nous sommes plus particulièrement intéressé à l'utilisation de corpus textuels à des fins didactiques, dans la perspective du *Data Driven Learning* (*DDL*), définie par Johns (1991) en ces termes comme « l'utilisation en salle de classe de concordanciers afin que les étudiants explorent les régularités des structures (*patterns*) de la langue cible, et le développement d'activités et d'exercices basées sur les sorties de ces concordanciers. »[39]

Dans son travail de pionnier, Tim Johns (1986) a été le premier à systématiser l'usage du concordancier en classe de langue. L'approche didactique est ici résolument constructiviste, l'apprenant devant prendre une place active dans la construction de ses connaissances, par des activités relativement autonomes. Comme le note Landure (1991 : 166), entre l'approche constructiviste en didactique et le DDL, on remarque des similitudes, notamment « au niveau des rôles ; dans ces deux approches, l'apprenant est défini comme un constructeur actif, un collaborateur et un chercheur et l'enseignant se voit attribuer les rôles de guide, facilitateur et conseiller. » La métaphore du « chercheur » a été proposée d'emblée, et avec certaine audace par Johns (1991) :

> Ce qui est nouveau dans le travail décrit dans cet article, est le parti-pris que « la recherche est une chose trop sérieuse pour être laissée aux seuls chercheurs » : que l'apprenant d'une langue est également, par essence, un chercheur dont l'apprentissage demande à être guidé par l'accès à des données langagières – d'où le terme d'apprentissage guidé par les données (*data driven learning*) pour désigner cette approche.[40]

39 "the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output."

À cette idée de participation active de l'apprenant dans la construction de son savoir, vient s'ajouter la recherche d'authenticité : plutôt que de fabriquer des exemples artificiels destinés à illustrer telle ou telle propriété lexicale ou syntaxique, on préfère rechercher des usages réels dans les textes, en mettant l'accent sur la fréquence observée des phénomènes dans ces usages. Cette approche *corpus driven* est la transposition dans le domaine didactique de la linguistique de corpus telle qu'elle a été défendue par Sinclair (1991), les corpus de référence présentant l'avantage de contenir des données authentiques, complètes, abondantes et neutres vis-à-vis des théories ou systématisations linguistiques (Hunston & Francis, 2000 : 15). Pour l'étude du lexique, une caractéristique fondamentale de cette approche est de ne jamais isoler les mots de leurs contextes. C'est ce qu'illustre la technique de la CPA (*Corpus Pattern Analysis)*, développée par Hanks (2004) en lexicographie, où les entrées lexicales sont décrites en fonction de leur contexte lexico-syntaxique. Par exemple, pour décrire un verbe, on associera ses différents sens à ses différentes constructions (valence) à et ses structures argumentales, en tenant compte aussi des valeurs sémantiques des actants potentiels. Parce que les *unités de sens*, pour reprendre le terme *units of meaning* de Sinclair (1994) n'ont pas toujours de frontière nette, et peuvent mêler à la fois des phénomènes collocationnels, colligationnels, ou des constructions (cf. l'exemple de « *naked eye* » donné par Sinclair, 2004), il nous paraît très intéressant d'embrasser les unités d'abord au sein de leurs contextes.

Dans cette optique, de nombreux travaux ont porté spécifiquement sur l'utilisation de bi-concordances. Par exemple St.John (2001) fournit une étude de cas autour d'activités centrées sur le lexique, pour un étudiant apprenant l'allemand. Cette étude est préliminaire, mais elle entend démontrer que les concordances peuvent être utiles même pour des débutants. Ici, pour réaliser les tâches demandées, l'étudiant sélectionne lui-même ses exemples pour constituer ses propres données. Dans d'autres contexte, les bi-concordances sont préparées par les enseignants et didactisées : les exemples sont sélectionnés, éventuellement annotés, puis intégrés dans diverses tâches. Le tableau 3.2 donne un exemple de bi-concordance didactisée, permettant d'illustrer les différents usages de la préposition *pour :*

40 "What is novel about the work reported in this paper is the perception that "research is too serious to be left to the researchers": that the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data hence the term "data driven learning" (DDL) to describe the approach."

| ***Original text*** | ***Translation*** |
|---|---|
| *1. Ainsi, quand il aperçut **POUR** la première fois mon avion […]* | 1. The first time he saw my aeroplane, for instance […] |
| *2. Alors elle avait forcé sa toux **POUR** lui infliger quand même des remords.* | 2. Then she forced her cough a little more **SO THAT** he should suffer from remorse just the same. |
| *3. -Approche-toi que je te voie mieux, lui dit le roi qui était tout fier d'être enfin roi **POUR** quelqu'un.* | 3. "Approach, so that I may see you better," said the king, who felt consumingly proud of being at last a king **OVER** somebody. |
| *4. Car, **POUR** les vaniteux, les autres hommes sont des admirateurs.* | 4. For, **TO** conceited men, all other men are admirers. |
| *5. C'est comme **POUR** la fleur.* " | 5. It is just as it is **WITH** the flower. |
| *6. C'est donc **POUR** ça encore que j'ai acheté une boîte de couleurs et des crayons.* | 6. It is **FOR THAT PURPOSE**, again, that I have bought a box of paints and some pencils. |
| *7. C'est le même paysage que celui de la page précédente, mais je l'ai dessiné une fois encore **POUR** bien vous le montrer.* | 7. It is the same as that on page 90, but I have drawn it again **TO** impress it on your memory. |
| *8. Elle ferait semblant de mourir **POUR** échapper au ridicule.* | 8. She would […] pretend that she was dying, **TO** avoid being laughed at. |
| *9. et c'était bien commode **POUR** faire chauffer le déjeuner du matin* | 9. and they were very convenient **FOR** heating his breakfast in the morning., |
| *10. Il commença donc par les visiter **POUR** y chercher une occupation et **POUR** s'instruire.* | 10. He began therefore, by visiting them, **IN ORDER TO** add to his knowledge. |
| *11. Il me fallut longtemps **POUR** comprendre d'où il venait.* | 11. It took me a long time **TO** learn where he came from. |
| *12. J'avais le reste du jour **POUR** me reposer, et le reste de la nuit **POUR** dormir…* | 12. I had the rest of the day **FOR** relaxation and the rest of the night **FOR** sleep." |
| *13. **POUR** toi je ne suis qu'un renard semblable à cent mille renards* | 13. **TO** you, I am nothing more than a fox like a hundred thousand other foxes |

*Tableau 3.2 : Un exemple de bi-concordance centrée sur "pour", extraite du* Petit Prince *(Antoine de Saint Exupéry) (Lamy & Klarskov Mortensen, 2012)*

Ces bi-concordances peuvent donner lieu à des activités de classement, de repérage (notamment pour repérer les traductions), voire à des exercices lacunaires, comme l'illustre l'exemple donné par Joseph Rézeau fourni en annexe (Annexe - 1., p. 166). Dans cette activité autour des différentes manières de rendre le pronom *on* en anglais, Rézeau propose une approche plutôt *corpus based* que *corpus driven*, pour reprendre la distinction établie par Tognini-Bonelli (2001) : il commence par énoncer un certain nombre de principes tirés d'une grammaire de Berland-Delépine, puis demande aux étudiants de classer le matériau empirique en fonction de ceux-ci. Dans un deuxième temps, les connaissances liées à cette catégorisation sont réinvesties dans des exercices lacunaires, où l'étudiant doit donner une traduction correcte de *on* en fonction du contexte.

On voit ici comment les approches *corpus-driven* et *corpus-based* sont en fait complémentaires : il peut être difficile pour un apprenant, qui n'est pas linguiste de formation,

de dériver lui-même une classification pertinente à partir des données. Cependant, le fait de confronter les données issues du corpus avec une classification pré-établie, permet de mieux intérioriser celle-ci, du fait de la multiplication des cas, et d'en saisir toutes les implications sur le plan syntaxique, sémantique, idiomatique et fonctionnel.

Prenons l'exemple ci-dessous, tiré de cette même activité :

> *19. - Comment peut-on posséder les étoiles?*
> *"How is it possible for _____ to own the stars ?"*

On voit ici que la traduction de *Comment peut ...* par *How is it possible for…*, qui correspond à des critères de nature idiomatique, impose le choix de *someone* ou *somebody, one* étant impossible pour des raisons syntaxiques (il est pronom sujet seulement). Or tous les exemples donnés illustrent cette propriété de *one*, sans que cette règle ait été explicitée. Le va-et-vient et la comparaison entre les exemples en contexte et l'exercice qui s'ensuit peut aider l'apprenant à intérioriser cette donnée, sans avoir à la formuler consciemment.

Wang (2001) relate une expérience conduite avec des apprenants chinois, là encore centrée sur l'apprentissage du lexique. Citant Rutherford, il souligne qu'un des intérêts du recours aux bi-concordances est de montrer que les langues peuvent recourir à des structures différentes, ce qui permet d'éveiller la conscience métalinguistique des apprenants :

> The main research interest in this paper is in the use of parallel concordancing in the teaching of languages, specifically in its use as a form of consciousness-raising, of making learners aware of the differences between the target language and their own language (Wang, 2001 : 174).

Malgré les difficultés techniques liées au traitement des caractères chinois, l'expérience semble concluante. Étonnamment, l'auteur valorise le caractère exploratoire de l'activité, qui oblige l'enseignant à faire face à l'imprévu :

> The distinctive feature of the Data-driven Learning approach to inductive language teaching is that the language data are primary, and the teacher does not know in advance exactly what rules or patterns the learner will discover.

Notons enfin que l'utilisation des bi-concordances, voire de simples concordances monolingues, n'est pas une panacée, et soulève de nombreuses questions. Chambers (2005), dans une étude qualitative sur l'usage direct de concordances avec des étudiants de licence (*undergraduate*), constate d'importantes variations concernant le style d'apprentissage, la

motivation, l'intérêt porté à ce type de travail en autonomie, les capacités d'analyse et la perception de la nature et des limitations du corpus :

> In addition to the variation in analytical ability, there was also considerable variation in the students' ability to reflect on the nature and limitations of the corpus, an ability which came easily to some students, but was totally lacking in others. (Chambers, 2005 : 119).

Globalement les évaluations des étudiants sont positives : ils apprécient le caractère « authentique », « réel » et « en lien avec l'actualité » ("up to date") du corpus, ce qui peut faciliter la « mémorisation » ; le fait d'avoir de très nombreux exemples, ce qui permet de mieux comprendre les critères utiles pour effectuer certains choix ; et la motivation liée à la découverte par soi-même ("I discovered that achieving results from my concordance was a highly motivating and enriching experience ", ibid. : 120). Mais un certain nombre de critiques sont récurrentes : les concordances ne peuvent se substituer à un manuel de grammaire traditionnel, auquel les étudiants accordent d'ailleurs une plus grande confiance ; pour observer certains phénomènes le corpus est trop limité ; le fait que l'analyse soit souvent ennuyeuse, longue et laborieuse ("tedious, time-consuming, and laborious" , ibid. : 120) ; enfin, le fait que cette approche exige une formation appropriée et des capacités analytiques ("training and appropriate analytical skills", ibid. : 120) dépassant parfois le niveau des étudiants. Chambers (2005 : 122), s'appuyant sur une revue assez complète des évaluations effectuées dans ce domaine, conclut sur le fait que l'usage des concordanciers peut trouver sa place dans la globalité d'un dispositif d'apprentissage, sans se limiter à la salle de classe, car il paraît adapté à des activités autonomes ou collaboratives.

La technicité des outils et l'aspect expérimental des méthodes didactiques ont jusqu'à présent freiné le développement de ces approches originales, qui restent assez marginales dans les pratiques pédagogiques, tout spécialement en France : mais nous croyons qu'elles sont appelées à se développer avec l'évolution des pratiques, notamment du fait de la place importante promise au numérique. Nous reviendrons sur ces aspects dans la partie 4 de ce travail.

## 3.4. Vers une cartographie sémantique ?

L'étude des corpus parallèles permet d'identifier des séries d'équivalences dont certaines, ont l'a vu, sont généralisables, c'est-à-dire peuvent être réutilisées dans de

nombreux contextes. Les équivalences ainsi identifiées constituent un réseau de correspondances qui peuvent mettre en lumière des propriétés sémantiques, comme nous le suggérions dans Kraif (2003a).

Par exemple, la polysémie d'une unité en langue source peut être manifestée par sa mise en correspondance avec des unités cibles appartenant à des champs sémantiques différents : l'italien *carta* sera souvent associé à *papier* et à *carte*, amorçant ainsi la structuration de la signification en deux acceptions principales dont une désigne un /matériau/ l'autre un /support d'inscription/. Par suite, la confrontation avec l'anglais permet d'enrichir cette décomposition du sens : *carta* est souvent associé à *paper*, *card* ou *map*. Une troisième distinction apparaît, entre /document topographique/ et /petit support rectangulaire/ (correspondant aux cartes à jouer, cartes de visite, cartes de crédit, etc.). On pourrait rétorquer que de telles relations ne nous permettent pas de distinguer entre la polysémie de *carta*, ou l'éventuelle synonymie de *paper*, *map* et *card*. Mais si l'on tient compte des correspondances de *paper*, *map* et *card*, dans d'autres langues, on obtiendra le plus souvent des équivalents différents (comme *papier*, *plan*, *carte*), ce qui permet d'affaiblir l'hypothèse de synonymie. Il est également possible de différencier polysémie et homonymie : dans la mesure où les liens polysémiques sont en partie motivés, il est fréquent de retrouver des polysémies parallèles (au moins partiellement) dans d'autres langues. Par exemple, les deux acceptions /document topographique/ et /petit support rectangulaire/ se retrouvent aussi bien dans le français *carte* que dans l'italien *carta*. Si ces deux sens correspondaient à des unités différentes homonymes, il serait étonnant que l'homonymie s'observe aussi bien en français qu'en italien, car l'homonymie est par définition fortuite (à la différence de la polysémie). Le schéma du tableau ci-dessous montre comment le repérage de traduction permet de structurer les significations, à la manière de Hjelmslev (1971 : 113) lorsqu'il comparait la distribution de *bois* avec l'allemand *Holz* et *Wald*, et le danois *trae* et *skov*.

| **Italien** | **Français** | **Anglais** |
|---|---|---|
| carta | papier | paper |
| | carte | map |
| | | card |

*Tableau 3.3 : Unités équivalentes à l'italien* carta

Comme le montre le schéma de la figure 3.7, on peut observer des configurations complexes qui mettent en relation des niveaux distincts :

– entre les langues : on constate par exemple que *paper* partage de nombreuses acceptions avec *papier* (ce qui pourrait indiquer des significations voisines), malgré quelques différences ;

– entre chaque langue et les *designata*[41] extra-linguistiques : cette relation, bien qu'invisible à l'intérieur des textes, peut être reconstruite grâce à certaines convergences (par exemple, lorsque *paper* est associé avec *article*, le *designatum* correspondant à 'article de presse' ou 'article scientifique' peut être déduit sans équivoque) ;

– entre les unités d'une même langue : on constate la possible synonymie de *papier* avec *article*, mais aussi la divergence de leurs autres acceptions ;

– entre les acceptions d'une même unité : la polysémie de *paper* ou de *papier* devient manifeste du fait de leurs multiples possibilités de traduction.

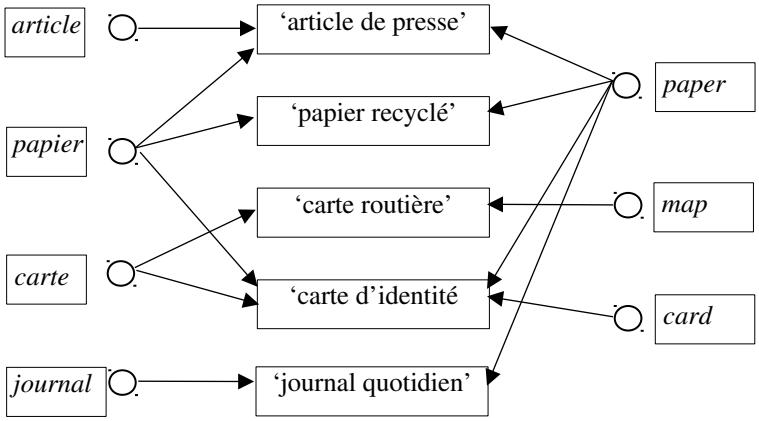


*Figure 3.7 : Réseau de relations interlingues manifestant les structurations sémantiques de chaque langue*

Comme le note Pergnier (1993 : 84), « la confrontation de signes appartenant à deux langues différentes révèle à la fois la polysémie de chacun (c'est-à-dire la diversité interne de leurs signifiés considérés du point de vue des concepts désignés) et la non-coïncidence de ces signifiés, c'est-à-dire le fait qu'ils sont polysémiques *différemment*. » La correspondance du

41 Par *designatum*, nous entendons l'objet extra-linguistique pointé par le signe linguistique, quelle qu'en soit sa nature (référent, classe d'objets, concept, représentation mentale, etc.). Pour éviter la confusion avec les /acceptions/, appartenant à la signification linguistique, nous noterons les *designata* par 'une glose entre guillemets simples'.

français *disque* avec l'anglais *record* n'est valable que dans certains contextes, dans la mesure où les deux unités sont toutes deux polysémiques mais véhiculent des significations différentes : par exemple, en référence à un CD, *disque* indique la forme de l'objet, tandis que *record* s'attache à la fonction d'enregistrement. Chaque langue s'attache à des traits référentiels arbitrairement choisis, et le repérage de traduction permet d'objectiver les différences de choix. L'organisation particulière de chaque « système classificateur », selon l'expression de (Pergnier 1993 : 109), devient ainsi manifeste : « Au niveau de son organisation interne, on pourrait dire que le signifié saisit les choses qu'il désigne non par leurs différences, mais par leurs ressemblances. Le passage de l'anglais au français n'a pas seulement pour effet de changer le signifiant ; il a pour effet de le faire changer de système classificateur. »

Ainsi, les correspondances lexicales permettent de comparer les codes, et de faire apparaître, pour chacun d'eux, des structurations sémantiques mises en lumière par la non-congruence des modes de désignation. Mais ce qui se dessine à travers cet entrelacs de liens interlinguistiques est extérieur aux codes eux-mêmes : ce sont les *designata* extra-linguistiques qui apparaissent en filigrane, puisqu'ils constituent – souvent – le pivot de la relation d'équivalence traductionnelle. Considérons l'énoncé italien : *Questa carta è vecchia*. L'ensemble des *designata* potentiels de *carta* est très étendu : 'papier', 'tapisserie', 'carte de crédit', 'carte à jouer', 'carte routière', 'carte de crédit', 'carte de visite'. Mais si l'on est en présence des traductions suivantes, l'ensemble des *designata* potentiels se réduit considérablement : *C'est une vieille carte* / *This is an old card* / *Esse bilhete e velho*. À l'intersection de toutes ces formulations linguistiques, toutes ambiguës si on les considère séparément, on trouve un *designatum* restreint, correspondant à 'carte d'identité'. Comme le proposaient déjà Dagan *et al.* (1991), il est donc envisageable d'élaborer des méthodes de désambiguïsation sémantique tirant parti des correspondances interlingues. Par exemple, Diab & Resnik (2002) décrivent une méthode de désambiguïsation non supervisée faisant intervenir un corpus multilingue traduit automatiquement.

### 3.4.1 Désambiguïsation lexicale

Pour notre part, nous avons tenté, lors de la campagne Senseval3, une approche basée sur la constitution de classes de synonymes (Moreau de Montcheuil *et al.*, 2004), obtenues à partir d'un corpus bilingue aligné (le corpus du projet Carmel). Les collègues du LIA avec qui

je travaillais alors, ont d'abord implémenté un système de désambiguïsation classique, puisant les indices de désambiguïsation dans le contexte proche des exemples d'apprentissage. Ce système résultait de la combinaison de 3 algorithmes : un arbre de classification sémantique, la méthode des *k* plus proches voisins et un modèle probabiliste basé sur la loi de Poisson. Par la suite, nous avons extrait des classes de synonymes en nous appuyant sur le corpus multilingue du projet Carmel : partant de ces classes, nous voulions généraliser et régulariser les contextes d'apprentissage afin d'améliorer la précision du système. Mais la constitution de ces classes n'a finalement pas permis d'améliorer les résultats : il s'est avéré que le corpus multilingue utilisé, constitué de récits de voyages assez anciens (XIXe et début XXe), était trop petit et inadapté pour cette tâche (par ailleurs, l'extraction de correspondances lexicales effectuée comportait trop de bruit, du fait de la trop petite taille du corpus).

Par la suite, afin de neutraliser les effets liés aux erreurs d'alignement et à l'inadéquation du corpus, nous avons cherché à identifier le gain d'une méthode de désambiguïsation basée sur les correspondances lexicales en nous basant préalablement sur une annotation manuelle (Haddara & Kraif, 2005). Des couples de mots ambigus (75 couples de noms, d'adjectifs et d'adverbes) ont été sélectionnés de manière aléatoire dans notre corpus, constitué de *The voyage of the Beagle*, de Darwin, et d'une traduction française de 1875. Le texte comporte environ 200 000 mots dans chaque langue, et a été aligné au niveau phrastique avec Alinéa. Les correspondances lexicales ont été extraites en utilisant *MotAMot*, développé par nous-même en collaboration avec Boxing Chen (Chen & Kraif, 2004). L'annotation s'est alors déroulée en deux étapes principales.

Tout d'abord, les unités ont été désambiguïsées manuellement sans recours au contexte. Pour ce faire, l'annotateur s'est contenté des informations fournies par l'appariement des unités équivalentes de notre corpus, et des listes de sens fournies par le dictionnaire. La tâche consistait donc à comparer les différents sens proposés des unités appariées afin de retenir les couples de sens qui semblaient les plus proches.

Les sens retenus ont ensuite été évalués en examinant les contextes séparément dans chaque langue. Les résultats figurent dans le tableau ci-dessous :

| Couples extraits de *The voyage of the Beagle*, de Darwin, et *Au Maroc*, de Loti | NOM | | ADJ | | ADV | |
|---|---|---|---|---|---|---|
| | % En | % Fr | % En | % Fr | % En | % Fr |
| Proportion d'unités totalement désambiguïsées | 28 | 35 | 32 | 19 | 21 | 36 |
| Précision (unités totalement désambiguïsées) | 100 | 100 | 83 | 75 | 100 | 100 |
| Proportion moyenne de sens éliminés | 42 | 38 | 35 | 27 | 33 | 35 |
| Précision globale | 96 | 96 | 84 | 88 | 100 | 100 |

*Tableau 3.4 : Résultat de la désambiguïsation bilingue manuelle*

D'après ces résultats, l'apport d'information d'une langue sur l'autre semble être plus déterminant pour les noms avec une réduction d'ambiguïté d'environ 40% pour l'anglais et le français, contre 31% et 34% respectivement pour les adjectifs et les adverbes. La précision indique la proportion des sens retenus après examen de l'équivalent traductionnel qui correspondent avec le(s) sens retenu(s) à la fin du processus de comparaison. Même si la méthode ne permet de désambiguïser complètement qu'une partie des unités ambiguës (entre 20 et 35%) la précision obtenue est très bonne (supérieure à 90% pour les noms et les adverbes).

Nous avons examiné si des critères linguistiques permettent à priori d'identifier les configurations les plus favorables – ou défavorables – pour ce type de désambiguïsation. Par exemple, on pourrait supposer que deux unités apparentées sont moins susceptibles de se désambiguïser mutuellement. Mais ce critère ne résiste pas à l'examen. On trouve par exemple les appariements *companion (EN)* ↔ *compagnon (FR)* et *region (EN)* ↔ *région (FR)*, qui sont aisément reconnaissables comme cognats. Dans l'exemple ci-dessous, on voit comment la comparaison des 5 sens liés à chaque unité pour *companion* et *compagnon* aboutit à une désambiguïsation complète et correcte puisqu'il y a un seul couple de sens compatibles (i.e. /ami/).

**Companion**
1 (friend) *compagnon/compagne* m/f; to be sb's constant companion [hunger, fear] *être le perpétuel compagnon de qn*; a companion in arms *un compagnon d'armes*;
2 (also paid companion) *dame f de compagnie*;
4 literature, publishing guide m; the fisherman's companion *le guide du pêcheur*;
5 nautical *capot* m.

**Compagnon**
1 (ami) *companion*; compagnon fidèle *faithful companion*;
2 (amant) *partner*;
3 (mâle) *mate*;
4 (artisan) *journeyman*;
5 (franc-maçon) *fellow of the craft*.

À l'opposé, les différents sens des unités *region* et *région* sont très proches, donnant lieu à plusieurs couples de sens possibles et donc à une désambiguïsation quasi-nulle.

Le multi-texte peut néanmoins donner des indices intéressants sur le pouvoir de désambiguïsation d'un mot sur un autre : il suffit d'examiner les appariements de ces deux mots dans une langue tierce. Par exemple, pour *scarcely* et *presque* on trouve dans le même corpus les appariements suivant avec l'espagnol (en considérant chaque mot du couple, indépendamment) :

*scarcely* → *tampoco, asegurar, casi, apenas*

*presque* → *casi*

Un simple filtrage des fréquences nous permet d'éliminer les alignements erronés tel que *asegurar*. On constate alors que les trois sens de *scarcely* indiqués par le dictionnaire, que l'on pourrait gloser par /presque pas/, /difficilement/, /à peine (sens temporel)/, se manifestent par des équivalents espagnols plus variés – et on voit assez clairement comment l'appariement avec *presque* (ou plus exactement *presque pas*) permet d'effectuer la désambiguïsation. On trouve ici une confirmation de notre idée initiale : chaque langue apporte une information supplémentaire, et le faisceau des correspondances interne au multitexte s'enrichit et s'affine avec l'ajout de nouvelles langues. Nous appelons *triangulation* ce type de désambiguïsation passant par la mise en correspondance avec une langue tierce.

Pour prédire s'il est judicieux ou non d'employer la triangulation, pour un couple donné, nous proposons de recourir à un critère numérique, comme l'indice de DICE : $s = \frac{2 \cdot card(ES(e) \cap ES(f))}{(card(ES(e)) + card(ES(f)))}$, où ES(e) et ES(f) représentent les ensembles d'équivalents dans la langue tierce (précédemment l'espagnol) dérivés de l'alignement pour les unités *e* et *f*. Calculé sur un corpus trilingue suffisamment important, nous pensons que *s*

peut être un bon indicateur de la similarité sémantique de deux unités : une valeur faible devrait indiquer de meilleures chances de désambiguïsation multilingue.

Enfin, pour valider d'une autre manière notre hypothèse de désambiguïsation sémantique par les équivalents traductionnels, nous avons mis en œuvre une méthode de désambiguïsation sémantique non supervisée. Comme Tufis *et al.* (2004), nous avons utilisé deux réseaux sémantiques (les lexiques français et anglais livrés avec EuroWordNet, à savoir FrWN et WordNet 1.5, cf. Vossen, 1998), afin de comparer les unités par le biais d'index interlingue (ILI), qui permettent d'établir des équivalences de sens entre des unités. L'algorithme de comparaison peut être décrit de la manière suivante :

*Pour* chaque paire ($U_s$, $U_c$) d'unités alignées {
    $S_s \leftarrow$ {ensemble des sens candidats pour Us}
    $S_c \leftarrow$ {ensemble des sens candidats pour Uc}
    SimMax $\leftarrow$ 0
    *Pour* chaque paire $(s_s, s_c) \in S_s \times S_c$ {
        calculer $Sim(s_s, s_c)$.
        *Si* $(Sim(s_s, s_c) > SimMax)$ *Alors* $SimMax \leftarrow Sim(s_s, s_c)$.
    }
    Enregistrer $Desamb(U_s, U_c) = \{ (s_s, s_c) \in S_s \times S_c \,/\, Sim(s_s, s_c) = SimMax \}$

$Sim(s_s,s_c)$ est une mesure de similarité des sens pouvant être calculée à partir du nombre de liens séparant chacun des ILI de leur plus proche parent commun dans la hiérarchie. Dans l'expérience ici décrite, afin de privilégier la précision (au détriment du rappel) et de s'approcher de la désambiguïsation manuelle effectuée précédemment, nous avons utilisé une définition maximaliste de la similarité, basée sur l'identité des ILI (donc avec une similarité de 0 si les ILI diffèrent, et de 1 sinon). Le corpus utilisé est à nouveau *The voyage of the Beagle* de Darwin. Pour qu'un couple d'unités soit partiellement ou complètement désambiguïsé, il faut que les deux unités alignées apparaissent chacune dans leurs réseaux respectifs. Seulement 21 133 couples de mots ont satisfait cette condition.

| | Anglais | Français |
|---|---|---|
| Proportion moyenne de sens éliminés | 63 % | 46 % |
| Unités totalement désambiguïsées | 34,6 % (7 316 / 21 133) | 22,7 % (4 804 / 21 133) |
| Précision estimée | 79 % | |

*Tableau 3.5 : Réduction des sens pour une méthode de désambiguïsation non supervisée*

Le tableau ci-dessus indique les résultats pour les 21 133 couples qui apparaissaient dans EWN. Les couples totalement désambiguïsés représentent environ 4,3 % de la totalité des mots, et 42 % des couples pour lesquels les deux réseaux n'étaient pas silencieux[42]. Pour estimer la précision des résultats, nous avons effectué un prélèvement aléatoire de 100 couples totalement désambiguïsés, qui ont été évalués manuellement par un seul annotateur. Par manque de moyens, seul l'anglais a été évalué. Ces résultats ne peuvent être comparés à ceux du tableau 3.4 sans précaution, car ils dépendent fortement de la couverture du réseau EWN. Entre autre, la proportion moyenne des sens éliminés est beaucoup plus importante : dans la mesure où elle dépend directement de l'identité des index interlingues (ILI) pour les unités comparées, certains sens ont été simplement éliminés du fait de l'incomplétude et du déséquilibre des réseaux (le réseaux français FrWN contient 22 745 sens contre 91 600 pour Wordnet 1.5). Notons néanmoins que la précision s'est maintenue à un bon niveau pour les couples totalement désambiguïsés.

Par ailleurs, on observe une certaine corrélation entre la similarité *s* obtenue par projection sur l'espagnol et la proportion de sens éliminés, ce qui confirme l'hypothèse de triangulation formulée précédemment.

| % sens éliminés | $0 \leq s < 0,25$ | $0,25 \leq s < 0,5$ | $0,5 \leq s < 0,75$ | $0,75 \leq s \leq 1$ |
|---|---|---|---|---|
| Anglais | 75 % | 65 % | 62 % | 60 % |
| Français | 60 % | 49 % | 43 % | 40 % |

*Tableau 3.6 : Corrélation entre s et la proportion des sens éliminés*

### 3.4.2 Construction d'une ressource multilingue de type WordNet pour l'arabe

Comme nous le notions dans un article publié en collaboration avec Authoul Abdulhay, la doctorante que j'ai encadrée (Abdulhay & Kraif, 2013) :

> Le réseau sémantique WordNet (Fellbaum, 1998) de l'université de Princeton est devenu un standard *de facto*, malgré certaines limites et imperfections qu'on peut lui reprocher, telles que ses incohérences, la confusion entre sens et concept ou l'inadéquation de son organisation des sens à d'autres langues que l'anglais (Mallak, 2011).

[42] Notons que les 34,6 % et 22,7 % du tableau 3.5 correspondent à deux ensemble de couples différents, dont l'intersection est relativement petite, car bien souvent seule l'unité en français ou en anglais était considérée comme ambiguë, ce qui explique qu'on obtienne au final un pourcentage de 42 % de couples désambiguïsés.

Une propriété intéressante de l'architecture de WordNet est son mode de représentation des sens : ceux-ci correspondent à des *synsets* – littéralement des ensemble de synonymes – en fait des groupes d'unités lexicales qui définissent en quelque sorte une acception par leur intersection.


*Figure 3.8 : Exemple de synsets de Princeton Wordnet (PWN) pour l'entrée* situation

À chaque *synset* est également lié une glose qui permet de préciser le sens. Notons que ce qui est visé ici, c'est un sens extra-linguistique (d'ailleurs nommé *concept*) : les unités lexicales se regroupent dans un synset parce qu'elles peuvent, potentiellement, *traduire* ce sens dans certains contextes. Ce qui relie les unités dans un même synset, c'est en quelque sorte une relation d'équivalence assez voisine de l'équivalence traductionnelle – le fait d'avoir une intersection sur le plan des *designata*.

Implicitement, les synsets définissent donc des liens de synonymie (pour des unités appartenant au même synset) et de polysémie, lorsqu'une même unité appartient à des synsets différents.

Or on a vu qu'il est possible, à partir des correspondances lexicales extraites d'un multitexte, d'obtenir une structuration assez similaire des unités : une unité fortement polysémique aura tendance à avoir des équivalents variés, et des unités voisines sémantiquement (c'est-à-dire en relation de synonymie partielle) partageront vraisemblablement les mêmes équivalents traductionnels.

La thèse d'Authoul Abdulhay s'appuie sur cette idée : utiliser un multitexte français, anglais, espagnol et arabe pour en extraire des synsets pertinents pour l'arabe, et mettre en relation ces synsets avec ceux de PWN (Abdulhay, 2012). L'établissement de ces relations permettant, dans un deuxième temps, de projeter certaines informations sémantiques, telles que les liens d'hyperonymie ou de méronymie, de l'anglais vers l'arabe.

Sagot et Fišer (2008) ont proposé une démarche similaire, pour constituer le Wolf, un réseau sémantique pour le français produit par extension de PWN (Vossen, 2008), i.e. en en traduisant les synsets. En partant d'un mot de PWN, ils considèrent deux cas de figure. Si ce mot est « monosémique » (i.e. rattaché à un seul synset), sa traduction en français est considérée comme triviale, et elle est effectuée via des lexiques bilingues (tirés de Wikipedia et du thésaurus EUROVOC20). Pour les mots polysémiques, les auteurs se basent sur une idée très voisine de celle que nous avons précédemment formulée :

> *Les différents sens des mots ambigus dans une langue donnée donnent souvent lieu à des traductions différentes dans une autre langue. À l'inverse, nous supposons que si deux mots ou plus sont traduits par le même mot dans une autre langue, ils partagent souvent un élément de sens. En outre, ces phénomènes sont renforcés par l'utilisation de plus de deux langues, d'où l'intérêt d'une approche par alignement multilingue.* (Sagot et Fišer 2008 :3)

Pour traiter les mots ambigus, ils utilisent le corpus parallèle CCR-Acquis19 comportant 5 langues alignées. Chaque mot simple français est mis en correspondance avec des équivalents en anglais, roumain, tchèque et bulgare, rattachables à un ou plusieurs ILI dans EWN et BalkaNet (les ILI, pour *Inter Lingual Index*, sont des identifiants numériques renvoyant à différents sens listés par PWN, éventuellement complétés par des sens supplémentaires pour traiter d'autres groupes de langues). En prenant l'intersection de ces ILI, les auteurs peuvent sélectionner le ou les sens rattachables au mot français, et compléter leur traduction des synsets de PWN. En utilisant EWN comme référence, les auteurs

obtiennent respectivement pour les noms et les verbes une précision de 77,2% et 65,8%, et un rappel de 68,7% et 54,7%.

L'approche par extension est selon nous assez contestable, car elle présuppose que l'organisation des sens de la langue cible soit isomorphe à PWN. De fait, bien que l'organisation des sens de PWN soit en principe indépendante de l'anglais (vu que les sens sont censés se baser sur un substrat extralinguistique), elle est fortement influencée par celle-ci, les sens qui y sont codés n'ayant rien de concepts universels. La meilleure preuve de cet enracinement dans la langue anglaise est l'organisation des sens en fonction des 4 parties du discours : noms, verbes, adverbes et adjectifs. Ces catégories ne représentent en rien des universaux – et l'arabe, par exemple, ne connaît que 3 catégories principales : noms, verbes et particules.

Pour notre part, dans le but de constituer des synsets pour l'arabe, nous nous sommes basés sur un modèle géométrique inspiré des atlas sémantiques de Ploux (2007). En s'appuyant sur des liens fournis par des dictionnaires de synonymes, celle-ci extrait des cliques, c'est-à-dire des ensembles de synonymes à l'intérieur desquels toutes les unités sont reliées entre elles (la clique étant définie formellement comme un graphe complet connexe). Les cliques maximales (c'est-à-dire qui ne sont pas incluses dans une clique de plus grande dimension) correspondent selon l'auteure à un découpage assez fin en sous-sens. En utilisant un dictionnaire bilingue, et une méthode de projection dans un espace sémantique commun (Ploux et Ji, 2003), l'auteure montre comment les cliques obtenues dans chaque langue peuvent être appariées, ce qui permet d'enrichir à la fois le dictionnaire bilingue, et d'identifier de nouveaux candidats synonymes dans chaque langue.

L'approche présentée par Abdulhay et Kraif (2013) est très voisine de ces travaux par sa représentation géométrique du sens. Mais c'est à partir de corpus multilingues parallèles, et non de dictionnaires, que nous avons cherché à extraire ce type de cliques. Ainsi, les cliques sont obtenues sur la base des correspondances lexicales et sont, par construction, multilingues.

Le corpus utilisé provient des archives des Nations Unies[43]. Les 185 textes téléchargés sont des rapports traitant de sujets divers (santé, commerce, droits des femmes, …), en 4 langues : anglais, arabe, espagnol et français. Ils ont été alignés avec Alinéa, étiquetés et lemmatisés avec Treetagger, sauf l'arabe qui a été étiqueté sans lemmatisation avec Amira 1.0

43 Téléchargé en 2008 depuis le site http://unbisnet.un.org

(Diab *et al.*, 2007). Après élimination des mauvais alignements et extraction des correspondances avec Giza++ (Och et Ney, 2003), on obtient entre 73 823 (fr-ar) et 98 303 (en-es) paires de mots.

On obtient par suite des cliques maximales de ce type :

> *(fr-Noun-question ar-Noun-msOlp en-Noun-matter es-Noun-cuestión en-Noun-issue)*[44]

Dans ce genre de clique multilingue, on suppose que la relation d'équivalence deux à deux correspond en fait à une relation *n* à *n*, liée à une intersection sémantique non nulle. C'est ce qu'on a nommé l'hypothèse de centralité des cliques. En d'autres termes, si on considère qu'il existe un sens commun au français *question* et à l'anglais *matter*, et un sens commun à l'anglais *matter* et à l'espagnol *cuestión*, on peut supposer que le même sens est également commun à *question* et *cuestión*. En revanche, pour les unités d'une même langue, comme ici *matter* et *issue*, deux cas de figure sont envisageables : soit ils ont des *designata* communs, et sont donc synonymes, soit ils correspondent à deux acceptions différentes du français *question* ou de l'espagnol *cuestión* – ils pourraient par exemple correspondre à un découpage plus fin en anglais être cohyponymes. Dans ce cas précis, nous penchons plutôt pour la synonymie, étant entendu que la relation de synonymie est le plus souvent très parcellaire.

Dans cette expérimentation, le corpus étant relativement modeste, nous avons dû procéder à une étape de clusterisation des cliques les plus ressemblantes, certains liens d'équivalence étant absents du fait du manque de données.

Dans un deuxième temps, les unités arables présentes dans les cliques ont été reliées aux synsets d'EuroWordNet, en appliquant un principe de clôture transitive : si toutes les unités d'une même clique partagent un et un seul sens d'EuroWordNet (via les ILI) alors la clique est désambiguïsée et on rattache l(es) unité(s) arabe(s) à ce sens commun. Par exemple, dans la clique (*EN-N-science FR-N-science ES-N-ciencia ar-N-* علم) les lexèmes anglais, français et espagnol sont tous les trois rattachés à un seul ILI glosé par */a particular branch of scientific knowledge*/. On peut donc également lui rattacher l'unité arabe, car il n'y a pas d'ambiguïté.

---

44 Le mot arabe est ici représenté en translittération ASCII Buckwalter.

Par ailleurs, on a également cherché à projeter sur l'arabe des relations sémantiques d'EWN : si deux cliques ont chacune été rattachées à un seul ILI, respectivement, et si pour une langue donnée il existe une relation sémantique entre deux unités appartenant à ces deux cliques, pour une acception liée au ILI retenu, alors la relation peut être étendue pour les unités arabes contenues dans ces cliques, sauf si une relation contradictoire peut être inférée à partir d'une autre paire de lexèmes.

Par exemple si on considère les deux cliques suivantes :: (*ar-N-* قسم *FR-N-fragment EN-N-snippet ES-N-recorte*) et (*ar-N-*حصة *FR-N-morceau ES-N-pedazo EN-N-piece*), sachant qu'on a une relation *'has_hyperonym'* entre EN-N-*snippet et EN-N-piece*, et qu'il n'existe pas de relation différente pour les unités des autres langues (il se trouve qu'on a la même relation pour le français et l'espagnol, même si ce n'est pas une condition nécessaire ici), on peut étendre la relation aux unités arabes *ar-N-* قسم et *ar-N-*حصة .

Une évaluation a été faite sur un échantillon de 200 clusters de cliques, pour les noms et pour les verbes. Dans un premier temps, on a évalué la proportion de clusters valides, c'est-à-dire reliés, de façon cohérente entre l'anglais, le français et l'espagnol, à un ou plusieurs ILI. Dans un second temps, on a évalué la validité des sens rattachés aux unités arabes des clusters : validité complète ou partielle – dans le cas où le sens est voisin (plus général ou plus spécifique) selon notre dictionnaire de référence (*Alwaseet*)

| | Nom | Verbe |
|---|---|---|
| Nb clusters traités | 100 | 100 |
| Nb clusters valides (désambiguïsés et non-désambiguïsés) | 56 | 29 |
| Nb lemmes arabes dans les clusters désambigüisés | 74 | 37 |
| Nb lemmes validés complètement (VC) | 59 | 21 |
| Nb lemmes validés partiellement (VP) | 8 | 6 |
| Nb lemmes non validés | 7 | 10 |
| Nb Total d'unités arabes validés (VC+VP) | 94 / 111<br>≈ 84,7% | |

On constate que les résultats sont assez bons en terme de précision, mais bien meilleurs pour les noms que pour les verbes. Parmi les rattachements valides, on trouve par exemple, pour le lemme arabe علم, deux clusters[45]:

*(ar-N-العلوم ar-N-العلم EN-N-science FR-N-science ES-N-ciencia)*
*(ar-N-علم ar-N-تعلم FR-N-apprentissage EN-N-learning ES-N-aprendizaje)*

Ces deux cliques correspondent bien à deux acceptions attestées par *Alwaseet*, glosées par « un groupe de connaissances scientifiques dans un domaine particulier » et « l'acquisition et la connaissance de la vérité des choses » (traduction de A. Abdulhay).

Globalement les résultats héritent des limitations des ressources mises en œuvre, autant au niveau du corpus, qui est trop petit pour réaliser toutes les virtualités sémantiques des unités, qu'au niveau des *wordnets* :

– de nombreux lexèmes, même courants, sont absents d'EWN : par exemple, dans FrWN, on ne trouve pas les verbes *adjoindre, s'approprier, figurer, spécialiser, ...*

– le rattachement aux sens de PWN est lacunaire : par exemple, la clique (*ar-N-*فلسفة *ES-N-filosofía FR-N-philosophie EN-N-philosophy*) est considérée comme monosémique, car FrWN ne retient qu'une seule acception pour *philosophie*, ce qui n'est pas le cas pour les autres langues. Pourtant, *philosophie* est bien polysémique en français (p.ex. en tant que synonyme de *flegme*).

– le découpage des sens, comme nous l'avons vu, est spécifique à l'anglais et présente de ce fait une part d'arbitraire. Par exemple, pour l'ensemble (*EN-N-fund FR-N-fonds ES-N-fondo*), on constate que les unités prises deux à deux partagent des sens, mais aucun de ces sens n'est commun au trois (ce qui contredit notre hypothèse de centralité des cliques). Voici les 3 sens en question :

- EN-N-fund ET FR-N-fonds : */a reserve of money set aside for some purpose & 03 1stOrderEntity 21 Artifact Function Money Representation Origin Possession/.*
- EN-N-fund ET ES-N-fondo : */a supply of something available for future use & 03 1stOrderEntity 21 Function Possession/.*
- FR-N-fonds ET ES-N-fondo : */assets in the form of money & 03 1stOrderEntity 21 Function Possession/*

[45] Toutes les unités en arabe *ar-N-*العلوم *ar-N-*العلم *ar-N-*علم *ar-N-*تعلم sont des formes fléchies de ce même lemme.

Ces variations de granularité dans le découpage des sens et leur rattachement apparaissent comme plutôt arbitraires. Ceci dit, on pourrait très bien trouver des contextes ou *fund*, *fonds* et *fondo* apparaissent comme équivalents, c'est-à-dire comme désignant la même chose (p.ex. *des fonds de pension*).

Par ailleurs, il faut noter que de nombreuses erreurs sont dues à la non reconnaissance des unités polylexicales. Considérons le cluster suivant : (*ar-N-*لغة *FR-N-langue EN-N-language ES-N-idioma FR-N-linguistique*). L'unité française *FR-N-linguistique* qui est monosémique (dans FrWN) et qui appartient à un *synset* totalement différent de celui de *fr-Noun-langue* a comme ILI : */the scientific study of language/.* Cette erreur est probablement liée à des appariements non reconnus entre des unités polylexicales (*language study*→ *linguistique*) ou à l'ambiguïté morphologique (p.ex. *language research*→*recherche linguistique*), l'adjectif *linguistique* étant par erreur étiqueté comme un nom.

Pour l'échantillon évalué, voici la répartition des causes de non-rattachement pour les noms (pour un total de 44%) :

- Insuffisance de couverture des WNs 18%
- Pas d'ILI commun à toutes les unités (unités polylexicales mal reconnues, mauvaises clusterisation, problème de découpage des sens) 9%
- Cliques ambiguës du fait de polysémies parallèles 17%

Quant aux verbes, les faibles résultats correspondent à la répartition suivante :

- Insuffisance de couverture des WNs 24%
- Pas d'ILI commun à toutes les unités (unités polylexicales mal reconnues, mauvaises clusterisation, problème de découpage des sens) 30%
- Cliques ambiguës du fait de polysémies parallèles 17%

Ces résultats faibles sont liés à l'incomplétude des ressources, ainsi qu'à la forte polysémie des verbes, à un découpage des sens propre à chaque langue, et aux nombreuses locutions verbales non reconnues.

### 3.4.3 Quelles sont les unités de sens ?

Afin de mieux comprendre les phénomènes mis en jeu par ce concept de clique multilingue, nous avons mené une expérimentation complémentaire en nous appuyant cette fois sur des données dictionnairiques[46]. De la sorte, on évite le bruit lié aux équivalences

46 Il s'agit de travaux non encore publiés.

erronées obtenues lors de l'extraction de correspondances lexicales, et l'on espère par ailleurs obtenir une meilleure couverture des principales acceptions en usage dans la langue générale.

Nous nous sommes basé sur les dictionnaires multilingues en ligne de Larousse (http://www.larousse.fr), avec les langues suivantes : DE, EN, ES, FR, IT. Ces dictionnaires bilingues concernent tous les couples pour ces 5 langues, sauf le couple ES-IT, qui en est absent.

Nous avons conçu un script permettant d'interroger le dictionnaire pour une entrée donnée et une paire de langues donnée, et de nous renvoyer une liste d'équivalents. Par exemple, pour le nom *économie*, on obtient la liste suivante, pour le couple FR-DE :

*FR-N-économie* ↔ *DE-N-Betriebswirtschaft* [47]
*FR-N-économie* ↔ *DE-N-Einsparung*
*FR-N-économie* ↔ *DE-N-Sparsamkeit*
*FR-N-économie* ↔ *DE-N-Volkswirtschaft*
*FR-N-économie* ↔ *DE-N-Wirtschaft*

En nous appuyant sur ces relations, nous avons lancé une extraction d'équivalents en 4 étapes :

1. On recherche tous les équivalents directs d'une entrée de départ (pour tous les couples considérés) en FR. P. ex. : *FR-N-économie*

2. On recherche, réciproquement, tous les équivalents de ces équivalents.

3. Parmi tous les équivalents obtenus en retour pour FR, on suppose qu'un certain nombre sont des pseudo-synonymes (par aller-retour). On réitère en cherchant à nouveau tous les équivalents de ces pseudo-synonymes supposés.

4. Enfin, on recherche à nouveau tous les équivalents de ces équivalents.

De la sorte on obtient un graphe de relations d'équivalence centré sur le mot initial et ses synonymes potentiels. Par ce double aller-retour, on doit *a priori* couvrir tous les équivalents de l'entrée et des mots portant un sens voisin. On peut dès lors extraire toutes les cliques qui contiennent l'entrée initiale, à partir de ce graphe.

47 Pour éviter les ambiguïtés, nous préfixons chaque lemme par le code langue et le code de sa catégorie (ici, N pour Nom).

Nous avons effectué cette extraction pour l'entrée : *FR-N-économie.* Concernant ce mot, le Larousse unilingue en ligne donne les sens suivants[48] :

> - Ensemble des activités d'une collectivité humaine relatives à la production, à la distribution et à la consommation des richesses.
> - Gestion où on réduit ses dépenses, où on évite les dépenses superflues. *Par économie il faisait le trajet à pied.*
> - Ce qu'on épargne, qu'on évite de dépenser.
> - Régulation, organisation visant à une diminution des dépenses, à une adaptation parfaite au but visé : *ce film a été réalisé avec une grande économie de moyens.*
> - Organisation des parties d'un ensemble, d'un système ; structure. *Ce trop long chapitre nuit à l'économie de l'ouvrage.*

Nous avons tenté de classer les cliques obtenues en fonction de ces différents sens. Par commodité nous avons regroupé les sens 2 et 4, qui nous semblent assez proches, et nous avons délaissé le dernier sens, absent des dictionnaires bilingues pour l'entrée *économie*. Nous avons ajouté un sens lié à la discipline académique (*sciences économiques*). Nous avons finalement retenu les sens suivants :

1. Système économique
2. Sciences économiques
3. Épargne
    a) Action d'économiser
    b) Produit de cette action d'économiser

Voici la liste des cliques obtenues classées en fonction de ces différents sens :

Sens 1

- *DE-N-Volkswirtschaft FR-N-économie FR-N-macroéconomie*
- *DE-N-Betriebswirtschaft FR-N-économie FR-N-gestion des entreprises FR-N-micro-économie*

Sens 2

- *EN-N-economics FR-N-aspect économique FR-N-économie FR-N-sciences économiques*
- *EN-N-economics EN-N-economy FR-N-économie IT-N-economia*
- *DE-N-Volkswirtschaft EN-N-economics EN-N-economy ES-N-economía FR-N-économie*

Sens 3.a

- *EN-N-economy FR-N-économie IT-N-economia IT-N-risparmio*
- *DE-N-Sparsamkeit EN-N-economy EN-N-thrift ES-N-economía FR-N-économie*
- *DE-N-Sparsamkeit FR-N-action d'économiser FR-N-économie*

48 cf. http://www.larousse.fr/dictionnaires/francais, consulté en juillet 2014.

- *EN-N-thrift FR-N-économie FR-N-esprit d'économie*

Sens 3.b

- *DE-N-Einsparung EN-N-saving FR-N-économie FR-N-épargne IT-N-risparmio*
- *EN-N-economy EN-N-saving FR-N-économie IT-N-risparmio*
- *DE-N-Einsparung EN-N-saving ES-N-ahorro FR-N-économie FR-N-épargne*
- *EN-N-saving ES-N-ahorro ES-N-economía FR-N-économie*
- *EN-N-saving ES-N-ahorro FR-N-économie FR-N-épargne FR-N-gain*
- *ES-N-ahorro FR-N-économie FR-N-économies FR-N-épargne FR-N-gain*

Cliques ambiguës

- *DE-N-Sparsamkeit DE-N-Wirtschaft EN-N-economy FR-N-économie IT-N-economia*
- *DE-N-Sparsamkeit DE-N-Volkswirtschaft DE-N-Wirtschaft EN-N-economy ES-N-economía FR-N-économie*
- *EN-N-economics EN-N-economy EN-N-saving EN-N-thrift ES-N-economía FR-N-économie*
- *DE-N-Betriebswirtschaft DE-N-Einsparung DE-N-Sparsamkeit DE-N-Volkswirtschaft DE-N-Wirtschaft FR-N-économie*
- *FR-N-action d'économiser FR-N-aspect économique FR-N-café FR-N-économie FR-N-économies FR-N-épargne FR-N-esprit d'économie FR-N-gain FR-N-gestion des entreprises FR-N-macroéconomie FR-N-micro-économie FR-N-restaurant FR-N-sciences économiques*
- *DE-N-Wirtschaft FR-N-café FR-N-économie FR-N-restaurant*

Ces résultats soulèvent un certain nombre de problèmes. Tout d'abord, on constate que les cliques ne concernant que deux langues sont peu fiables, car la présence d'un homonyme peut aboutir à l'union de sens totalement étrangers. C'est le cas pour l'allemand *DE-N-Wirtschaft* qui non seulement est très polysémique dans son sens économique, mais qui possède aussi le sens de /taverne/. On aboutit ainsi aux deux dernières cliques, peu cohérentes :

- *FR-N-action d'économiser FR-N-aspect économique FR-N-café FR-N-économie FR-N-économies FR-N-épargne FR-N-esprit d'économie FR-N-gain FR-N-gestion des entreprises FR-N-macroéconomie FR-N-micro-économie FR-N-restaurant FR-N-sciences économiques*
- *DE-N-Wirtschaft FR-N-café FR-N-économie FR-N-restaurant*

Même si la micro-structure du dictionnaire permet de distinguer entre les homonymes, il n'existe pas d'indice simple pour déterminer lequel correspond bien au sens visé, surtout quand on change de direction, ou de couple de langues…

Pour éviter ces incohérences, on peut ne considérer que les cliques incluant plus de 3 langues. Mais ce faisant, on perd le sens 1 (un des principaux sens), qui ne se retrouve plus que dans les cliques ambiguës. On obtient alors :

Sens 2

- *EN-N-economics EN-N-economy FR-N-économie IT-N-economia*
- *DE-N-Volkswirtschaft EN-N-economics EN-N-economy ES-N-economía FR-N-économie*

Sens 3.a

- *EN-N-economy FR-N-économie IT-N-economia IT-N-risparmio*
- *DE-N-Sparsamkeit EN-N-economy EN-N-thrift ES-N-economía FR-N-économie*

Sens 3.b

- *DE-N-Einsparung EN-N-saving FR-N-économie FR-N-épargne IT-N-risparmio*
- *EN-N-economy EN-N-saving FR-N-économie IT-N-risparmio*
- *DE-N-Einsparung EN-N-saving ES-N-ahorro FR-N-économie FR-N-épargne*
- *EN-N-saving ES-N-ahorro ES-N-economía FR-N-économie*
- *EN-N-saving ES-N-ahorro FR-N-économie FR-N-épargne FR-N-gain*

Cliques ambiguës

- *DE-N-Sparsamkeit DE-N-Wirtschaft EN-N-economy FR-N-économie IT-N-economia*
- *DE-N-Sparsamkeit DE-N-Volkswirtschaft DE-N-Wirtschaft EN-N-economy ES-N-economía FR-N-économie*
- *EN-N-economics EN-N-economy EN-N-saving EN-N-thrift ES-N-economía FR-N-économie*

Il paraît au final vraiment difficile de s'appuyer sur ces cliques pour structurer les différents sens de notre entrée initiale : d'une part, pour un même sens, on peut obtenir un foisonnement de cliques, sans qu'il y ait de critère simple pour les regrouper par clusterisation : c'est le cas pour le sens 3.b. D'autre part, certains sens, parmi les plus fréquents, n'apparaissent qu'au sein de cliques ambiguës, étant portés par des unités très polysémiques : c'est le cas du sens 1, porté par *EN-N-economy, FR-N-économie* ou *IT-N-economia*. Effectuer une clusterisation ne ferait qu'aggraver ces ambiguïtés.

Les résultats ne sont donc pas, loin s'en faut, plus facilement exploitables dans le cas du dictionnaire que dans le cas des cliques obtenues à partir d'un corpus. On obtient les mêmes problèmes de variation de granularité sémantique entre couples de langues. Pour un couple de

langues, dans une certaine direction, on a parfois beaucoup plus d'équivalents, avec un degré de détail plus important, que pour un autre couple de langues. Considérons par exemple les traductions du français et de l'italien vers l'allemand :

*FR-N-économie* ↔ *DE-N-Betriebswirtschaft*
*FR-N-économie* ↔ *DE-N-Einsparung*
*FR-N-économie* ↔ *DE-N-Sparsamkeit*
*FR-N-économie* ↔ *DE-N-Volkswirtschaft*
*FR-N-économie* ↔ *DE-N-Wirtschaft*

*IT-N-economia* ↔ *DE-N-Sparsamkeit*
*IT-N-economia* ↔ *DE-N-Wirtschaft*
*IT-N-economia* ↔ *DE-N-Ökonomie*

Pour obtenir un degré de détail équivalent, il faut en fait considérer les entrées composées du côté italien, car on a aussi :

*IT-N-economia aziendale* ↔ *DE-N-Betriebswirtschaft*
*IT-N-economia nazionale* ↔ *DE-N-Volkswirtschaft*
*IT-N-economia politica* ↔ *DE-N-Volkswirtschaft*

Dans ce cas, la répartition des équivalents est à peu près parallèle en français et en italien, à part pour *DE-N-Ökonomie* et *DE-N-Einsparung*. On a donc deux types de variation qui aboutissent à une fragmentation des cliques : d'une part, des listes équivalents plus ou moins complètes, d'autre part, différentes manières d'organiser la polylexicalité, qui figure soit dans le découpage des entrées, soit simplement dans les exemples donnés pour une entrée simple. Ces variations sont démultipliées par le nombre important de couples de langues à considérer, ce qui explique la dispersion d'un même sens sur plusieurs cliques.

Ces variations sont inhérentes à la fabrication d'un dictionnaire bilingue, et ne peuvent être simplement considérées comme des erreurs ou des incohérences. Il est notoire que les dictionnaires bilingues ne sont pas réversibles, car comme l'explique Corréard (1998 :23), le fait de partir d'une langue implique un point de vue particulier :

> L'effort du traducteur lexicographe porte sur la traduction de chaque mot-vedette en particulier et des problèmes liés à ce mot-vedette. Cette approche donne une saveur particulière aux traductions et rend leur utilisation dans le sens inverse (L2 vers Ll) extrêmement délicate.

Par ailleurs, la persistance de cliques ambiguës s'explique ici par le fait que les ambiguïtés sont parallèles pour toutes les langues (FR, EN, ES, IT) sauf une (DE). Et si les

unités en allemand apparaissent ici comme moins ambiguës, c'est aussi dû au fait qu'en allemand les composés sont soudés, et sont par conséquent moins sujets à variation dans le découpage des entrées.

Dans certain cas, l'accumulation des ambiguïtés et les variations de découpage aboutissent à une situation parfaitement illisible. C'est le cas par exemple des cliques obtenues pour *FR-N-espèce* qui apparaissent comme extrêmement fragmentées :

- *DE-N-Art EN-N-kind EN-N-sort EN-N-species EN-N-type ES-N-especie FR-N-espèce IT-N-specie*
- *DE-N-Art EN-N-sort EN-N-type ES-N-tipo FR-N-espèce FR-N-genre FR-N-nature FR-N-sorte IT-N-tipo*
- *DE-N-Art EN-N-kind EN-N-type ES-N-tipo FR-N-espèce FR-N-genre FR-N-sorte FR-N-type IT-N-tipo*
- *DE-N-Art EN-N-type ES-N-tipo FR-N-caractère FR-N-espèce FR-N-genre FR-N-nature FR-N-sorte FR-N-type IT-N-tipo*
- *DE-N-Art DE-N-Gattung DE-N-Sorte EN-N-kind ES-N-clase ES-N-especie FR-N-espèce FR-N-genre FR-N-sorte IT-N-tipo*
- *DE-N-Art EN-N-kind EN-N-type ES-N-clase ES-N-especie ES-N-género ES-N-tipo FR-N-espèce FR-N-genre IT-N-tipo*
- *DE-N-Art DE-N-Gattung EN-N-kind EN-N-type ES-N-clase ES-N-especie FR-N-espèce FR-N-genre FR-N-sorte IT-N-tipo*
- *DE-N-Art DE-N-Sorte EN-N-kind EN-N-sort ES-N-clase ES-N-especie FR-N-espèce FR-N-genre FR-N-sorte IT-N-tipo*
- *DE-N-Art EN-N-kind EN-N-sort EN-N-type ES-N-clase ES-N-especie FR-N-espèce FR-N-genre IT-N-genere IT-N-specie IT-N-tipo*
- *DE-N-Art EN-N-kind EN-N-sort EN-N-type ES-N-clase ES-N-especie ES-N-tipo FR-N-espèce FR-N-genre FR-N-sorte IT-N-tipo*
- *DE-N-Art DE-N-Gattung EN-N-kind EN-N-type ES-N-clase ES-N-especie ES-N-género FR-N-espèce FR-N-genre IT-N-genere IT-N-specie IT-N-tipo*

Certains sens sont totalement absents, car traités au niveau d'une autre entrée (*payer en espèces*), et de nombreux synonymes (*type, genre, classe, sorte*) existent parallèlement dans les autres langues : d'où cette prolifération de cliques au sens assez vague.

Pour conclure, nous pensons que l'échec de la méthode des cliques est révélatrice de la très grande complexité qui se cache derrière la notion d'équivalence traductionnelle. On sait que pour exprimer une idée, toutes les langues ne disposent pas des mêmes outils, et ne lexicalisent pas de la même manière : ainsi, suivant le couple de langues considéré, l'équivalence ne se situera pas au même niveau de détail. Le dictionnaire ne donne qu'un seul équivalent français à *IT-N-economia*, tandis qu'il en donne 6 en allemand, dont 3 impliquant

des unités polylexicales. Réciproquement, il ne donne que 2 équivalents italiens à *FR-N-économie*, alors qu'il en donne 6 vers l'allemand, dont une qui met en jeu la forme plurielle *FR-N-économies*. Peut-on dire pour autant que la situation est plus simple entre l'italien et le français ? L'équivalence *IT-N-economia* ↔ *FR-N-économie* cache en fait un réseau d'équivalence complexe, qu'elle subsume et dissimule sous son apparente simplicité. Il n'y a qu'à regarder de près quelques expressions en français pour voir que la situation n'est pas plus simple du français vers l'italien que du français vers l'allemand : *l'économie réelle, faire des économies, réaliser une économie, une grande économie de moyens, un professeur d'économie, l'économie des ménages, l'économie industrielle*…

La polysémie et la polylexicalité sont donc tantôt occultées par l'apparente transparence sémantique de certains couples d'équivalents, tantôt révélées par d'autres couples qui fonctionnent différemment.

Plus fondamentalement, ce que révèle ici la dimension multilingue, c'est le caractère illusoire du sens lexical. On croit naïvement que le mot signifie par lui-même, isolément. On lui attache une idée, ou une chose, à la manière des logiciens d'Aristote, à Peirce ou même Lyons, dans les différentes versions de ce que Rastier (1990) nomme triade sémiotique. Mais ce sens lexical n'est que l'effet d'une illusion, peut-être lié à la prégnance psychologique d'une forme de prototype (Rosch, 1975). Dans la réalité des systèmes linguistiques, on ne peut faire abstraction de la manière dont les mots sont utilisés. Les unités de sens, telles que Sinclair (2004) les identifient, ne s'attachent pas aux mots, ni même aux expressions polylexicales possédant un certain degré de figement, mais à la manière dont ces unités fonctionnent dans la phrase et dans le texte : leur régime, leur détermination, leurs modifieurs, la valeur de leurs arguments, mais aussi les fonctions discursives, les routines phraséologiques, les prescriptions génériques. Ce que Sinclair nomme *la* collocation, c'est cette propriété des mots à prendre leur sens en fonction de leurs voisins. Et la difficulté de construire des cliques multilingues cohérentes est selon nous une preuve de cette impossibilité d'isoler les mots autour de leur signification. La signification d'un mot pris isolément est donc en grande partie une illusion, et comme le disait le disait Wittgenstein (1953, 1958 : 20), c'est l'usage – i.e. la façon dont on utilise le mot dans un certain contexte, dans les différents types de « jeux de langage » que la langue permet – qui constitue la signification : « Pour la plupart

des cas de figure où nous employons le mot 'signification' – mais pas pour tous –, on peut le définir ainsi : la signification d'un mot, c'est son usage dans la langue[49]».

Ainsi, dans l'exercice de traduction, le problème de la compositionnalité se pose au préalable pour chaque langue prise séparément. Il résulte de l'équilibre entre les deux principes antagonistes décrits par Sinclair, le principe de l'idiome et le principe de libre choix (1991). Si on veut observer les contrastes et les équivalences entre les langues, on ne peut faire l'économie d'une étude approfondie des unités au sein de leur système linguistique, et de la description de l'idiome en tant que tel. La partie suivante sera consacrée à l'étude des corpus monolingues et comparables, et aux outils que nous avons développés dans cette perspective.

49 « For a large class of cases –though not for all– in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language »

# 4. Des corpus parallèles aux corpus comparables

> Le traducteur ne peut évidemment rien laisser en suspens de qui lui semble obscur. Il doit abattre ses cartes. Il y a, certes, des cas limites, dans lesquels l'original contient quelque chose d'obscur (même pour le premier lecteur). Mais c'est justement dans de tels cas limites d'interprétation qu'apparaît clairement la contrainte qui pèse toujours sur le traducteur. Il lui faut ici prendre son parti et dire clairement comment il comprend. […] Toute traduction qui prend sa tâche au sérieux est plus claire et plus plate que l'original.
>
> H.-G. Gadamer, *Vérité et Méthode*, 1960

La compositionnalité traductionnelle, tout comme la notion d'unité de traduction qui en découle, est, on l'a vu, instable par nature : si parfois les unités de traduction révèlent la non compositionnalité de certaines expressions non traduisibles mot-à-mot, leur périmètre peut se révéler très variable suivant la langue cible et suivant les choix particuliers du traducteur. Nos recherches nous ont donc conduit, progressivement, à raffiner la caractérisation des unités pour des corpus monolingues, et notamment à proposer des outils pour étudier la combinatoire des unités dans une langue donnée.

Tout comme pour les corpus parallèles, nos travaux se sont portés principalement sur le développement d'outils pour l'observation linguistique, et secondairement sur les applications didactiques. Mais avant d'aborder ces travaux, dans les parties 4.2 et 4.3 ci-dessous, il paraît nécessaire d'ouvrir une parenthèse afin d'examiner dans une perspective contrastive les

caractéristiques et les avantages respectifs – ainsi que peut-être les limites – des corpus multilingues comparables et parallèles.

## 4.1. Corpus parallèles *vs* corpus comparables

Comme le note Teubert (1996 : 247), il y aurait quelque chose de suspect et de fondamentalement biaisé dans les corpus parallèles – à tel point qu'ils seraient à bannir de ce qu'on nomme un « corpus de référence », au sens de Sinclair (1996)[50]:

> Il y a une objection essentielle aux corpus parallèles. Les traductions, quelles que soient leur qualité et leur quasi-perfection (ce qui est cependant rare), ne peuvent donner qu'une image déformée de la langue qu'elles représentent. Les linguistes ne devraient jamais se fier à des traductions lorsqu'ils décrivent une langue. C'est pourquoi les traductions n'ont pas leur place dans les corpus de référence. Plus qu'elles ne représentent la langue dans laquelle elles sont écrites, elles donnent une image en miroir de leur langue source.[51]

Une des principales raisons pour s'affranchir des corpus parallèles serait donc leur manque de fiabilité, la traduction produisant des énoncés artificiels portant l'empreinte – plus ou moins visible – des structures de la langue source. L'alternative serait à chercher dans les corpus dits « comparables », c'est-à-dire multilingues mais ne rassemblant que des textes originaux, en s'affranchissant de la relation d'équivalence traductionnelle. En quoi consiste la la « comparabilité » d'un corpus multilingue ?

Certains auteurs, spécialistes du TAL, donnent une définition assez restrictive de ce qu'est un corpus comparable, la comparabilité étant essentiellement une propriété liée à la couverture du vocabulaire. Par exemple, pour Déjean & Gaussier (2002) « deux corpus de deux langues L1 et L2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue L1, respectivement L2, dont la traduction se trouve dans le corpus de langue L2, respectivement L1. » Nous préférons la définition plus générale de Teubert (1996 : 245), pertinente dans le domaine de la linguistique de corpus, où la comparabilité est située au niveau des critères d'échantillonnage du corpus :

50 A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. (Sinclair, 1996 : http://www.ilc.cnr.it/EAGLES96/corpustyp/node18.html, consulté en juin 2014).

51 There is one essential objection to parallel corpora. Translations, however good and near-perfect they may be (but rarely are), cannot but give a distorted picture of the language they represent. Linguists should never rely on translations when they are describing a language. That is why translations have no place in reference corpora. Rather than representing the language they are written in, they give a mirror image of their source language.

> Des « corpus comparables » sont des corpus en deux langues ou plus composés de façon identique ou similaire. Les textes qu'ils contiennent peuvent être classés selon une variété de traits intralinguistiques et extralinguistiques. Le domaine, par exemple, peut être une caractéristique pertinente pour la composition du corpus.[52]

La liste des critères sur lesquels s'appuie la comparabilité selon Teubert reste ouverte et n'est pas exhaustive : aux domaines on peut ajouter les thèmes, les genres textuels, la période, etc. La définition donnée par Sinclair (1996) est très proche, et suggère également une certaine ouverture dans les critères de comparabilité :

> Un corpus comparable est un corpus composé de textes similaires dans une langue ou une variété. Il n'y a pas à ce jour d'accord sur la nature de la similarité, car il y a encore peu d'exemples de corpus comparables.[53]

La comparabilité impose par ailleurs des critères implicites : si on veut par exemple se baser sur des données fréquentielles, il importera que la taille des corpus soit comparable. De même, dans une perspective synchronique, on cherchera plutôt à comparer des textes contemporains entre eux, plutôt que des textes très éloignés dans le temps, afin de réduire les contrastes observés à la dimension interlingue. Un critère selon nous central pour la constitution d'un corpus comparable est celui du *genre* textuel, parce qu'il existe des genres comparables à travers les langues (p.ex. article scientifique, article de presse, roman) et parce que le genre subsume tout un ensemble de prescriptions normative sur le plan de la textualité. Comme le résume Rastier (2006), « instance stratégique de normativité, le genre détermine l'essentiel de la sémiosis textuelle ».

Par rapport aux corpus parallèles, les corpus comparables présentent un double avantage :

– Ils sont beaucoup plus faciles à constituer, car les textes non traduits sont beaucoup plus nombreux. En effet seule une petite fraction des textes accessibles existe en traduction. Dans certains domaines, les traductions sont quasi inexistantes – c'est par exemple le cas pour les articles scientifiques, qui sont de plus en plus souvent rédigés

---

52 'Comparable corpora' are corpora in two or more languages with the same or similar composition. All corpora have an explicit or implicit composition. The texts they contain can be classified according to a variety of intralinguistic or extralinguistic features. Domains, for instance, can be a feature relevant to the composition of a corpus.

53 A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora.

directement en anglais dans de très nombreuses disciplines. Ils sont en outre plus faciles à traiter, car ils ne nécessitent pas de mettre en œuvre une étape d'alignement.

- Ils sont a priori plus fiables, parce que dénués de biais traductionnels.

Ainsi, tandis qu'en TAL on se rabat généralement sur les corpus comparables du fait de la difficulté à obtenir des corpus parallèles de grande dimension, pour des linguistes comme Sinclair ou Teubert, on constitue des corpus comparables à dessein pour éviter l'écueil des corpus parallèles, susceptibles de refléter une image fausse de la langue : « Un corpus comparable doit permettre de comparer différentes langues ou variétés dans des circonstances de communication similaire, mais en évitant les inévitables distorsions dues à la traduction dans les corpus parallèles » (Sinclair, 1996)[54]

De quelle sorte de distorsions s'agit-il ? Teubert donne deux exemples pour étayer cette position, un concernant le lexique et l'autre concernant la syntaxe. D'abord, il affirme qu'un lexème typiquement allemand, telle que *Schadenfreude,* qui n'a pas d'équivalent lexical en anglais, apparaîtra vraisemblablement rarement dans une traduction allemande d'un texte anglais. Ensuite, sur le plan de la syntaxe, il suppose qu'une construction propre à l'allemand, telle que le passif impersonnel (« Es wurde viel getrunken », littéralement en anglais « It was drunk a lot ») sera en général absente dans une traduction allemande de l'anglais. L'argument de Teubert repose donc sur l'idée de l'absence supposée de certaines constructions ou expressions typiques de l'idiome d'arrivée, étant donné que ces constructions ou expressions sont à priori absentes de la langue source. Une traduction est donc supposée plus pauvre, car ne réalisant pas toutes les potentialités de la langue cible. Nous parlerons désormais de l'hypothèse d'appauvrissement.

### 4.1.1 Hypothèse d'appauvrissement

Nous avons voulu vérifier cette hypothèse en utilisant les corpus parallèles à notre disposition, issus du projet Emolex (cf. chapitre 4.3.2 ci-dessous).

Nous avons ainsi interrogé un corpus parallèle français allemand de 18 298 453 occurrences (dans les deux langues), constitué essentiellement de textes littéraires

[54] The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus. URL: http://www.ilc.cnr.it/EAGLES96/corpustyp/node21.html, consulté en juin 2014.

contemportains (avec quelques textes du XIX[è]). On y trouve 28 occurrences de *Schadenfreude*, ainsi réparties dans le corpus en fonction de la langue source :

| *Langue source* | *Occurrences de Schadenfreude* | *Taille du corpus parallèle* |
|---|---|---|
| *de* | 1 | 628 029 |
| *en* | 5 | 87 596 |
| *fr* | 21 | 17 079 606 |
| *sw* | 1 | 444 076 |

*Tableau 4.1 : Répartition des occurrences de* Schadenfreude *en fonction de la langue source*

Si on suppose que le nombre d'occurrences de *Schadenfreude* dans le texte allemand est indépendant de la langue source, on doit s'attendre à trouver une valeur proche de 28 * 628 029 / 18 298 453 = 0,96 occurrences en allemand : l'hypothèse d'indépendance est donc bien vérifiée pour l'allemand. On constate en revanche une surreprésentation pour les textes en langue source anglaise : il s'agit en fait de romans de Tom Clancy, tous traduits par la même équipe de traducteurs. Ainsi, les aléas de la distribution d'un tel lexème semblent liés à des facteurs idiosyncrasiques tels que l'identité de l'auteur et/ou du traducteur, plutôt qu'à la direction de traduction. L'hypothèse d'appauvrissement, pour ce lexème-ci du moins, est donc contredite par les observations du corpus : le caractère idiomatique de *Schadenfreude* en allemand n'aboutit pas à un affaiblissement de sa fréquence au sein des textes traduits.

Teubert donne un autre exemple de nom typiquement allemand, et réputé sans équivalent stable : *Missgunst*, dont les équivalents proches en français seraient *jalousie* ou *envie*. On en trouve deux occurrences : la première dans une traduction de *Madame de Pompadour*, des frères Goncourt, dans un titre de chapitre ajouté par le traducteur ; la seconde dans la traduction d'un roman de Jean Echenoz :

> *Aus ihren Blicken sprachen nichts als Eifersucht und Missgunst.*
> *les regards qu'ils échangeaient ne dénotaient qu'envie et jalousie.*
> Jean Echenoz (1999) *Je m'en vais*

Même si ces fréquences sont trop faibles pour pouvoir en tirer une conclusion générale (le corpus parallèle contenant essentiellement des textes français en langue source, il est assez prévisible d'obtenir plus d'occurrences dans cette partie du corpus), on peut en conclure néanmoins que les deux exemples cités *Schadenfreude* et *Missgunst* sont bien représentés

dans les textes traduits. On ne peut certes en tirer de conclusion définitive sur l'hypothèse d'appauvrissement sur le plan lexical : ces deux exemples étaient peut-être simplement mal choisis. Nous avons effectué un test sur un autre lexème typiquement germanique, *Gemütlichkeit*, qui désigne une situation de confort, de tranquillité, de bien-être domestique. Cette fois, sur 22 occurrences, nous en avons trouvé 12 dans les textes originaux allemands, soit 54,5% des occurrences dans une partie du corpus qui en représente 3,43 %. Dans ce cas, il y a donc bien une sous-représentation du lexème dans les textes allemands traduits. Ce que montrent ces exemples, c'est que l'hypothèse d'appauvrissement lexical doit être nuancée : il existe sans doute un biais fréquentiel pour certaines unités assez rares, et dont il faudrait étudier le périmètre plus précisément – mais il semble qu'aucune forme de l'idiome d'arrivée ne puisse être a priori exclue d'une traduction. Sur le plan lexical tout au moins, rien ne permet d'affirmer qu'un traducteur n'aura pas recours à toute l'étendue du matériau linguistique dont il dispose.

Pour vérifier cette hypothèse d'un point de vue plus général, nous pouvons comparer l'accroissement du vocabulaire pour le texte original et sa traduction, de la même manière que Fleury (2009) pour un corpus anglais-français ou Miao et Salem (2009) pour un corpus français-chinois. Dans ces travaux on constate que la courbe de l'accroissement du vocabulaire du texte traduit se situe au-dessus de celle de l'original : mais on ne peut en tirer de conclusion, dans la mesure où le comptage du vocabulaire est étroitement lié aux opérations de segmentation et de lemmatisation des unités lexicales. Les différences observées peuvent être imputées à des propriétés linguistiques (variations morphosyntaxiques des unités) et à des artefacts liés à la segmentation et à la normalisation des unités, autant qu'à des distorsions traductionnelles. Pour vraiment identifier un effet traductionnel, il faut pouvoir considérer deux bi-textes comparables représentant les deux directions de traduction : on peut alors comparer les courbes d'accroissement pour une même langue, suivant qu'il s'agit d'un original ou d'une traduction.

Dans ce but, nous avons constitué deux corpus parallèles comparables, le premier constitué de 14 œuvres littéraires allemandes (nous noterons DE-Source) et de leurs traductions en français (nous noterons FR-cible), et le second de 14 œuvres en français (FR-Source) avec leurs traductions en allemand (DE-cible)[55]. Pour que la période temporelle soit la

[55] Sans présumer du sous-genre ni de la qualité littéraire : il s'agit d'œuvres ayant rencontré un certain succès public, ce qui les rend plus aisément disponibles en traduction et en version numérique. La liste de ces

même entre les corpus sources et les corpus traduits, nous n'avons sélectionné que des œuvres contemporaines récentes (dans l'intervalle 1977-2000 pour le corpus DE-Source, et 1986-2006 pour FR-Source). Nous avons tenté d'équilibrer au mieux le corpus au niveau des auteurs (13 auteurs différents pour chaque corpus), des tailles et des genres, et avons obtenu la composition suivante :

| Corpus DE-FR<br>Romans, récits en allemand | DE-Source | FR-Cible |
|---|---|---|
| Nombre d'occurrences | 756 969 | 909 124 |
| Nombre de caractères | 3 394 817 | 3 508 746 |

| Corpus FR-DE<br>Romans, récits en français | DE-Cible | FR-Source |
|---|---|---|
| Nombre d'occurrences | 866 189 | 905 596 |
| Nombre de caractères | 3 862 320 | 3 426 452 |

*Tableau 4.2 : Composition des corpus parallèles comparables DE-FR et FR-DE*

D'un point de vue général, on observe pour tous les couples de textes que l'allemand est plus économe en occurrences, avec un accroissement du vocabulaire beaucoup plus rapide. Ceci peut s'expliquer par le phénomène de soudure graphique des noms composés en allemand, qui aboutit à des mots plus longs représentant moins d'occurrences et un vocabulaire plus variés (les mots composés français, sans soudure graphique, n'étant pas comptés dans le vocabulaire). De fait, le plus grand nombre d'occurrences en français n'implique pas nécessairement des textes plus longs en caractères. Si on examine le nombre de caractères, on constate qu'il est dans tous les cas supérieurs dans les textes traduits, par rapport aux originaux, qu'il s'agisse du français ou de l'allemand. Quelle que soit la direction, la traduction aboutit donc, de façon générale, à une augmentation de la taille du texte traduit, en nombre de caractères : c'est de qu'on observe pour 24 textes sur 28. Cela peut s'expliquer, parfois, par la nécessité pour le traducteur d'expliciter certaines informations dans le contexte et la culture d'arrivée[56]. Par ailleurs, assez fréquemment, la transposition d'une construction

œuvres est donnée en Annexe - 2.p. 172

56 Et parfois d'ajouter des explications en note, même si cela reste marginal : dans tout le corpus FR-cible, on ne compte que 15 occurrences de la forme *N.D.T.* Notons que nous avons conservé les quelques notes de bas de page, peu nombreuses, mais que nous avons supprimé manuellement tous les éléments péritextuels plus volumineux susceptibles de brouiller le parallélisme : table des matières, préface, postface, etc.

grammaticale en une construction équivalente implique ce que Vinay & Darbelnet (1958) appellent l'« étoffement », c'est-à-dire l'ajout d'un syntagme servant de support à un pronom, un adverbe ou une préposition (p.ex. *the charge against him* → l'accusation *portée* contre lui).

Cette légère inflation textuelle n'est pas en elle-même porteuse de biais traductionnel, mais si on compare l'accroissement du vocabulaire pour les textes sources et cibles pour une même langue, on doit bien admettre que la situation est assez contrastée :

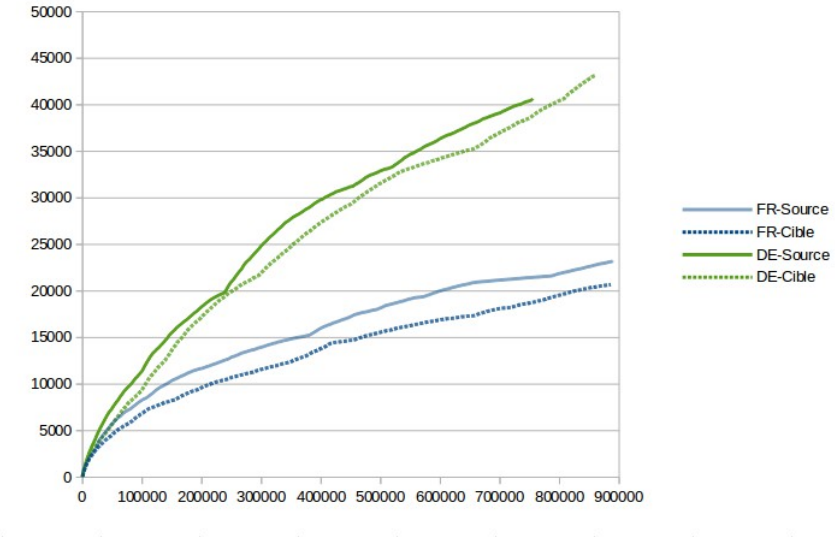


*Figure 4.1 : Accroissement du vocabulaire (lemmes) comparé entre textes orignaux et traductions*

La figure ci-dessus montre l'évolution du vocabulaire au niveau des lemmes[57]. Les irrégularités de ces courbes sont liées aux différences de richesse lexicale dans les différents textes (voire dans les différents passages du corpus). Pour rendre ces courbes plus facilement comparables nous avons procédé à un mélange aléatoire des occurrences. Toutes les œuvres étant ainsi mélangées, le profil de la courbe est alors plus régulier :

[57] La segmentation et la lemmatisation ont été produites par le logiciel Connexor (Tapanainen & Järvinen, 1997).

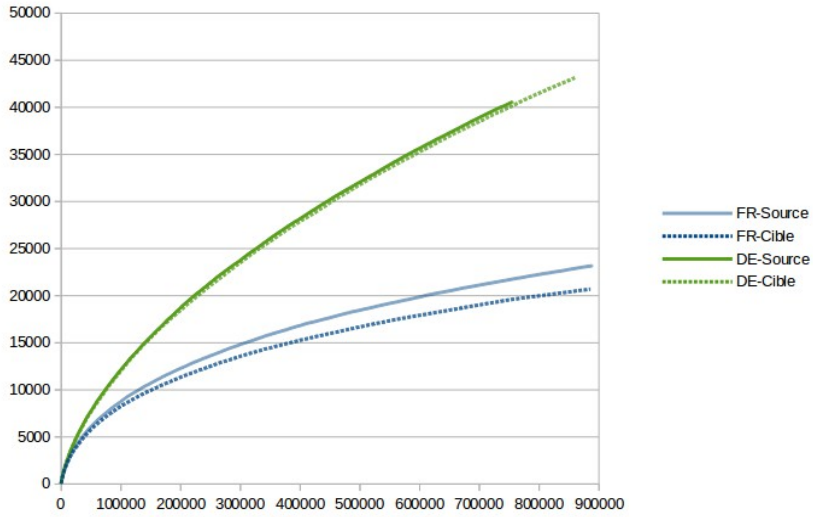


*Figure 4.2 : Accroissement du vocabulaire (lemmes) comparé entre textes orignaux et traductions (lissé par mélange aléatoire)*

Il est frappant de constater que les courbes pour l'allemand sont rigoureusement superposées, alors que pour le français on constate que l'accroissement du vocabulaire est inférieur en traduction. Pour vérifier qu'il ne s'agissait pas là d'un biais lié aux opérations de segmentation et de lemmatisation[58], nous avons effectué la même extraction pour les formes, en segmentant de façon « brutale » au niveau des espaces, des tirets, des apostrophes et de tout signe de ponctuation. On obtient alors les courbes ci-dessous :

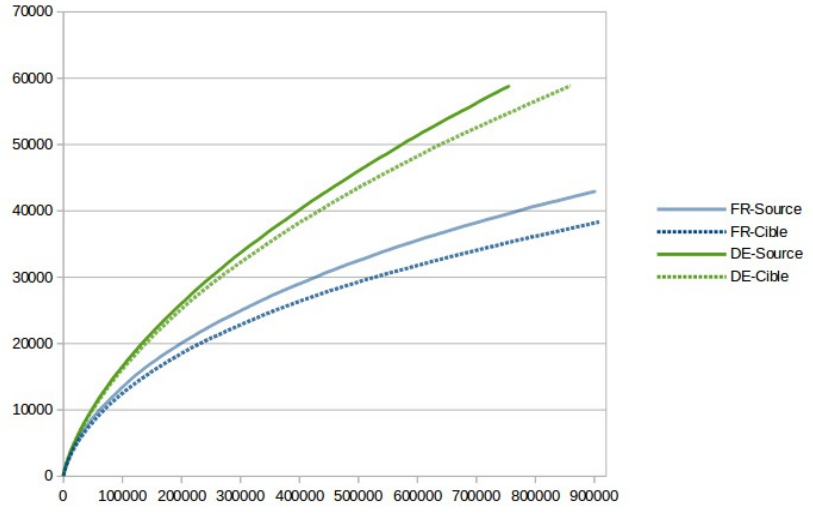


*Figure 4.3 : Accroissement du vocabulaire (formes) comparé entre textes orignaux et traductions (lissé par mélange aléatoire)*

58 p.ex. la locution « de temps en temps » est identifiée par Connexor comme une seule unité.

Ces courbes ont de quoi laisser perplexe[59] : cette fois l'écart entre traduction et original s'observe pour les deux langues, bien que de façon moins marquée pour les corpus allemands. Pour ceux-ci, la raison est purement morphosyntaxique : pour le même ensemble de lemmes, on observe en moyenne une plus grande variété de formes dans le corpus original. À titre d'illustration, voici quelques exemples pris au hasard :

| | DE-Source | DE-Cible |
|---|---|---|
| *losgehen* | *losgegangen, losgehen, losging, losgingen, loszugehen* | *losgegangen, losgehen, losgeht, losging* |
| *dämlich* | *dämlich, dämliche, dämlichen, dämliches, dämlichste* | *dämlich, dämliche, dämlichen* |
| *fortsetzen* | *fortgesetzt, fortsetze, fortsetzen, fortsetzt, fortzusetzen* | *fortgesetzt, fortsetzen, fortsetzten, fortzusetzen* |
| *trinken* | *Trank, Trink, Trinken, Trinkst, getrunken, trank, tranken, trink, trinke, trinken, trinkst, trinkt* | *Trinken, getrunken, trank, tranken, trink, trinke, trinken, trinkt* |

*Tableau 4.3 : Quelques exemples de variations morphologiques*

Bien entendu, il arrive aussi qu'un lemme du corpus DE-Cible possède plus de variantes morphologiques que dans le corpus DE-Source : n'oublions pas que ces deux corpus rassemblent des textes différents. Mais en moyenne, il y en a plus dans le corpus source : pour 40 589 lemmes on a 60 927 formes, soit environ 1,50 formes par lemme, tandis que dans DE-cible on trouve pour 44 694 lemmes 64 468 formes différentes, soit environ 1,44 formes par lemme. Pour l'allemand on peut donc faire ce double constat : le corpus d'originaux et le corpus de traductions présentent la même richesse lexicale (en terme d'accroissement du vocabulaire de lemmes), mais le corpus de textes originaux est légèrement plus varié sur le plan de la morphologie flexionnelle. Pour expliquer ce phénomène, il faudrait un examen plus poussé sur le plan traductologique, que nous ne sommes pas en mesure d'effectuer ici. Quant au corpus français, on constate un léger appauvrissement du lexique pour le corpus de textes traduits. S'agit-il d'un épiphénomène lié aux particularités de notre corpus ? Est-ce dû au fait que pour ces 14 textes, nous n'avons que 12 traducteurs différents ? La traduction en français,

[59] Nous avons vérifié en détail le parallélisme des corpus, et relancé plusieurs fois nos calculs, afin de vérifier l'absence de biais expérimental.

pour ce type de textes littéraires, opère-t-elle une forme de normalisation lexicale, aussi légère soit-elle ? Les traditions de traductions sont-elles différentes Outre-Rhin ? Les questions soulevées sont nombreuses, et appellent toutes de plus amples études, à la fois traductologiques et textométriques, sur de plus grandes quantités de données. Peut-être doit-on partager le constat de Gadamer (1960) : « Toute traduction qui prend sa tâche au sérieux est plus claire et plus plate que l'original. » ?

Mais quand bien même on généraliserait le constat d'un léger appauvrissement de la variété lexicale dans les corpus de traduction, cela suffirait-il à les disqualifier, au point de les bannir à priori d'un corpus de référence ? Si l'on compare l'accroissement du vocabulaire pour 10 des textes originaux qui compose notre corpus FR-Source, voici ce qu'on obtient :

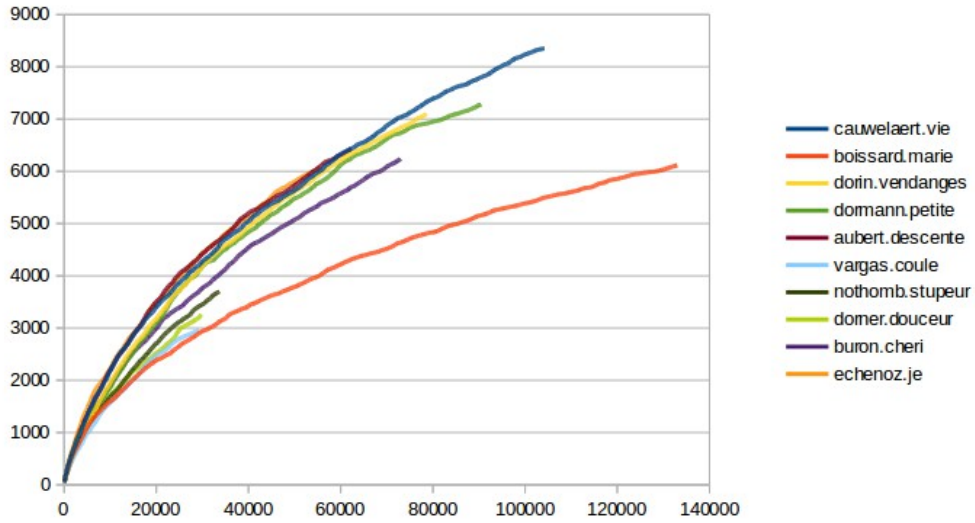


*Figure 4.4 : Accroissement comparé du vocabulaire (lemmes) pour des textes de FR-Source*

Ce n'est pas une grande découverte : on constate que l'accroissement du vocabulaire, et donc la richesse lexicale, est très variable d'un texte à l'autre, avec des écarts bien plus importants, en proportion, avec ceux observés jusqu'ici. Cette variation est intrinsèque à la constitution de tout corpus composé de textes différents, et ne constitue aucunement un critère pour exclure certains textes *a priori* : nous pensons qu'il en va de même, tout au moins sur le plan de la richesse lexicale, pour les corpus de traduction.

Qu'en est-il sur le plan syntaxique ? Nous avons également cherché, dans le même corpus parallèle, des occurrences du passif impersonnel mentionné par Teubert. Nous avons cherché les deux expressions suivantes :

*es* + *wurde* + *ADV* + *ParticipePassé* et *es* + *wurde* + *ParticipePassé*

Nous n'avons retenu que les occurrences de cette expression dont le sujet *es* n'avait pas d'antécédent anaphorique (p.ex. *Er war so geschickt, dass man kaum mitbekam, was er tat , aber es wurde gesehen*[60]) – afin de n'avoir que des tournures impersonnelles –, et dont le verbe n'avait ni complément d'objet direct ni subordonné complétive (p.ex. *es wurde angenommen daß…*[61]) ou infinitive (*es wurde beschlossen, abzuwarten und bei Tageslicht weiterzusuchen*[62]).

Au final, nous n'avons trouvé que deux occurrences correspondant à la structure recherchée, et toutes deux sont des traductions :

> *Vorbeigehenden Besatzungsmitgliedern entging das nicht, und es wurde bereits getuschelt.*
> *Des hommes d'équipage s'en rendirent compte au passage, et les murmures commencèrent.*
> Tom Clancy (1984) *The Hunt for Red October*

> *Wir stritten uns über den jeweiligen Kurs, es wurde gelacht und gesungen.*
> *On disputa des courses, il y eut des rires, des chansons.*
> Simone de Beauvoir (1958) *Mémoires d'une jeune fille rangée*

Là encore les fréquences sont trop faibles pour en tirer une conclusion générale. Notons toutefois qu'aucun de ces deux exemples en français n'utilise la tournure avec le pronom indéfini *on*, qui semble être la plus proche, syntaxiquement, du passif impersonnel allemand. La tournure impersonnelle avec *il y eut* impose une nominalisation du procès. La construction allemande apparaît donc bien « spontanément » dans la traduction, et l'hypothèse d'appauvrissement s'en trouve affaiblie aussi sur le plan syntaxique, bien qu'on ne puisse exclure que les traductions aboutissent à des biais sur le plan fréquentiel pour certaines constructions précises.

---

60 Il était si adroit qu'on avait du mal à deviner ce qu'il faisait, mais il le fit, et il a été vu.

61 … on supposait que…

62 … il a été décidé d'attendre le jour pour faire le gros du boulot.

On pourra nous reprocher le fait que les exemples ci-dessus sont tirés de traductions littéraires. Or la traduction littéraire implique nécessairement un travail de recréation, qui impose au traducteur d'être en quelque sorte lui-même écrivain. Comme le note Meschonnic (1999 : 85) « Qu'on puisse parler du Poe de Baudelaire et de celui de Mallarmé montre que la traduction réussie est une écriture… » (Meschonnic, 1999 : 85). Berman (1988 : 24) remarque d'ailleurs que dans la tradition occidentale, l'acte de traduire et celui d'écrire sont inextricablement liés : « Origine de l'écriture, la traduction est aussi son horizon. Pour un homme du XVI$^{e}$ siècle, écrire n'est jamais bien loin de traduire. Non seulement l'écriture vient de la traduction, mais elle ne cesse d'y *retourner*. » La traduction littéraire implique donc à la fois une grande maîtrise de langue cible et des qualités de créativité littéraire, comme le signale Motoc (2002) : « Il y a autant de maîtrise, sinon plus, dans l'acte de traduire : ce travail de l'écriture auquel les écrivains se soumettent déjà à la force, ne fait que « se reconstruire » dans la traduction. ».

Pour d'autres domaines de la traduction, comme la traduction technique ou la traduction juridique, le traducteur, sans cesser d'être interprète, n'a pas les marges que confère la licence poétique. Les contraintes de productivité et le recours à des outils de traduction assistée par ordinateur (TAO), par ailleurs, donnent à la pratique de la traduction spécialisée un tour plus mécanique, du moins en partie. On peut dès lors craindre que des distorsions apparaissent, non plus sous une forme négative (par l'absence supposée d'unités lexicales ou de constructions de la langue cible), mais sous une forme positive, par la projection sur la langue cible des structures de la langue source. Voilà sans doute une sérieuse raison de suspecter les textes traduits de constituer un « miroir de la langue source » (Teubert, Ibid.).

### 4.1.2 Présence de calques et d'emprunts

Selon la typologie de Vinay et Darbelnet (1958 : 47-52), désormais devenue classique, le « calque » désigne le fait d'« emprunte[r] à la langue étrangère le syntagme, mais [de] tradui[re] littéralement les éléments qui le composent ». Par extension, chez de nombreux auteurs, le calque désigne le transfert d'un procédé de construction, qui peut se situer à différents niveaux : on parle de calque syntaxique, sémantique (Chuquet et Paillard, 2004 : 223-224), morphologique (Di Spralo *et al.*, 2010), morphosyntaxique, morpho-sémantique, etc.

À partir d'un vaste corpus de textes parallèles de l'Union européenne, Manuel Torrellas Castillo (2009) a consacré sa thèse à une analyse minutieuse des interférences linguistiques dans les textes espagnols de l'UE. Le recours à des corpus massifs (le *JRC-Acquis*) et l'utilisation d'outils de traitement de corpus parallèles tels qu'Alinéa lui ont permis d'identifier des phénomènes d'interférence assez ténus, difficilement repérables par un dépouillement manuel.

Dans une publication commune (Duchet *et al.*, 2008), il mentionne de nombreux types d'interférences :

- « emprunt lexical » (ibid. : 138) : p.ex. *délocalisation* = * *deslocalización* (avec glissement sémantique de *localizar*)
- « emprunt de collocations » (ibid. : 139-140) : p. ex. *prestataires de service* = * *prestatario de servicios* (avec glissement sémantique de *prestatario*), *indemnité journalière* = ?*indemnización diaria* (au lieu de *dieta*, l'équivalent le plus conforme)
- « calques constructionnels » (ibid. : 141) : p. ex. *susceptible de* + *V*, comme dans *susceptible de provoquer* → *susceptibles de provocar*, alors qu'en espagnol *susceptible de* doit en principe être suivi par un verbe de sens passif ou une nominalisation avec déterminant zéro (*susceptible de recurso*).
- « calques syntaxiques » (ibid. : 142) : syntagme prépositionnel vs gérondif en apposition : *sur la base de* = *en base a*, au lieu du gérondif en apposition *basándose en* plus idiomatique.

Dans le même article, Jean-Louis Duchet (ibid. : 144) signale un autre type de calque : *permit* + *prédicat nominalisé* en anglais (p. ex. *permit the gradual implementation (...)* ↔ *permettre la mise en œuvre progressive (...)*). Il note que cette construction est peu naturelle, quoique permise dans un registre étroitement spécialisé. Il y a bien calque, car on constate une différence de « degré de banalisation » entre les constructions anglaises et françaises, différence qui s'apparente à une dérive sémantique.

Outre ces interférences, ces corpus sont par ailleurs marqués par une certaine forme d'appauvrissement, lié au caractère répétitif des traductions :

> L'état de langue que nous avons observé, très fortement marqué par l'activité des traducteurs, manifeste aussi une restriction des choix lexicaux (..). Cette tendance est confortée par l'effet des mémoires de traduction (…) qui peuvent imposer pendant une longue période une traduction exclusive aux dépens de traductions équipossibles, pouvant aller jusqu'à la fossilisation d'erreurs reprises par tous les utilisateurs d'une même mémoire.
>
> Torrellas Castillo (2009 : 302)

Il faut toutefois relativiser ces constats : les interférences identifiées manifestent des phénomènes ténus, « à bas bruit » dirait-on en médecine, inhérents au contact de langue dans un cadre professionnel spécialisé. Dans l'exercice de leur profession, l'exigence de qualité oblige les traducteurs de ce type de texte à remettre en question leurs choix et à s'assurer de l'idiomaticité du texte produit. Mais cette recherche d'idiomaticité est contrecarrée par la spécialisation du discours : or, les discours professionnels n'ont pas vocation à rechercher cette forme de « banalité » ou de « généralité » inhérente à l'idiome : bien au contraire, ils forgent des usages qui tendent à se démarquer – qu'il s'agisse d'affirmer une identité socio-professionnelle ou de se forger des termes clairement identifiés. Quand le contact de langue est inhérent à la profession, comme c'est le cas dans les institutions internationales comme l'UE, ou encore dans le monde des affaires, alors la convergence linguistique que l'on observe parfois (souvent très fortement marquée par l'anglais, quoique la situation soit plus nuancée en ce qui concerne l'UE, vu la forte influence du français) fait partie intégrante de ce qu'on peut définir comme un *technolecte*. Les usages spécialisés dans le monde des affaires et du commerce constituent une bonne illustration de ce type de convergence : la forte empreinte de l'anglais dans les échanges internationaux se manifeste par de nombreux calques et emprunts, par exemple dans la composition nominale, avec des termes tels que : *communication produit*, *responsable produit*, *responsable marketing*, *service communication*, *business modèle*, etc.

Comme le note Goffin (1994 : 642) parlant des écrits communautaires : « Par sa nature, ses origines, ses modes de formation et son fonctionnement, ce langage – auquel on peut conférer la dignité *d'eurolecte* – ne se démarque aucunement des règles qui gouvernent toute langue de spécialité. » L'exigence de convergence économique et politique, et la recherche de transparence des écrits officiels, explicitement formulée au sommet d'Edimbourg de 1992, font qu'il est parfois difficile de distinguer entre interférence traductionnelle et spécialisation du discours – comme dans les exemples précédents donnés par Torrellas Castillo (2009).

Au terme de cette longue discussion, nous réaffirmons que l'appréhension assez générale des linguistes vis-à-vis des textes traduits est injustifiée – voire irrationnelle. Rien ne permet d'écarter *a priori* un texte d'un corpus au prétexte qu'il s'agit d'une traduction, et qu'en tant que tel, on ne peut lui attribuer le caractère d'authenticité que tout autre texte, quelle qu'en soit la qualité, se voit attribuer spontanément. La traduction est une activité de communication comme toutes les autres activités langagières, et mérite en tant que telle d'être réintégrée dans le champ de la linguistique. Il existe de mauvaises traductions, tout comme il existe des textes mal rédigés, au regard des normes de l'écrit. Il existe des traductions émaillées de calques et d'interférence tout comme il existe des textes rédigés par des locuteurs non-natifs, ou dans des situations de contact linguistique telles qu'ils produiront spontanément quantité de calques ou d'interférences. Bref, un texte traduit n'est pas plus suspect que n'importe quel texte écrit de première main, surtout s'il est le fait d'un traducteur professionnel que l'on peut considérer à juste titre comme un expert de la langue cible (en principe sa langue maternelle).

On voit parfois la traduction, à tort, comme une opération de transcodage, visant à établir des équivalences entre des unités sources et cible : il s'agit peut-être là d'une réminiscence de lycéens – nous avons tous pratiqué la traduction naïvement, et en tant qu'apprenants. Dans cette perspective du transcodage, on ne peut nier, en effet le caractère artificiel du résultat. Mais comme l'écrit Rastier (2006) « la question de la traduction spécifie une question générale qui concerne non les rapports de langue à langue, mais les rapports de texte à texte, puisque tout texte en transforme d'autres : quels sont les rapports sémiotiques entre deux textes qui dérivent l'un de l'autre, qu'il s'agisse de réécriture créatrice, de commentaire ou de traduction ? » Traduire, c'est donc écrire un texte qui s'inscrit dans un corpus intertextuel, et qui participe au devenir de la langue : « En outre, en élargissant le corpus, [la traduction] fait évoluer la langue : le corpus des textes traduits s'intègre au corpus de la langue. » (Rastier 2006)

### 4.1.3 Complémentarité

La principale raison, selon nous, de recourir à des corpus comparables, est leur disponibilité : il est bien plus facile de regrouper de vastes corpus multilingues lorsqu'on s'affranchit de la contrainte de la traduction.

Cela explique sans doute, outre les réserves méthodologiques précédemment discutées, que la plupart des travaux en lexicographie bilingue s'appuient aujourd'hui sur des corpus comparables. Notons que depuis les travaux de pionnier de John Sinclair (1991) et la sortie du *Collins COBUILD English Language Dictionary* (1987), basé sur le corpus COBUILD de 7 millions de mots, le recours à de grandes bases textuelles dans le processus de la rédaction de dictionnaires est devenu incontournable. Comme le note Sinclair, l'introspection seule ne peut être une source fiable pour déterminer ce qu'est l'usage réel de la langue :

> (…) l'écart entre le sentiment linguistique des locuteurs, concernant les détails de la langue, et les faits récoltés objectivement à partir des textes est énorme et systématique. Il nous conduit à émettre l'hypothèse que l'intuition humaine à propos de la langue est spécifique aux individus, et qu'elle ne peut pas du tout constituer un bon guide pour décrire ce qui se passe réellement lorsque ces mêmes individus font usage de la langue (Sinclair 1991 : 4)[63]

Dans le cadre de la lexicographie bilingue, le dépouillement de corpus en langue source et cible est également très profitable. Pour l'*Oxford-Hachette French Dictionary*, deux corpus ont été utilisés : l'*Oxford Pilot Corpus* pour l'anglais (60 M de mots) et un corpus de français moderne de 10 M de mot réunis pour le projet par Oxford University Press. Comme le note Grundy (1998), le corpus permet de donner des réponses à de nombreuses questions précises concernant les usages, notamment sur le plan de leurs fréquences :

> (…) aucune équipe de lexicographes ne peut espérer mener à bien ce travail herculéen de documentation et d'analyse sans avoir accès à des textes. Quelle est la fréquence de telle ou telle unité lexicale et quelle est son importance pour la communication ? Quelle place doit-on lui accorder dans un dictionnaire ? Quels sont les modèles syntaxiques de base qu'elle exploite ? Quelle acception est la plus fréquente ? Quels sont les exemples les plus typiques de son utilisation dans chacune de ses acceptions ? Quelles acceptions sont devenues vieillies ou obsolètes ? Qu'est-ce qui constitue un véritable changement de sens et qu'est-ce qui relève simplement d'une préférence contextuelle ? (Grundy, 1996 : 131)

63 "(…) the contrast exposed between the impressions on language detail noted by people, and the evidence compiled objectively from texts is huge and systematic. It leads one to suppose that human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language "

Ainsi, lors des étapes d'*analyse* (analyse des mots-vedettes en langue source) et de *transfert* (recherche des équivalents et traduction des exemples en langue cible)[64] ce sont des corpus monolingues qui sont utilisés – et non des corpus multilingues parallèles.

La terminologie est également un domaine où l'on fait abondamment usage de corpus comparables. Des recherches en TAL ont montré, depuis un certains temps déjà, comment extraire des lexiques bilingues à partir de corpus comparables (Daille *et al.* 1994, Rapp, 1999, Déjan & Gaussier 2002). Ces recherches s'appuient généralement sur une identification préalable des termes candidats (simples ou complexes) dans chaque langue séparément, puis sur l'appariement des termes en se basant sur une comparaison de leurs contextes (ou des contextes de leurs voisins distributionnels), les vecteurs contextuels étant traduits avec des ressources dictionnairiques bilingues (Morin *et al.*, 2004). Pour les termes complexes, on peut combiner une approche compositionnelle, consistant à traduire séparément, au moyen d'un dictionnaire, chaque composant du terme complexe, et l'approche contextuelle, consistant à traduire les vecteurs de contextes des composants du terme complexe, lorsqu'ils n'ont pas d'équivalents directs dans le dictionnaire (Morin & Daille, 2011, 2012).

Pourtant les corpus parallèles peuvent aussi apporter des informations utiles, notamment pour guider la phase de transfert, en indiquant les traductions les plus communes pour un mot-vedette, une expression ou un terme. C'est ce que nous avons montré en collaborant avec Anaïch Le Serrec (Le Serrec *et al.*, 2010), qui a travaillé sur un corpus parallèle issus d'une organisation internationale, le GIEC. Le couplage d'un outil de détection des termes comme TermoStat (Drouin & Doll, 2008) et d'un outil d'alignement bilingue des termes simples comme Alinéa, s'est révélé pertinent pour servir d'appui à l'élaboration d'une ressource terminologique. Le Serrec *et al.* (2010) suggèrent d'ailleurs que des fonctionnalités d'identification de terme et d'alignement bilingue puisse être intégrées dans un même outil, ce qui n'a encore pas été fait à notre connaissance.

D'un point de vue général, Teubert (1996 : 248) note avec justesse que dans la mesure où il n'existe pas d'équivalence simple entre les codes, tant sur le plan du lexique que de la

64 Grundy écrit plus loin (1996:134) : « La méthodologie mise en œuvre pour la création d'un dictionnaire varie considérablement en fonction de l'ampleur du projet, de la nature des données linguistiques disponibles, et des ressources, essentiellement financières et matérielles, qui ont été allouées. Néanmoins, quelle que soit l'ampleur ou la complexité de l'entreprise, trois processus distincts entrent en jeu (…) Ces trois étapes sont les processus que Atkins désigne respectivement par *analyse, transfert et synthèse (analysis, transfer and synthesis)*, Atkins, 1990). »

syntaxe, la recherche d'équivalent de traduction implique des configurations complexes, qui mettent souvent en jeu des expressions polylexicales et des constructions étendues. Ces configurations font partie du savoir implicite des traducteurs, et on ne les trouve pas ailleurs que dans des corpus parallèles :

> « Les traductions sont rendues possibles par le fait que les traducteurs en savent habituellement plus que ce qu'ils trouvent dans les grammaires, ou dans les dictionnaires monololingues ou bilingues (…). Si l'on veut capter la connaissance implicite que les traducteurs ont des équivalents traductionnels, il n'y a pas d'autre choix que d'analyser les traductions. Elles constituent l'archivage des appariements qui ont été proposés, testés et établis au fil du temps. Les corpus parallèles sont une des sources les plus précieuses pour la recherche des équivalents traductionnels. »
>
> Teubert (1996 : 248-249)

Il en découle que des corpus parallèles suffisamment grands permettent également d'établir inductivement, par la variété des exemples qu'ils donnent, des généralisations concernant « les conditions sémantiques qui doivent être remplies (…) dans le contexte pour que le mot *a* de la langue *A* soit traduit par le mot *b* de la langue *B*. » (Ibid.) Cette approche inductive qui cherche à identifier, par la récurrence des exemples, l'ensemble des caractéristiques contextuelles qui déterminent tel ou tel phénomène n'est ni plus ni moins que l'approche *corpus driven* appliquée en lexicographie par Hanks ou Sinclair, mais étendue aux corpus de traduction. Cela rappelle également l'extraction de correspondances lexicales telle que nous l'avons mise en œuvre (cf. partie 3 de cette synthèse), mais sous une forme moins formalisée et plus complexe, car intégrant tous les traits synctatico-sémantiques du contexte. D'ailleurs, notons qu'aucune des 25 thèses de Teubert (2005) sur la linguistique de corpus n'exclut les corpus de traduction.

Dans une recherche exploratoire, Bertels & Verlinde (2011) montrent comment les corpus comparables et les corpus parallèles fournissent des approches méthodologiques complémentaires et convergentes, qui peuvent être utiles pour la lexicographie bilingue, notamment dans une perspective didactique. Ils identifient deux sortes de « profils » : les corpus comparables permettent de caractériser les « profils combinatoires » des unités et d'analyser les collocatifs pertinents, alors que les corpus parallèles permettent d'extraire ce

qu'ils appellent le « profil de traduction » d'une unité – les deux étant, naturellement, intimement liés.

Nous avons nous même tenté de montrer, dans une collaboration avec Elena Melnikova et Iva Novakova, comment articuler la complémentarité entre les deux types de corpus (Melnikova *et al.*, 2009) : tandis que les corpus parallèles permettent d'identifier, de manière directe des équivalents fonctionnels, entre unités et constructions, les hypothèses émises doivent ensuite être vérifiées sur des corpus comparables de grande taille, mieux à même de fournir des données fiables sur le plan fréquentiel.

Les deux dernières parties de cette synthèse seront consacrées à des développements centrés sur l'observation de phénomènes internes aux langues, concernant les corpus monolingues et comparables.

## 4.2. Des corpus aux applications didactiques

À la suite de mes travaux sur les applications didactiques des corpus bilingues (cf. partie 3.3, p. 78), j'ai travaillé, parallèlement au développement des méthodes d'interrogation et de concordance, dans deux directions :

- d'une part, l'aide à la sélection des textes, en vue de la constitution de ressources didactisées (textes complets ou concordances) ;
- d'autre part, la correction et le diagnostic d'erreur automatisé pour des activités impliquant des réponses ouvertes courtes (ROC), par exemple dans le cadre de questions de compréhension ou d'exercices lacunaires (tels que ceux présentés en Annexe - 1. p. 166).

Concernant l'aide à la sélection de textes pour les enseignants, j'ai travaillé avec des étudiants sur le développement d'un projet en 2007 dans le cadre du Master 2 Industries de la langue. Ce projet a ensuite été finalisé lors du stage d'un étudiant nommé Ralf Baumbach, qui a poursuivi les développements du site. L'idée consistait à télécharger périodiquement des textes sur le Web, à partir de différents types de sources (littérature, blog, chansons, articles de presses), et à les analyser afin de fournir une indexation de ces ressources pertinentes pour le choix d'un texte. Divers critères ont été mis en œuvre, comme le montre la figure 4.5 :

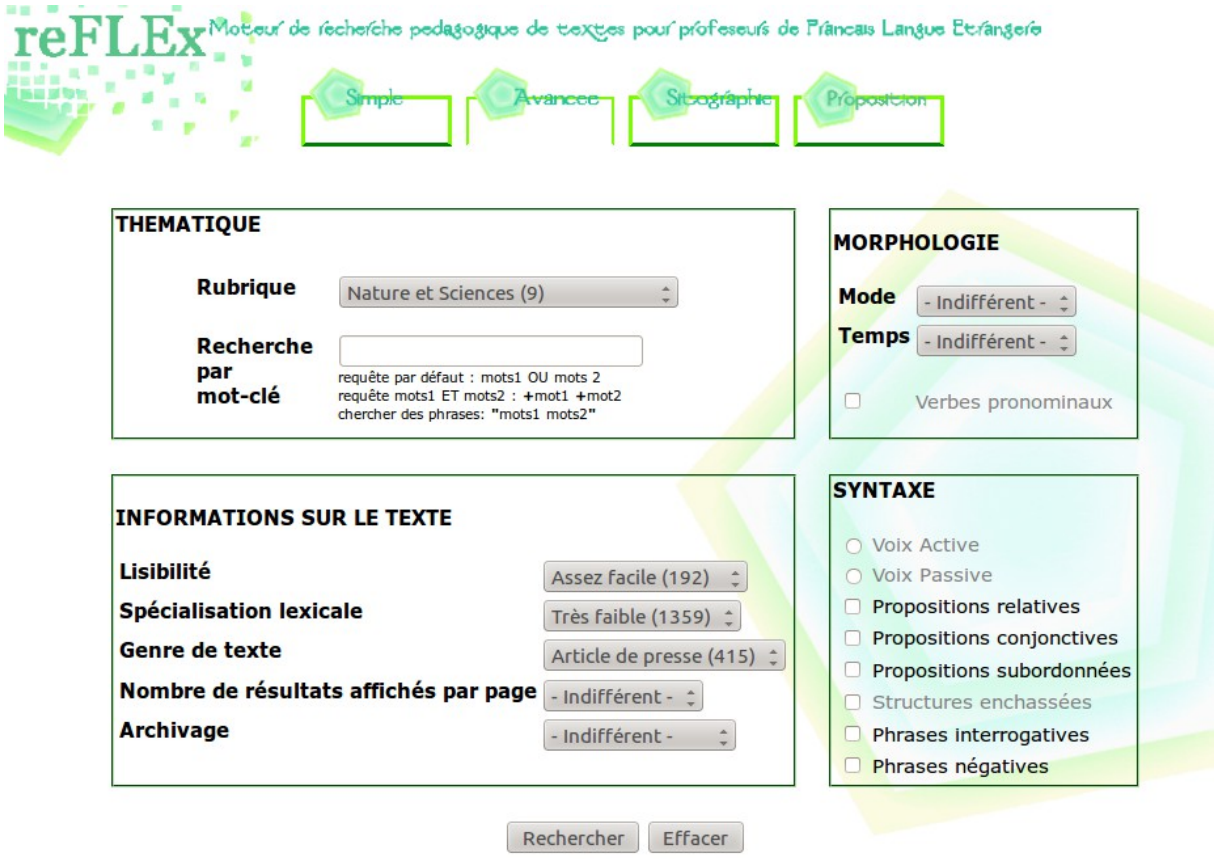


*Figure 4.5 : critères de sélection de texte dans l'interface de reFLEx*

Aux critères classiques liés au genre textuel et à la thématique (en s'appuyant sur des mots-clés ou les rubriques des articles), s'ajoutent des critères purement textométriques tel que l'indice de lisibilité de Kandel & Moles (1958) et un coefficient de spécialisation du lexique (basé sur une comparaison avec les fréquences trouvées dans Frantext). Enfin, et c'est là l'originalité d'un tel outil, on y intègre également des critères liés à la morphosyntaxe, en s'appuyant sur les sorties de *Treetagger*.

Une fois les textes sélectionnés à partir de ces critères, on donne la possibilité de visualiser différentes « facettes » (pour reprendre le terme de Loiseau *et al*., 2010) pour guider le choix définitif, comme le montre la figure ci-dessous.

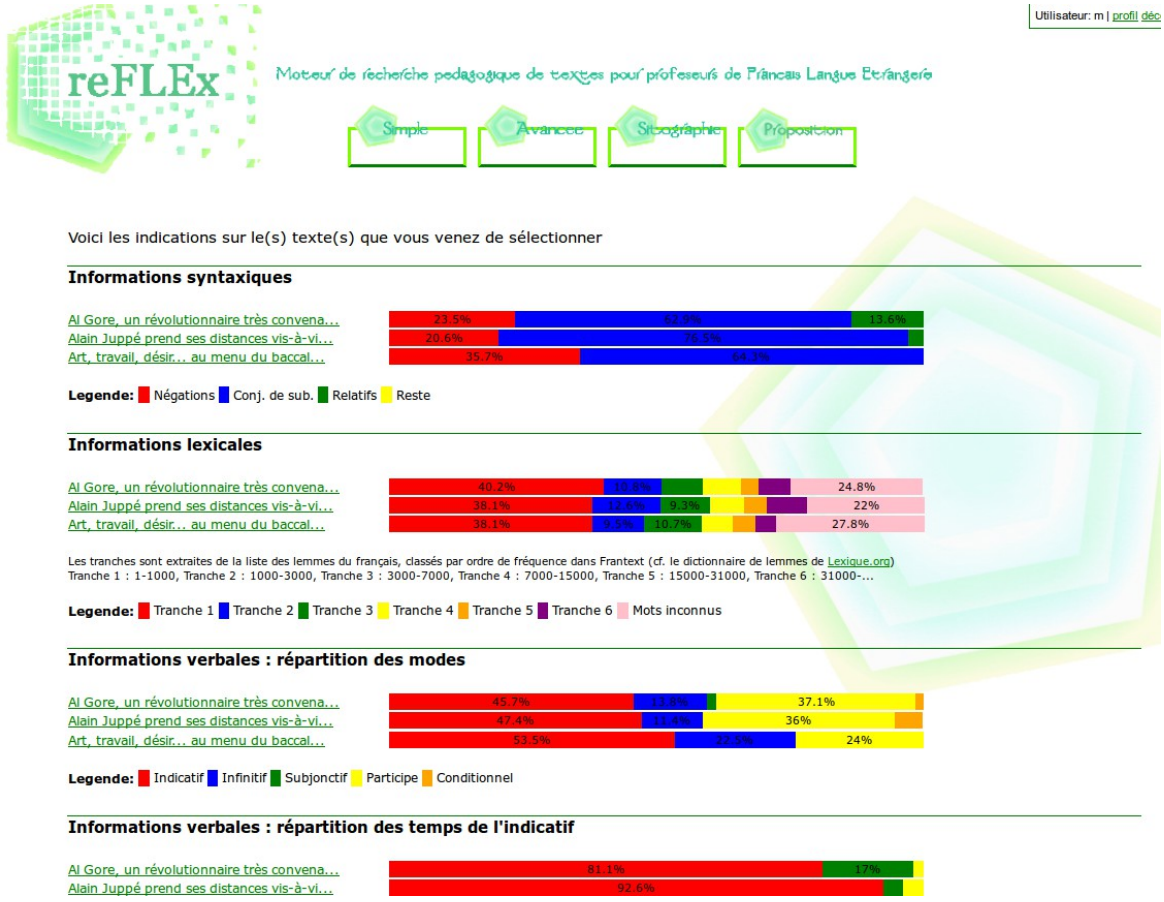


*Figure 4.6 : visualisation comparatives de différentes « facettes » des textes choisis*

Ces propositions semblent cohérentes avec les pratiques pédagogiques : en effet, en s'appuyant sur les résultats d'une enquête passée auprès de 130 enseignants, Loiseau *et al.* (2010), citent les 4 exemples suivant de critères usuels parmi ceux-ci :

- le choix d'un auteur et d'un type de texte en fonction des structures que l'enseignant attribue à son « style » d'écriture ;
- le choix d'un journal en fonction de la lisibilité présumée de ses articles (…) ;
- le choix d'un numéro de périodique en fonction d'un type de texte attendu ;
- le choix d'un type de texte (…) par rapport à un type d'activité (exercices lacunaires).

Cet outil original est malheureusement resté à l'état de prototype, et n'a pu être testé auprès du public visé : comme il impliquait un archivage des contenus (notamment des articles de presse), nous nous sommes heurtés à des questions de propriété intellectuelle. Le projet est resté en suspens, mais nous aimerions y retravailler dans les années qui viennent, notamment pour caractériser un peu plus finement l'étude du lexique – en nous appuyant sur les listes de vocabulaire fondamental (Gougenheim *et al.*, 1964), et des indices plus fins pour

mesurer la diversité et la densité lexicale (Read, 2000). Par ailleurs, j'ai commencé à collaborer avec des collègues de la société Pearson (Londres), pour mesurer automatiquement la richesse phraséologique de productions d'apprenants (Benigno *et al.*, 2014). Pour ce faire, nous avons projeté des ressources phraséologiques (une liste de collocations dites « académiques », et une liste collocations de langue générale tirés de deux dictionnaires généralistes[65]). Cette première étude a montré une corrélation assez claire entre l'utilisation des collocations académiques et le niveau des apprenants. La même méthode pourrait être employée à rebours, pour sélectionner des textes avec un niveau de difficulté adapté en termes de richesse et de spécialisation de la phraséologie.

À travers nos recherches sur les applications didactiques du TAL, nous avons montré comment bâtir des ressources pour la génération d'activités auto-correctives à partir de corpus de texte authentiques : nous ne développerons pas ici ces aspects, qui sont détaillés dans Kraif & Ponton (2007) et Blanchard *et al.* (2009), afin de nous concentrer sur la piste de recherche que nous avons principalement approfondie par la suite : la recherche d'expressions complexes et l'étude des profils combinatoires de ces expressions.

## 4.3. Développement d'outils pour la recherche d'expressions

Certaines des idées décrites dans Kraif & Tutin (2006) dans une perspective d'aide à la rédaction en langue seconde, ont pu être concrétisées dans un cadre monolingue, avec le développement du projet Scientext (Falaise *et al*., 2011) autour de l'écrit scientifique. L'interface développée pour ce projet permet d'interroger un corpus d'articles, de thèses et de mémoires, structuré par discipline et par genre. Elle permet notamment de rechercher des collocations dans le corpus par une entrée sémantique, avec 5 grandes classes liées aux notions de lexique et de phraséologie transdiciplinaire (Tutin, 2007) : */Dénomination/Autour des hypothèses/Evaluation et opinion/Auteurs cités/Propositions propres de l'auteur/.* Chaque classe se subdivise ensuite en sous-catégories lexicales, liées à la fois à des critères fonctionnels et syntaxiques. On a par exemple, pour l'expression du positionnement de l'auteur, les sous-catégories suivantes (Tutin, 2010) :

*/Evaluation et opinion/ → Verbes d'opinion, Verbes modaux d'opinion, Adjectifs d'évaluation, Adjectifs d'opinion, Noms d'opinion, Adverbiaux d'opinion.*

[65] 110 000 collocations issues du *Longman Dictionary of Contemporary English* (LDOCE) et du *Longman Collocations Dictionary and Thesaurus* (LCDAT)

Ces classes lexico-sémantiques correspondent à des expressions de recherche complexes traduites dans le langage d'interrogation de ConcQuest. Une méta-grammaire, spécialement élaborée pour Scientext, permet de définir des variables contenant des listes de critères, de définir des relations syntaxiques par composition, et d'assembler simplement plusieurs expressions de ConcQuest. Par exemple, pour rechercher les adjectifs d'évaluation qui portent sur les noms scientifiques, les contributeurs du projet Scientext ont codé les expressions suivantes :

```
(ATTRIB,#2,#1) = (SUJ,#3,#1) (ATTS,#3,#2) ;
$eval=acceptable,adéquat,aisé,ambitieux,approximatif,bon,central,clair,classique,cohérent,complet,complexe,concis,confus,convaincant,correct,crucial,déterminant,difficile,discutable,effectif,efficace,encourageant,épineux,erroné,essentiel,excellent,facile,faible,fin,flou,fondamental,important,innovant,insuffisant,intéressant,invalide,irréprochable,judicieux,majeur,mauvais,meilleur,important,nouveau,original,passable,passionnant,performant,pertinent,principal,prometteur,riche,rigoureux,satisfaisant,séduisant,sérieux,significatif,solide,souhaitable,stimulant,suffisant,vague,valable,valide,véritable,vrai
$theo=analyse,approche,article,caractéristique,cas,choix,communication,concept,conception,contribution,critère,définition,description,donnée,élément,étude,exemple,facteur,fonction,idée,méthode,modèle,notion,objectif,outil,paramètre,phénomène,principe,problème,projet,proposition,qualité,question,réflexion,résultat,rôle,solution,structure,système,terminologie,test,théorie,traitement,travail
Main = <lemma=$eval,#1> && <lemma=$theo,#2> :: (ATTRIB,#1,#2) OR (ADJ,#1,#2);
```

La première ligne permet de définir la relation d'attribution entre le sujet de la copule comme la composition de deux relations de surface, entre le sujet et la copule, et entre la copule et l'adjectif. Les variables $*eval* et $*theo* permettent d'exprimer les listes des adjectifs évaluatifs et les noms transdisciplinaires auxquels ils s'appliquent. La règle *Main* permet ensuite d'exprimer les nœuds correspondants (#1, et #2) et les arcs (ATTRIB ou ADJ) dans l'arbre de dépendance.

### 4.3.1 Interface de requête

Une difficulté inhérente à ce type d'outil est de trouver le bon équilibre entre la puissance expressive du formalisme et la simplicité d'utilisation : pour rencontrer ses utilisateurs, un outil ne doit pas nécessiter une formation trop spécialisée ni des compétences avancées en informatique. Dans Kraif (2008b), nous avons effectué quelques propositions dans ce sens, en proposant un langage d'interrogation similaire à CQP, intégrant en outre un système de contraintes syntaxiques à l'instar de *TigerSearch* (König & Lezius, 2003) – mais

de façon très simplifiée. Le formalisme que nous proposons permet une certaine progressivité dans le raffinement des requêtes : une requête peut être formulée comme une simple concaténation de formes, mélanger des formes et des lemmes, intégrer des insertions facultatives, permettre des choix alternatifs (disjonctions), spécifier des traits morphosyntaxiques, et enfin définir des relations syntaxiques complexes. Voici quelques exemples de requêtes de complexité croissante, avec des occurrences issues du corpus Emolex :

- concaténation *: il est tard → « Lorsqu'il se réveille,* ***il est tard****. »*
- lemmatisation *: il %être tard → « en rouspétant parce qu'****il était tard*** *et qu'il avait envie de se coucher. »*
- insertion facultative *: il %être <>? tard → « Lorsqu'il reprend le contrôle de sa trajectoire,* ***il est trop tard****. »*
- formes alternatives *: (il|%ce) %être <>? <>? tard → « Le championnat,* ***ce sera pour plus tard****. »*
- traits morphosyntaxiques *: il <l=être,f=.*impf.*> <c=ADV>? <c=ADJ>? tard → « Elle était exténuée et* ***il était déjà tard****. »*
- relations syntaxiques : *il <l=être,f=.*impf.*,#1> <c=ADV>? <c=ADJ>? tard && <#2,c=V>:: (comp,1,2*) *→ « Elle avait essayé de lui parler mais i****l était trop tard*** *pour* ***entrer*** *dans le vif du sujet. »*

Ce formalisme intègre en outre, de façon facultative, le langage des expressions régulières, tant au niveau des formes que des valeurs de traits. Nous pensons ainsi qu'il est possible, pour un utilisateur, de se former progressivement au langage, à mesure que ses besoins se précisent.

Pour l'aider dans cette démarche, nous avons développé en 2007 un prototype d'assistant graphique, permettant d'ajouter des tokens le long de l'axe syntagmatique, ainsi que sur l'axe paradigmatique pour les formes alternatives. Une illustration de cet assistant est fournie dans la figure 4.7 ci-dessous.

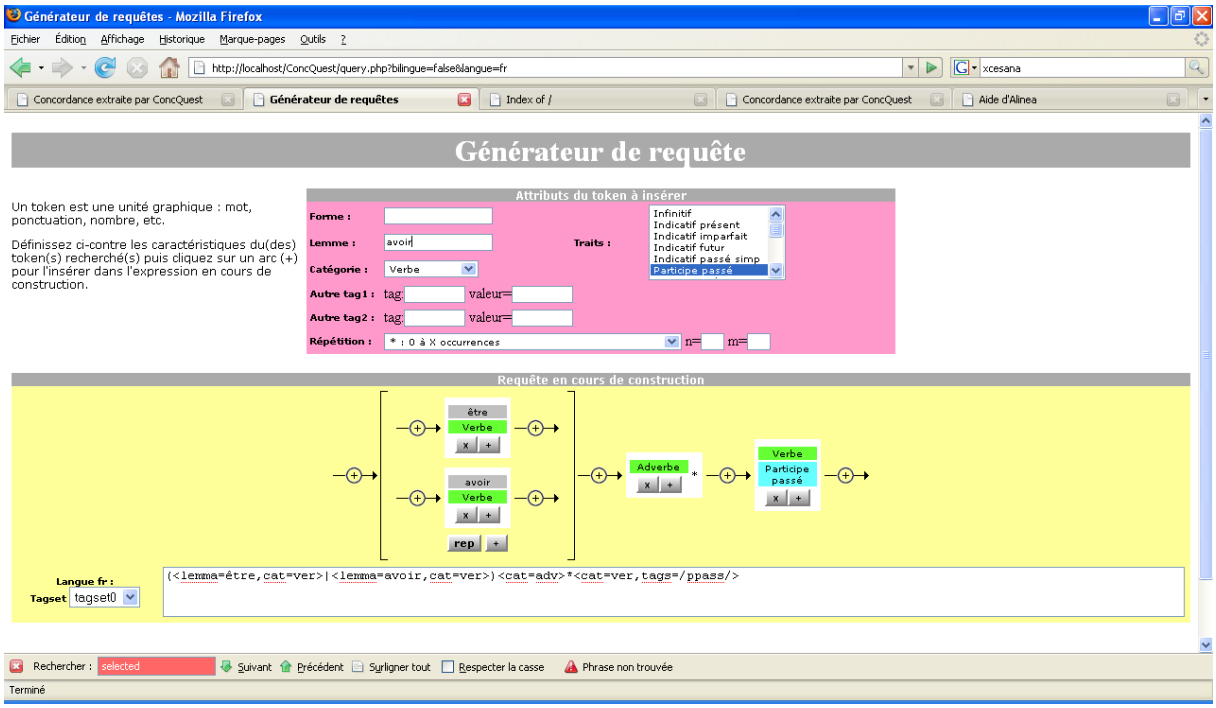


*Figure 4.7 : Assistant graphique pour la construction des requêtes dans ConcQuest*

Ces propositions ont été affinées dans l'élaboration de l'interface du corpus Scientext, conçue et développée par Achille Falaise (Falaise *et al.*, 2012) :


*Figure 4.8 : Interface de recherche simple pour le corpus Scientext*

### 4.3.2 Étude des profils combinatoires : le projet Emolex

Le formalisme de requête élaboré pour ConcQuest a également été réinvesti dans les outils de concordance développés pour le projet Emolex, dirigé par Peter Blumenthal et Iva Novakova, et centré sur l'étude de la combinatoire du lexique des émotions. L'objectif de ce projet était d'analyser d'un point de vue contrastif, dans une perspective formulée par Sinclair (2004) ou encore Hoey (2005), les valeurs sémantiques et les rôles discursifs à partir de la combinatoire du lexique des émotions, afin d'élaborer une cartographie permettant de mieux structurer ce champ lexical, avec des applications en lexicographie mais aussi en didactique des langues et traductologie.

Dans le cadre du projet, nous avons réuni un corpus comparable de grande dimension, intégrant également un sous-corpus parallèle. Ces corpus sont interrogeables en ligne par le biais d'une plate-forme nommée *EmoBase*[66], qui donne accès à trois outils d'interrogation : *EmoProf*, dédié aux applications didactiques du projet ; *EmoLing*, une base de données enregistrant la modélisation de la combinatoire (champ sémantiques, actants, etc.) du lexique des émotions, tel qu'elle a été codée par les linguistes du projet ; et enfin, *EmoConc*, l'outil d'interrogation des corpus comparables et parallèles, sur lequel nous avons concentré nos efforts. Les corpus d'EmoBase ont été rassemblés dans cinq langues : le français, l'allemand, l'anglais, l'espagnol et le russe. Les corpus comparables comprennent environ 140 millions de mots par langue : des textes journalistiques pour un total d'environ 120 millions de mots, et des textes littéraires représentant environ 20 millions de mots (pour l'essentiel des romans des années 1950 à nos jours). Le corpus parallèle a une taille d'environ 78 millions de mots au total et comprend uniquement des textes littéraires (des romans du XIX$^{è}$ et du XX$^{è}$ siècle, la plus grande part étant constituée de romans contemporains) qui ont été alignés avec leur traduction respective à l'aide du programme Alinéa.

Dans EmoConc, pour caractériser le profil combinatoire d'une entrée, nous avons repris le concept de lexicogramme, introduit par Maurice Tournier et repris dans le logiciel WebLex (Heiden, Tournier 1998) : il s'agit d'établir, pour un pivot donné, la liste de ses cooccurrents les plus fréquents, à gauche et à droite, en faisant l'extraction des fréquences de cooccurrence et en calculant des mesures d'association statistiques (telles que rapport de vraisemblance ou t-score). Pour construire ces lexicogrammes, nous avons proposé un modèle de cooccurrence

66 URL : http://emobase.u-grenoble3.fr, consulté en juin 2014.

flexible permettant à l'utilisateur de définir lui-même les unités de cooccurrences : formes, lemmes, catégories morphosyntaxiques, traits additionnels (p.ex. sémantiques), relations syntaxiques (dans le cas des colligations) ou des combinaisons de ces informations. La possibilité de faire intervenir des combinaisons de ces traits nous semble importante pour permettre à l'utilisateur d'ajuster la focale de ses observations en allant du général au particulier (ou vice-versa), de préciser des contraintes pour désambiguïser certains contextes, et de combiner les aspects lexicaux et syntaxiques dans ses observations. Par ailleurs nous préconisons une caractérisation flexible de l'espace de cooccurrence, qui conditionne les points de rencontre entre pivot et collocatifs, ainsi que la manière de les dénombrer. On peut par exemple définir la cooccurrence à l'intérieur d'un empan de largeur fixe, éventuellement différente à droite et à gauche du pivot. Mais on peut aussi rechercher la cooccurrence syntaxique, à l'instar de Kilgariff et Tugwell (2001) ou Charest *et al.* (2010), mise en jeu lorsqu'une relation fonctionnelle (du type sujet, complément d'objet, modifieur, etc.) a été identifiée entre deux unités. Evert (2007) signale l'intérêt de ce type de cooccurrence en termes de bruit et de silence : « à la différence des cooccurrences de surface, [la cooccurrence syntaxique] ne fixe pas une limite de distance arbitraire entre deux cooccurrents, tout en introduisant moins de " bruit " que dans la cooccurrence textuelle »[67]. Pour la cooccurrence syntaxique, nous exploitons les relations de dépendance obtenues grâce à différents analyseurs : XIP pour l'anglais (Aït-Mokhtar *et al.*, 2001), Connexor pour l'allemand, le français et l'espagnol (Tapanainen & Järvinen 1997), DeSR pour le russe (Attardi *et al.*, 2007), basé sur un modèle stochastique créé à partir du corpus arboré SyntagRus (Nivre *et al.*, 2008). Nous avons par la suite complété ces relations pour obtenir des dépendances plus pertinentes sur le plan sémantique (p. ex. sujet profond dans les constructions passives, etc.).

Avec le modèle de cooccurrence ainsi défini, on peut viser des aspects très génériques de la combinatoire (par exemple : quels sont les principaux collocatifs de la forme *surprise* toutes relations confondues) ou beaucoup plus spécifiques et circonscrits (par exemple : quels sont les principaux collocatifs verbaux à l'imparfait du nom lemmatisé *surprise* pris en tant qu'objet direct). Le tableau 4.4, repris de Kraif & Diwersy (2012), montre un tel lexicogramme :

[67] « (…) unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less "noise" than textual cooccurrence »

| l1 | l2 | f | f1 | f2 | loglike |
|---|---|---|---|---|---|
| surprise_N | créer_V | 614 | 2098 | 21658 | 4548,4333 |
| surprise_N | réserver_V | 230 | 2098 | 2869 | 2143,50164 |
| surprise_N | avoir_V | 484 | 2098 | 423602 | 627,503103 |
| surprise_N | constituer_V | 94 | 2098 | 13778 | 406,792757 |
| surprise_N | éviter_V | 43 | 2098 | 16296 | 109,29478 |
| surprise_N | manifester_V | 22 | 2098 | 2424 | 106,621896 |
| surprise_N | causer_V | 19 | 2098 | 2210 | 90,0605475 |
| surprise_N | ménager_V | 15 | 2098 | 1495 | 75,5763954 |
| surprise_N | exprimer_V | 23 | 2098 | 6186 | 72,5375788 |
| surprise_N | provoquer_V | 23 | 2098 | 10551 | 50,6130103 |
| surprise_N | feindre_V | 9 | 2098 | 676 | 50,3068057 |

*Tableau 4.4 : extrait du lexicogramme pour le nom lemmatisé surprise pris en tant qu'objet direct (f=fréquence de cooccurrence, f1=fréquence de l1, f2=fréquence de l2*

Le *loglike* mentionné dans ce tableau, ou rapport de vraisemblance, est une mesure d'association classique qui exprime l'invraisemblance d'obtenir un certaine niveau de cooccurrence par le simple jeu du hasard. En effet, plus l'association entre deux unités est forte, plus cette mesure d'invraisemblance est élevée.

#### 4.3.2.1. Visualisation des profils

À partir de ces lexicogrammes, nous offrons différentes modalités d'exploration :

– Pour l'analyse linguistique, le « retour au texte » est indispensable : un simple clic sur un collocatif permet de retrouver, sous forme de concordance, tous les contextes de cooccurrence avec le pivot.

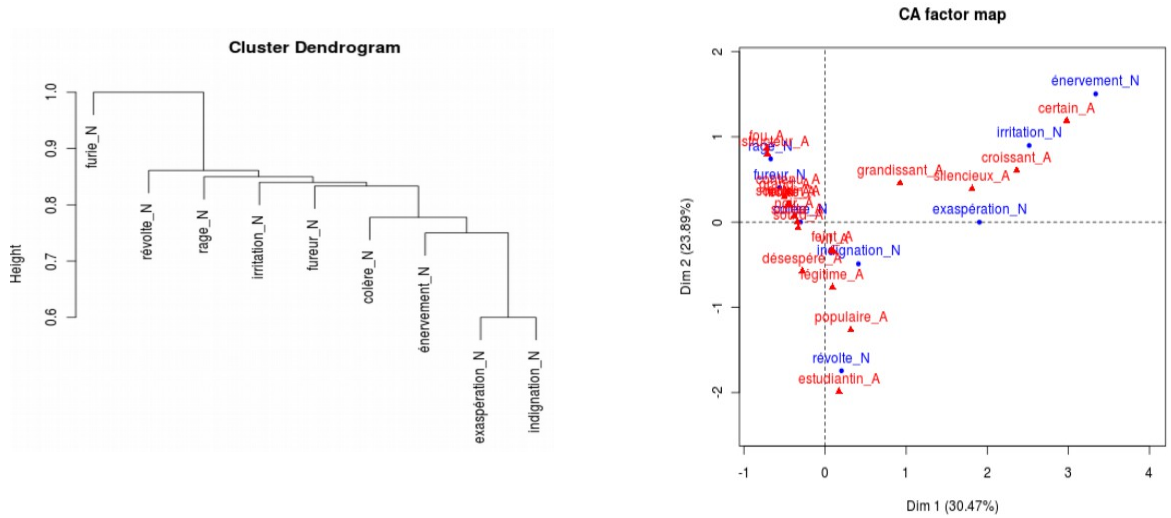


*Figure 4.9 : Classification hiérarchique et AFC (domaine sémantique de la 'colère')*

#### 4.3.2.2. Prise en compte des *pivots complexes*

– Pour comparer de manière synthétique divers profils combinatoires, nous proposons d'identifier les lexicogrammes à des points dans un espace vectoriel, en ne retenant que la mesure jugée la plus pertinente (fréquence, loglike, t-score, etc.). Il est dès lors possible d'utiliser des méthodes d'analyse de données pour visualiser les similarités entre pivots : analyse factorielle des correspondances (AFC), échelonnement multidimensionnel (MDS) ou classification hiérarchique ascendante (hClust). La figure 4.9, reprise de Kraif et Diwersy (2012), montre ces sorties pour des unités du domaine sémantique de la 'colère' (obtenues grâce aux modules du projet 'GNU R'). La classification, réalisée pour la relation "objet", indique une hiérarchisation assez bien corrélée à l'intensité du sentiment. Quant à la *factor map*, réalisée pour des relations quelconques concernant des collocatifs adjectivaux, elle permet de distinguer trois groupes : *révolte*, *indignation* - souvent lié à la sphère publique et politique ; *fureur*, *rage*, *colère* - lié à l'expression ponctuelle et plus ou moins intense de l'affect ; enfin *énervement*, *irritation*, *exaspération* - qui concernent plutôt des états émotionnels précurseurs de cette manifestation. Ces cas montrent de façon assez éclairante le lien entre les valeurs sémantiques et la combinatoire lexico-syntaxique.

L'aspect exclusivement binaire des relations de dépendance directe peut aboutir à un rétrécissement du contexte des observations et faire manquer des phénomènes intéressants sur le plan phraséologique. Ces limitations empêchent notamment l'extraction automatique de séquences polylexicales à valeur d'unité minimale de sens (les « *extended units of meaning* » selon Sinclair 2004), qui peuvent présenter une variabilité considérable sur le plan de l'expression.

Cependant, en ce qui concerne les « collocations lexicales », Tutin (2008) affirme que la plupart d'entre elles ont une structure binaire, même pour celles qui s'étendent à plus de deux éléments, car elles correspondent sémantiquement à une structure prédicat-argument : « Les collocations peuvent être considérées comme des structures prédicat-argument, et comme telles sont, de façon prototypique, des associations binaires, où le prédicat est le collocatif et l'argument la base. La plupart des collocations ternaires (et au-delà) sont des collocations combinées (*collocational cluster*) ou récursives. »[68]

[68] "Collocations can be considered as predicate-argument structures, and as such, are prototypically binary associations, where the predicate is the collocate and the argument is the base. Most ternary (and over)

On note par ailleurs que de nombreux travaux dédiés à l'extraction de collocations étendues à plus de deux mots se basent sur des modèles binaires, appliqués à deux éléments composés : collocation d'arbres syntaxiques (Charest et al., 2010), construction itérative de cooccurrence multimots à partir de cooccurrences binaires (Seretan et al., 2003), ou encore calcul de mesure d'association multimots en combinant des mesures à deux termes.

De la même manière, il est possible d'étendre notre architecture pour le calcul des lexicogrammes d'un pivot donné, en la généralisant à des configurations plus complexes : la solution consiste à définir le pivot non plus seulement à partir d'une forme prise isolément, mais comme *une forme associée à un certain contexte lexico-syntaxique*. Une fois déterminé ce contexte, il est possible de calculer le tableau de contingence comme précédemment, le pivot et son contexte formant en quelque sorte une nouvelle unité pour laquelle il est possible de calculer à la fois les fréquences de cooccurrence (en se basant sur les relations du pivot) et la fréquence marginale dans le corpus.

Pour l'écriture des contextes, nous utilisons le formalisme de méta-expressions régulières déjà mentionné plus haut (p. 73). Par exemple, pour rechercher le pattern *avouer* + DET(poss.) + N, nous définissons le contexte suivant :

*pivot : #1= avouer_V*
*contexte : <c=N,#2> && <l=son,#3>::(obj,1,2)(det,2,3)*

Le calcul est seulement un peu plus long à mettre en œuvre, car les pivots multimots n'étant pas connus a priori, il n'est pas possible de les indexer tels quels. Seuls les tokens (formes ou lemmes) composant le contexte, ainsi que les relations de dépendance entre deux tokens définis, sont indexés, ce qui permet de réduire significativement l'ensemble des phrases à analyser. Pour des expressions comportant plusieurs relations, comme c'est l'intersection des phrases indexées pour chaque relation qui est retenue, la recherche est plus rapide : en d'autres termes, plus un pivot complexe est long, plus sa recherche est rapide. Dans le tableau 4.5 ci-dessous, on constate que pour le contexte donné en exemple, la mesure du *log-likelihood* fait clairement ressortir deux expressions récurrentes : *avouer son impuissance* et *avouer son admiration.*

| **l1** | **l2** | **f** | **f1** | **f2** | **loglike** |
|---|---|---|---|---|---|

collocations are merged collocations (collocational clusters) or recursive collocations."

| | | | | | |
|---|---|---|---|---|---|
| avouer_V | impuissance_N | 10 | 226 | 2868 | 142,0125 |
| avouer_V | admiration_N | 9 | 226 | 4016 | 119,8055 |
| avouer_V | crime_N | 6 | 226 | 26464 | 52,3355 |
| avouer_V | peur_N | 6 | 226 | 28357 | 51,5103 |
| avouer_V | faute_N | 5 | 226 | 15441 | 47,1415 |
| avouer_V | goût_N | 5 | 226 | 25267 | 42,2369 |
| avouer_V | participation_N | 5 | 226 | 28769 | 40,9463 |

*Tableau 4.5 : extrait de lexicogramme pour le pivot complexe* avouer son + N

Ainsi conçue, l'extraction des lexicogrammes pour les pivots complexes se veut surtout un outil d'observation permettant aux utilisateurs, par complexification progressive, de mieux préciser le contexte des phénomènes qui les intéressent (comme ici, en précisant la détermination ou la structure prépositionnelle). Par exemple, le corpus nous permet de constater que dans la plupart des cas, l'expression *avouer son admiration* attend la réalisation d'un troisième actant, le plus souvent introduit par la préposition *pour*.

#### 4.3.2.3. Extraction automatique d'expressions polylexicales

Cette approche qui va du simple vers le complexe peut néanmoins, d'une certaine manière, s'automatiser. Partant d'un pivot simple, on peut retenir ses collocatifs les plus saillants pour former de nouveaux pivots complexes. Et l'on peut réitérer l'opération de manière récursive sur les nouveaux pivots, jusqu'à une taille limite fixée arbitrairement. La figure ci-dessous montre comment un sous-arbre récurrent a été extrait pour identifier, de façon totalement automatique, l'expression *vouer une admiration sans borne*.

Julien Corman (2012), dans une recherche sous la direction d'Agnès Tutin et moi-même, a proposé une autre méthode pour l'extraction de ce type de sous-arbres récurrents dans un corpus. Plutôt que de construire les sous-arbres de façon progressive, en partant d'un pivot initial, la méthode proposée par Corman se base sur une énumération exhaustive de tous les sous-arbres syntaxiques jusqu'à une certaine taille *n* (comptée en nombre de nœuds) dans le corpus, avec le calcul d'un score d'association globale au niveau des sous-arbres, afin de retenir les récurrences significatives sur le plan statistique. Cette méthode est intéressante, et probablement plus complète dans ses résultats : théoriquement un sous-arbre peut obtenir un bon score d'association globale sans qu'il y ait une association statistique forte entre les éléments qui le compose pris deux à deux. Elle s'expose cependant à une explosion de la combinatoire pour des valeurs élevées de *n*, et nécessite une pré-indexation des résultats, et par conséquent un coût de stockage important nécessitant une optimisation des structures de

données[69]. Jusqu'à présent, nous avons préféré explorer la méthode itérative exposée ci-dessus, plus flexible, qui permet de fournir à la demande des sous-arbres d'une longueur non-bornée sans requérir d'autre prétraitement que l'analyse syntaxique en dépendance du corpus.

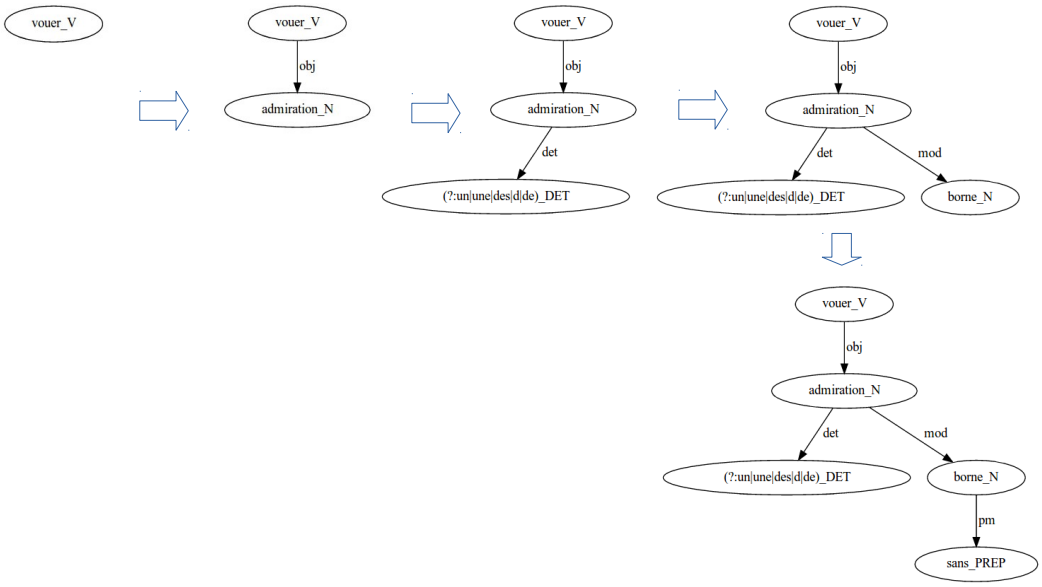


*Figure 4.10 : Extraction itérative d'une expression complexe*
(vouer une admiration sans borne)

Nous avons effectué une telle extraction pour le pivot *colère* pris en tant qu'objet direct, en ne retenant que les collocatifs obtenant un *loglike* supérieur ou égal à 5, et une fréquence de cooccurrence au moins égale à 3. On obtient la liste des expressions ci-dessous (partiellement lemmatisées et regroupées), qui constitue un « instantané » assez riche illustrant la combinatoire du pivot étudié :

69 Par exemple, en recourant à des tableaux de suffixe (Wing-Kai *et al.* 2003)

| | |
|---|---|
| provoquer la/une colère | tenter de calmer la colère |
| provoquer la colère des syndicats/du président/du gouvernement | pour calmer la colère |
| | calmer sa colère |
| l'annonce avait provoqué colère | attiser la colère |
| susciter la colère d'une partie | laisser éclater sa colère |
| susciter la colère des associations | manifester sa/leur colère |
| pour exprimer leur/sa colère | pour manifester leur colère |
| exprimer sa/leur/une colère | venir manifester leur colère |
| avoir exprimé hier colère | ne pas cacher sa/leur colère |
| déclencher la colère | crier sa/leur colère |
| piquer une/des colère/s | ravaler sa colère |
| apaiser la colère | ruminer sa colère |
| tenter d'apaiser la colère | contenir sa/la colère |
| pour apaiser la colère | avoir du mal à contenir colère |
| calmer la colère | déchaîner la colère |

*Tableau 4.6 : Liste des expressions polylexicales extraites pour* colère *pris en tant qu'objet direct (corpus de presse)*

On le voit, cette méthode permet d'extraire des collocations, comme *ruminer* + *colère*, mais aussi des constructions plus larges qui dépassent le cadre restreint d'une phraséologie qui se limiterait aux seuls critères du figement et de la non-compositionnalité. Une expression comme « avoir du mal à contenir sa colère » pourrait être identifiée comme une *routine*, typique d'un certain type de discours (ici le discours journalistique), s'inscrivant dans une conception élargie d'une phraséologie fondée sur les « deux seuls critères fédérateurs » que retient Tutin (2010 :179), à savoir « la polylexicalité et le caractère préconstruit et mémorisé (…) ».

Nous avons par la suite effectué plusieurs types de sondage avec notre méthode sur le corpus des noms d'affect. En effectuant l'extraction automatique des expressions polylexicales sur une cinquantaine de noms d'affect, nous avons noté que certaines expressions correspondaient à des schémas génériques très répandus pour l'ensemble de ces noms.

Par exemple, sur nos 39 pivots ayant suffisamment d'occurrences dans le corpus pour avoir permis d'extraire des expressions polylexicales, 15 ont été identifiés dans la construction *ne pas cacher* + *Det_poss* + *N* (avec les seuils de significativité que nous avons imposés, i.e. un nombre d'occurrences égal ou supérieur à 3 et un *loglike* supérieur à 5).

Cette construction semble donc assez générale dans ce champ sémantique. Si réciproquement, en partant de cette construction prise comme pivot complexe, on cherche tous les collocatifs nominaux en position d'objet direct, dans la même démarche que celle effectuée plus haut, alors on trouve non seulement une grande variété de noms d'affect, mais ces noms sont presque *tous* des noms d'affect (nous avons souligné les deux seuls intrus) :

*inquiétude, satisfaction, déception, admiration, ambition, joie, intention, agacement, scepticisme, sympathie, amertume, volonté, préférence, colère, intérêt, pessimisme, embarras, hostilité, irritation, enthousiasme, désir, exaspération, fierté, mécontentement, impatience, émotion, étonnement, souhait, soulagement, mépris, aversion, crainte, désarroi, jubilation, perplexité, plaisir, bonheur, réticence, préoccupation, envie, réserve, goût, doute, espoir, jeu*

On a donc trouvé une *construction*, dont les unités prises isolément ont peu à voir avec le sémantisme des affects, mais dont la cooccurrence avec les noms d'affect montre une grande spécialisation sémantique. Ce type de cooccurrence évoque ce que Stefanowitsch & Gries (2003) nomment des *collostructions*.

Il apparaît que dans ce type de construction, les variables déterminantes sont de nature grammaticale : ici le déterminant possessif et la négation.

Pour évaluer plus finement l'impact de ces variables, nous avons comparé les occurrences d'expressions de *cacher + Det_poss + N* avec et sans négation. Dans le premier comptage, nous avons éliminé toutes les expressions comportant une forme de négation sémantique au sens large : *ne pas cacher, ne jamais cacher*, *sans cacher*, *avoir du mal à cacher.* Dans le deuxième comptage nous n'avons retenu que les occurrences de *ne pas cacher + Det_poss + N*. Pour chacune de ces deux extractions, nous avons établis la liste des noms obtenus, sans appliquer de seuil de filtrage, et les avons classés en deux catégories : *Emo+* pour les noms possédant des occurrences correspondant à la classe des noms d'affect et *Emo-* pour les autres. Pour délimiter cette classe nous nous sommes inspiré de la définition de Tutin et al. (2006 : 32) :

> « La classe des noms d'affect (…) regroupe des noms pouvant se combiner avec les supports *avoir*, *ressentir* ou *éprouver* et apparaître en cooccurrence avec le nom *sentiment (de)*. Sémantiquement, ces noms caractérisent un processus psychologique plus que physique et requièrent obligatoirement un actant humain dans le rôle d'agent ou d'expérienceur. »

Tout d'abord, nous avons vérifié d'après le contexte de l'occurrence qu'il s'agissait bien d'un processus ou d'un état psychologique – et dans un second temps nous avons vérifié (à

partir de recherches sur le Web) la possibilité de trouver des occurrences de *sentiment de + N*, *ressentir + N* ou *éprouver + N*. On accepte par exemple les occurrences suivantes (en gras).

> *J'ai éprouvé un **vide**, le même **vide** que j'ai ressenti après ma première collection.*
>
> *Scène chaotique, traces omniprésentes, massacre sans contrôle apparent pouvaient induire un loup-garou, donc un individu mentalement perturbé, ayant du mal à cacher son **instabilité**, voire une personne déjà suivie psychiatriquement.*

En revanche l'occurrence suivante n'a pas été comptée dans la classe *Emo+*, bien que l'on rencontre des expressions telles que « éprouver une faute », ou un « sentiment de faute ». En effet le nom est polysémique, et l'occurrence précise ramenée par notre requête ne fait pas référence à un état psychologique :

> *Ils étaient jugés pour avoir désactivé les sécurités du tapis roulant peu avant le drame, et tenté de cacher leur **faute** en les réactivant tout de suite après.*

Certaines de nos extractions ont renvoyé des noms (comme *bouffée, sens, état, regain, liens*) faisant partie d'un syntagme nominal plus complexe (*bouffée de désir, sens de l'honneur, état de transe, état d'ivresse, regain d'angoisse, liens d'amitié*) : dans ces cas, nous avons substitué le nom support, par le nom de l'affect correspondant (*éprouver un lien d'amitié → éprouver + amitié*). En revanche, les occurrences de *éprouver un besoin de + V*, non pas été retenue pour la classe, bien qu'on puisse imaginer une acception de *besoin* correspondant à un affect. De même on n'a pas retenu des cas limites exprimant des états psychologiques compatibles avec nos différents tests : *conviction*, *désaccord*, *détermination,* ces unités appartenant à d'autres champs sémantiques connexes comme *opinion*, *volonté, intention, souhait, ...* Les noms de sensations physique (*brûlure*, *morsure*) n'ont pas été intégrés, sauf dans d'éventuels emplois métaphoriques. Enfin, on retient également dans *Emo+* les noms génériques, qui ne peuvent passer le test *sentiment de + N* : *sentiment, émotion, humeur, …*

Au final, avons constitué des listes de types[70], détachés de leurs occurrences, qui nous ont fourni les statistiques du tableau ci-dessous :

70 cf. Annexe - 3. p. 174

| Constructions | Emo+ | Emo- |
|---|---|---|
| *cacher + Det_poss + N sans négation* | 47<br>26,4 % | 131<br>73,6 % |
| *ne pas cacher + Det_poss + N* | 109<br>67,7 % | 51<br>32,3 % |

*Tableau 4.7 : Influence de la négation dans la construction* cacher + DetPoss + N *vis-à-vis de la classe des noms d'affect*

Il apparaît que la négation dans cette construction est plus fréquente que l'absence de négation, et qu'elle est fortement corrélée avec le champ sémantique des affects, avec presque 7 noms sur 10, tandis que le verbe *cacher* sans adverbe de négation n'est que faiblement lié à ce champ.

Nous avons également évalué le rôle de la détermination du nom dans la construction canonique *éprouver + N*, avec les articles définis, indéfinis, et éventuellement l'ajout d'un adjectif :

| Constructions | Emo+ | Emo- |
|---|---|---|
| *éprouver ArtDef + N* | 24<br>55,8 % | 19<br>44,2 % |
| *éprouver ArtIndéf + N* | 157<br>89,7 % | 18<br>10,3 % |
| *éprouver ArtIndéf + N + Adj* | 94<br>91,26 % | 9<br>8,74 % |

*Tableau 4.8 : Influence de la détermination dans la construction* éprouver + N

La construction avec l'article défini apparaît comme moins fréquente, et fonctionne le plus souvent avec des noms prédicatifs (*éprouver le besoin de + ..., éprouver le désir de..., éprouver le sentiment de ...*). Bien qu'elle soit liée au champ sémantique des affects, ce lien est assez lâche et représente seulement un peu plus de la moitié des occurrences. En revanche avec une détermination indéfinie, la construction est à la fois fréquente, diversifiée et hautement spécialisée, avec presque 90 % des noms. Si on ajoute un adjectif (en position d'épithète, antéposé ou postposé), la spécialisation est encore plus marquée, avec un peu plus de 91 % des noms de la liste.

Pour vérifier si d'autres constructions pouvaient déboucher sur ce type de paradigme, nous avons opéré une généralisation à partir des expressions polylexicales issues de

l'extraction automatique, en recherchant tous les noms apparaissant dans le même contexte. Par exemple, à partir de l'expression *en concevoir une amertume*, correspondant au sous-arbre de la figure 4.11, on considère le pivot complexe obtenu en substituant *amertume* par un nom quelconque, et on généralise la requête afin de chercher tous les collocatifs nominaux qui entrent en cooccurrence avec ce pivot.

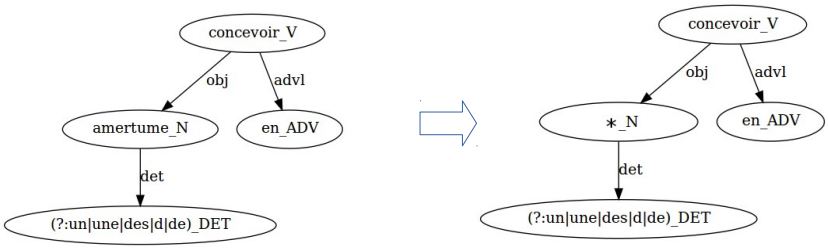


*Figure 4.11 : Généralisation d'une expression polylexicale dans une requête soumise à EmoConc*

Pour cet exemple précis, on obtient un paradigme assez restreint, et assez homogène sur le plan sémantique, avec une polarité plutôt négative :

***en concevoir un/une* + *N*** *: [amertume, chagrin, déception]*

Ces constructions ont donc des productivités variables, et délimitent des classes sémantiques plus ou moins restreintes, tout comme certains collocatifs simples – ce que note Tutin (2010 : 53), avec l'exemple de *grièvement,* qui manifeste une relative productivité en se combinant avec différentes bases telles que : *atteint*, *offensé*, *blessé*, *touché*, ...

D'autres expressions issues de nos extractions ont permis d'obtenir des résultats comparables (avec les seuils de significativité précédemment évoqués : *f* ≥ *3, loglike* > *5)* :

***pour éviter un/une nouveau* + *N*** *: [désillusion, déconvenue, dérapage, crise]*
***exprimer son/sa* + *N* à l'égard :** *[déception, défiance, compassion]*
***pour calmer le/la* + *N* :** *[colère, jeu, esprit, grogne, tension, surchauffe, ardeur, inquiétude, mécontentement, fronde, ire, impatience, douleur, crainte, prix, monde, crise]*
***ne pas cacher son/sa* + *N* de voir** *: [déception, satisfaction, souhait, espoir]*
***exprimer son/sa* + *N* de voir** *: [souhait, déception, satisfaction, désir]*
***laisser éclater sa* + *N*** *: [joie, colère]*

Toutes ces constructions affichent une nette attirance pour des noms d'affect (nous avons souligné les intrus) et présentent chacune des traits sémantiques particuliers rendus manifestes par ces différents paradigmes (aspect ponctuel, polarité, attente, etc.). Ces

paradigmes manifestent par ailleurs des préférences sémantiques marquées : p. ex. dans la classe *[souhait, déception, satisfaction, désir]* on voit se dessiner une idée d'*attente*, satisfaite ou non.

Enfin, on trouve également des expressions très récurrentes mais caractéristiques d'un nom en particulier, et typique du genre textuel. Par exemple toutes les occurrences ci-dessous ont été identifiées sur la partie journalistique de notre corpus (le nombre d'occurrences figure entre parenthèse) :

*provoquer la colère des syndicats (16)*
*provoquer la colère de l'opposition (5)*
*provoquer la colère des habitants (4)*
*provoquer la colère d'une partie (4)*
*provoquer la colère des salariés (4)*
*provoquer la colère des autorités (4)*
*provoquer la colère du gouvernement (4)*
*provoquer la colère du président (4)*

Certaines expressions stéréotypées affichent encore un degré de spécialisation supérieur, comme **pour ne pas connaître une désillusion**, qui n'apparaît dans notre corpus que dans les articles sportifs :

> *Mais une fois l'ennui d'un match à zéro essai digéré, il faut reconnaître que, face à ces Argentins toujours aussi pénibles et embrouilleurs, il valait mieux remiser ambitions offensives et grain de folie* ***pour ne pas connaître une*** *nouvelle et cruelle* ***désillusion****.*
>
> *Finalement, le Stade Rennais n'aurait- il pas intérêt à gérer au mieux sa fin de saison tout en s'activant à préparer la suivante* ***pour ne pas connaître une*** *nouvelle grande* ***désillusion****?*
>
> *Mais les hommes d'Oswald Tanchot devront se montrer prudents* ***pour ne pas connaître une*** *réelle* ***désillusion****, devant une équipe qui a, selon le coach vitréen,« un fort potentiel offensif ».*

Pour voir si ces constructions polylexicales pouvaient fonctionner comme marqueur du genre textuel, nous avons examiné la répartition des constructions précédemment listées au sein des deux composantes du corpus *Emolex*, littéraire et journalistique. Afin d'équilibrer ces deux composantes, nous n'avons utilisé que le sous-corpus *Libe07* (16 741 875 de tokens) d'une taille comparable à celle du corpus littéraire (15 978 230 de tokens).

Le tableau ci-dessous montre les répartitions obtenues pour quelques expressions (nous n'avons retenu que celles qui obtenaient plus de 10 occurrences dans le corpus). Seules les occurrences concernant les noms d'affect ont été considérées.

| Constructions | Presse | Littéraire |
|---|---|---|
| *ne pas cacher DetPoss + N* | 79<br>76 % | 25<br>24 % |
| *en concervoir un/une + N* | 0<br>0 % | 10<br>100 % |
| *pour calmer le/la + N* | 8<br>44,4 % | 10<br>55,6 % |
| *exprimer DetPoss + N* | 62<br>51,7 % | 58<br>48,3 % |

*Tableau 4.9 : Répartition des constructions en fonction du genre, dans un échantillon de 16 millions de mots du corpus Emolex (articles de presse vs romans)*

Certaines constructions apparaissent comme étant très spécialisées : c'est le cas de *en concevoir un/une + N*, qui ne s'observe que dans les romans.

Les constructions génériques, plus fréquentes globalement, ont des profils différents : *ne pas cacher DetPoss + N* est assez nettement caractéristique des articles de presse, tandis que *exprimer DetPoss + N* apparaît comme neutre vis-à-vis du genre. Toutefois, une observation plus fine des occurrences montre que ces constructions génériques sont employées dans des contextes assez marqués suivant le genre : dans la presse *ne pas cacher DetPoss + N* s'emploie typiquement avec un modifieur du nom d'affect (p. ex. « il ne cache pas sa sympathie pour la candidate socialiste ») ou du verbe (p. ex. « il ne cache pas son scepticisme par rapport à sa propre expertise »), tandis que les textes littéraires privilégient les tournures où le nom n'est pas modifié et apparaît en fin de phrase, avec 17 occurrences sur 25 (p. ex. « L'infirmière ne cachait pas son admiration. » ), ce que montrent les statistiques du tableau 4.10 :

| Constructions | Presse | Littéraire |
|---|---|---|
| *ne pas cacher DetPoss + N + modifieur* | 23<br>85,2 % | 4<br>14,8 % |
| *ne pas cacher + adv + DetPoss + N* | 52<br>88,1 % | 7<br>11,9 % |
| *ne pas cacher DetPoss + N* (fin de phrase) | 21<br>54 % | 18<br>46 % |

*Tableau 4.10 : Répartition des constructions avec* ne pas cacher DetPoss + N *en fonction du genre (articles de presse vs romans)*

Pour l'autre construction générique ici étudiée, *exprimer DetPoss + N,* on observe également des contextes d'apparition fortement liés au genre : la construction apparaît de préférence en fin de phrase dans les textes littéraires (p.ex. « Elle rougit, comme si son cerveau se servait de son apparence pour exprimer sa déception. »), et ceci de façon très marquée (52 occurrences sur 58), alors que c'est plutôt rare dans les articles de presse (4 occurrences sur 62). Par ailleurs, certaines constructions moins fréquentes, comme *exprimer Det Poss + N + à l'égard de,* semblent rester l'apanage des textes de presse.

| Constructions | Presse | Littéraire |
|---|---|---|
| *exprimer Det Poss + N* (fin de phrase) | 4<br>7,1 % | 52<br>92,9 % |
| *exprimer Det Poss + N + à l'égard de* | 3<br>100 % | 0<br>0 % |

*Tableau 4.11 : Répartition des constructions avec* ne pas cacher DetPoss + N *en fonction du genre (articles de presse vs romans)*

L'étude des profils combinatoires nous a donc conduit à l'identification de constructions polylexicales récurrentes, qui mettent en évidence le caractère préconstruit des discours, illustrant ainsi le principe de l'idiome de Sinclair. Plus précisément, les outils développés dans le cadre du projet Emolex nous ont permis de mettre en évidence les différentes facettes des « *extended units of meaning* » décrites par Sinclair (1996), à savoir les *collocations* (associations privilégiées entre mots), les *colligations* (associations privilégiées entre les mots et leur environnement syntaxique, p. ex. détermination, actance, complémentation, etc.), les *préférences sémantiques* (« *semantic preference* ») illustrées par les paradigmes restreints liés à certaines constructions, et la *prosodie sémantique* (« s*emantic prosody* ») manifestant, au

niveau du discours, une certaine polarité (p. ex. la polarité négative associée à *en concevoir un/une + N*).

L'étude de la combinatoire nous a donc permis d'identifier des unités étendues jouant un rôle précis sur le plan discursif, notamment par l'inscription plus ou moins marquée dans des usages liés au genre textuel. Comme le résume très justement Tutin (2010 : 180) : « La combinatoire lexicale et la phraséologie au sens large constituent une porte d'entrée particulièrement intéressante pour l'analyse du discours, en permettant de recontextualiser le lexique. »

# 5. Perspectives

Au terme de cette synthèse, il me paraît difficile de terminer par des conclusions : celles-ci seraient de nature à donner un caractère de point final à mes différentes recherches entamées depuis presque vingt ans, et j'ai plutôt été frappé, en revenant sur mes travaux, par leur état d'inachèvement. Cet état est certes inhérent à la recherche scientifique, qui ne cesse de progresser et de se renouveler, mais plus encore aux aspects ingénieriques de ces recherches – qui m'ont pris tant de temps au fil des ans. Que l'on constitue des corpus ou des logiciels, ces produits de l'ingénieur – linguiste ou informaticien – vieillissent mal, et leurs imperfections sont toujours gênantes – bogues, fonctionnalités limitées, coquilles, erreurs d'annotation, méta-données manquantes, taille insuffisante du corpus, documentation lacunaire… Ces défauts sont trop visibles pour ne pas être source de frustration, mais ces frustrations sont compensées sans doute par la satisfaction concrète de voir ces outils, et ces données, servir à des collègues dans leur propre recherche, au moins pour un temps.

Au-delà de ces outils, l'exercice de cette synthèse m'a appris à reconnaître, voire à découvrir mes propres obsessions : pour paraphraser le titre d'un article de Michael Stubbs (2009) en hommage à John Sinclair, je pense qu'une grande part de mon énergie a été consacrée à la « recherche des unités de sens ». Partant de l'étude des corpus parallèles, je me suis assez tôt rendu compte que la notion de compositionnalité traductionnelle permettait parfois de révéler (et parfois aussi d'occulter, comme nous l'avons vu) les unités de la langue et du discours.

L'idée de la multi-textualité a été un de mes principaux guides : invoquer plusieurs langues, si possible un grand nombre, permet de démultiplier cet effet révélateur de la traduction, et d'établir des réseaux de correspondance convergents, susceptibles de mettre en lumière les unités et les valeurs sémantiques. La traduction, en tant qu'elle est une forme de paraphrase, s'inscrit de plain-pied dans la sémantique – d'une sémantique à la fois structurale et référentielle, puisqu'elle permet de briser l'autarcie du « système où tout se tient ».

Cette intuition, je n'ai réussi à la mettre en œuvre que de façon embryonnaire : en reprenant des travaux laissés en suspens il y a une dizaine d'années, j'ai montré qu'au niveau superficiel de l'alignement phrastique, le recours à la multi-textualité permettait de consolider les réseaux de correspondances. Mais si l'on s'intéresse aux unités de sens, la situation devient plus complexe, car leur contour est fuyant : l'extraction des cliques à partir des dictionnaires, tout comme à partir d'un petit corpus multilingue, montre que la difficulté de mettre en correspondance les unités autour de sens partagés, en passant par l'équivalence traductionnelle, se heurte au problème du contour même de ces unités. La perception naïve qu'un sens (ou plusieurs sens) puisse(nt) être attaché(s) à un mot est une illusion, un effet prototypique : c'est ce que nous ont appris nos échecs dans la tentative d'extraire des cliques cohérentes.

Je ne quitte par pour autant le terrain de la traduction et de la multi-textualité : le développement rapide d'outils comme *Linguee* montre qu'il y a un vrai besoin dans le domaine de la lexicographie et des aides à la traduction. Je pense que l'articulation entre le dictionnaire, d'une part, qui donne une structuration normative du sens et du régime des unités, et des corpus, d'autre part, qui révèlent par analogie toute la richesse de la combinatoire et des valeurs sémantiques implicites, est une piste de recherche très prometteuse. Je garde à l'esprit les projets que nous avions échafaudés avec Agnès Tutin en 2006, et espère pouvoir poursuivre nos développements dans cette direction, en intégrant des corpus parallèles ou comparables. Notamment, les méthodes d'extraction automatique d'expressions ou de constructions polylexicales, décrites dans la partie 4 de cette synthèse, pourraient s'intégrer avec bénéfice dans des applications d'aide à la rédaction.

Un autre parti-pris réitéré avec force dans cette synthèse est l'absence de frontière marquée entre corpus parallèles et comparables. En revendiquant leur complémentarité, nous

cherchons également à réinscrire les phénomènes traductionnels dans l'horizon de la linguistique de corpus. Comme le note très justement Rastier (2006) :

> *La traduction pourrait enfin révéler la linguistique à elle-même . Il faudrait éviter une disciplinarisation autonome de la traductologie, car sa vocation reste de renouveler la linguistique de l'intérieur : la question de la traduction peut et doit y devenir centrale dès lors qu'on quitte la problématique du signe pour celle du texte. Elle permet en effet de réintroduire pleinement l'activité interprétative dans la communication linguistique, en ouvrant la voie à sa reconception comme une interaction au sein du texte et de l'intertexte.*

La dimension textuelle peut notamment être abordée sous le prisme du genre, central chez Rastier. Or, mes recherches les plus récentes, encore très « inchoatives » sur ce terrain, indiquent une relation entre expressions, constructions polylexicales et appartenance générique : l'inscription du texte dans un genre se manifeste également par le recours à des expressions pré-construites dont la valeur fonctionnelle est conventionnelle et fortement stéréotypée. L'exploration de ce lien fait partie de nos projets à court terme, notamment dans le cadre du dépôt d'un projet ANR-DFG sous la direction d'Iva Novakova et de Dirk Siepmann. Ce projet, baptisé *PhraséoRom*, vise à étudier de façon comparative, la phraséologie et les motifs textuels (au sens de Longrée et Mellet, 2013), à travers un corpus d'œuvres littéraires appartenant à des sous-genres bien codifiés (roman policier, science-fiction, etc.). Dans ce cadre, nous avons l'intention de développer la technique d'extraction de constructions polylexicales, en nous basant sur une généralisation des expressions polylexicales récurrentes. De la sorte, nous espérons pouvoir aller plus loin dans l'étude phraséologique, en identifiant des constructions typiques des genres, et peut-être aussi des auteurs. Par exemple, Legallois (2012) relève le motif *il V le NC de DETPOSS NC ADJ,* caractéristique du style de Zola selon le calcul des spécificités utilisé en textométrie (en comparaison avec un corpus de romans du XIX$^{e}$ s.). Il cite les exemples suivants *:* « Elle ne parla plus, elle s'abattit près du brancard, dont *elle écarta les toiles de ses mains tremblantes*. » ou « Alors, ils cessèrent de rire, penchés au-dessus de la Bible antique, dont *elle tournait les pages, de ses doigts minces*. ») On le voit, ces motifs sont difficilement repérables à l'œil nu : la perspective de pouvoir fournir aux linguistes, et aux spécialistes de littérature, des outils pour mettre au jour des régularités cachées me paraît enthousiasmante. Par ailleurs, sur le plan linguistique, une extraction automatique des constructions permettrait également d'identifier automatiquement des patrons de sous-catégorisation, qui pourraient guider les travaux des lexicographes en structurant plus facilement les sorties de concordances

pour une entrée lexicale donnée (par exemple, dans la perspective de la *Corpus pattern analysis*, cf. Hanks, 2004).

Une dernière piste à court terme concerne le développement d'outils pour l'interrogation. Les sorties de l'extraction automatique d'unités polylexicales donnent lieu à l'affichage de sous-arbres, accompagnés d'occurrences en contextes et d'information complémentaires, comme le paradigme des noms qui peuvent se substituer au pivot au sein de la même construction, ainsi que le montre la figure suivante :

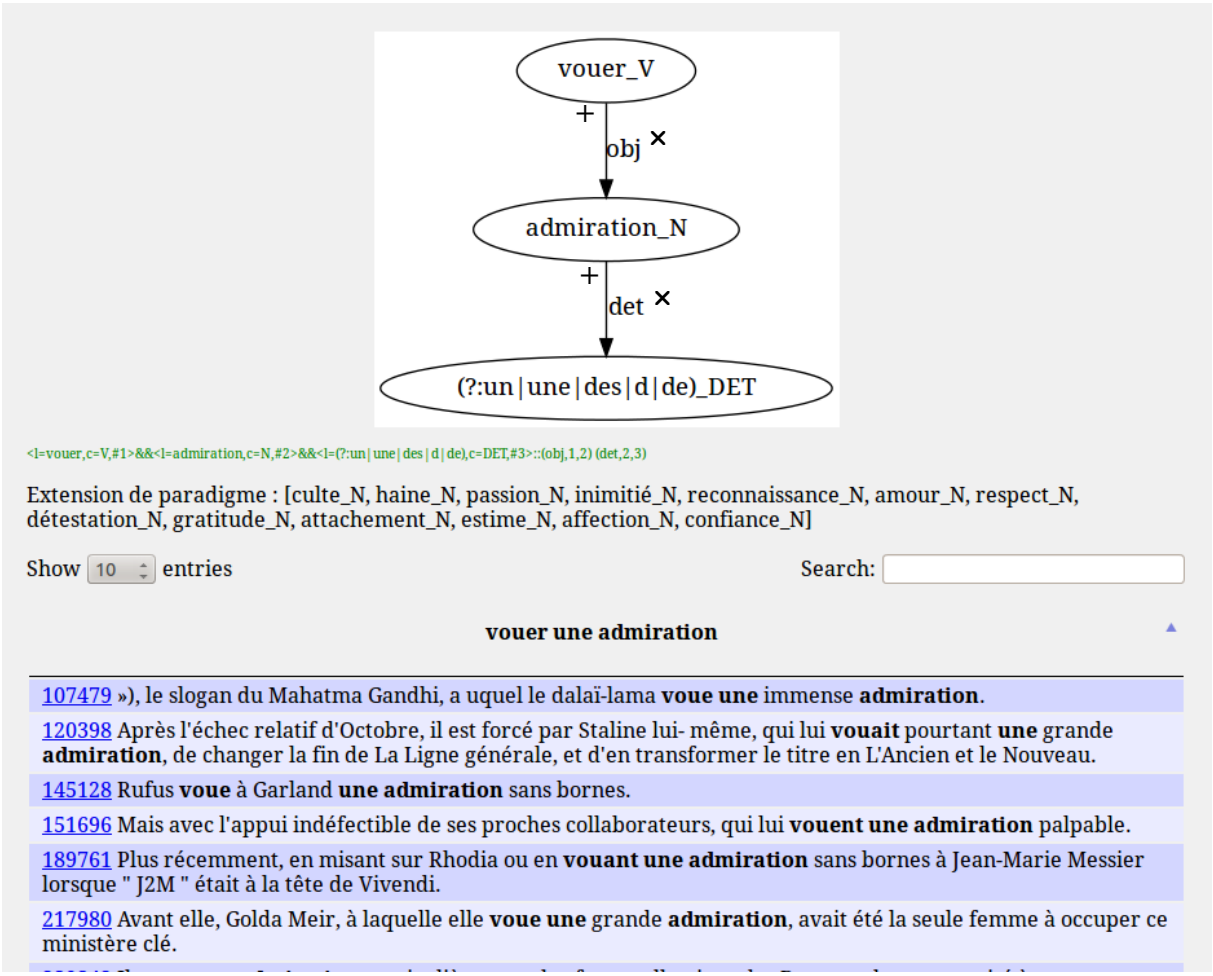


*Figure 5.1 : Extractions d'expression polylexicales et affichage statique des résultats*

Or, l'architecture actuelle d'EmoConc permet une recherche rapide, à la demande, de ce genre de sous-arbre. Je prévois donc de rendre dynamique ce type d'affichage. L'utilisateur, en cliquant sur un nœud, pourra voir la liste des unités et catégories susceptibles d'y figurer (comme ici la liste *culte*, *haine*, *passion*, etc. qui peut se substituer à *admiration*). En sélectionnant une unité, un sous-groupe d'unités ou une catégorie, il pourra modifier dynamiquement le sous-arbre, et les sorties correspondantes. Il pourra également modifier les

relations et développer (ou contracter) l'arbre en ajoutant (ou en supprimant) des relations, par simple clic sur le signe + (ou la croix ×) rajouté sur la figure. Cette interface est en cours de développement : elle constitue selon moi une évolution intéressante des outils d'interrogation, dans la continuité des progrès apportés par le développement de l'interface de Scientext. Ce type de dispositif vise à *faire voir* des phénomènes jusqu'à présent invisibles, car noyés dans la masse de corpus trop grands pour être embrassés du regard de manière critique. Si l'on considère l'« instrument » comme « objet technique qui permet de prolonger et d'adapter le corps pour obtenir une meilleure perception » (Simondon, 1989 :114, cité par Habert, 2005), peut-être y verra-t-on une étape supplémentaire dans le développement d'un instrument utilisé pour l'exégèse depuis le Moyen Âge, le concordancier (Kraif, 2011).

Comme l'écrivait Wittgenstein (1961 : 125) dans ses *Investigations philosophiques* : « il est d'innombrables et diverses sortes d'utilisation de tout ce que nous nommons "signes", "mots", "phrases". Et cette diversité, cette multiplicité n'est rien de stable, ni de donné une fois pour toutes ; mais de nouveaux types de langage, de nouveaux jeux de langage naissent, pourrions-nous dire, tandis que d'autres vieillissent et tombent en l'oubli... » Wittgenstein, pour sa part, explorait les fondations logiques du langage au moyen d'un type de « jeu » très particulier, les expériences de pensées. Pour qui s'intéresse au langage et à l'informatique, le TAL appliqué à de grands volumes de données langagières apparaît comme le moyen de démultiplier les perspectives sur cet objet à la fois intime et méconnu qu'est la langue, et à en objectiver de nouvelles facettes. Avec le développement de cette instrumentation, peut-être assiste-t-on à une nouvelle manière de « jouer » avec le langage, mais empiriquement cette fois, afin d'en approfondir la connaissance ?

# 6. Références


Abdulhay, A. (2006). *Le repérage et l'alignement d'entités nommées dans un corpus bilingue français - arabe, Mémoire de master*, sous la dir. de Olivier Kraif, Université Stendhal Grenoble 3, 107 pp.

Abdulhay, A., Kraif, O. (2013). Constitution d'une ressource sémantique arabe à partir de corpus multilingue aligné, *Actes de TALN 2013*, Les sables d'Olonnes, pp. 299-312.

Aït-Mokhtar, S., Chanod, J.-P., Roux C. (2002). Robustness beyond Shallowness: Incremental Deep Parsing, *Natural Language Engineering*, 8 :121-144.

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé T., Ponton C. (2005). Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO, *Alsic Apprentissage des Langues et Systèmes d'Information et de Communication, Vol. 8*, pp. 65-79, [URL : http://alsic.u-strasbg.fr/v08/antoniadis/alsic_v08_04-rec4.htm, consulté en juin 2014].

Atkins, S. (1993). Theoretical Lexicography and its relation to Dictionary-making. In: *Dictionaries: the Journal of the Dictionary Society of North America,* (guest editor) W. Frawley, DSNA, Cleveland Ohio. pp. 4-43.

Attardi, G., Dell'Orletta, F., Simi, M., Chanev, A., Ciaramita, M. (2007). Multilingual Dependency Parsing and Domain Adaptation using DeSR", in *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague.

Barlow, M. (1996). Parallel texts in language teaching. In S. Botley, J. Glass, A. M. McEnery, & A. Wilson (Eds.), *Proceedings of teaching and language corpora 1996* (UCREL Technical Papers Volume 9; pp. 45-56). Lancaster, UK: University Centre for Computer Corpus Research on Language.

Barlow, M. (2008). Parallel texts and corpus-based contrastive analysis, In: Gómez González, M., Mackenzie, L. and González Alvarez, E. (eds.), *Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives.,* Benjamins, 101-121.

Benigno, V., Hancock J., Pawlak K., Kraif O. (2014). The use of academic collocations in essays in a test of academic English, *LTRC 2014*, 4 - 6 June 2014, VU University : Amsterdam

Berman, A. (1988). De la translation à la traduction, in *TTR : traduction, terminologie, rédaction*, vol. 1 n°1, pp. 23-40

Bertels, A., Verlininde, S. (2011). La lexicographie et l'analyse de corpus : nouvelles perspectives, *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 56, n° 2, 2011, pp. 247-265.

Blanchard, A., Kraif, O., Ponton, C. (2009). Mastering Overdetection and Underdetection in Learner-Answer Processing: Simple Techniques for Analysis and Diagnosis. *Calico Journal*. Vol. 26, No. 3 (May 2009).

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, vol. 19, n. 2, pp. 263-311.

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

Catford, J. C. (1965) *A Linguistic Theory of Translation*. London : Oxford University Press.

Caviglia, F. (2005). Students' diverse appreciation of text corpora as writing aids. In T. Caudery (ed.), *Proceedings of the Ninth Nordic Conference for English Studies* (NAES 2004), Aarhus, Denmark, 27-29 May 2004.

Chang, J. J. S., Ker, S. J. (1996). Aligning More Words with High Precision for Small Bilingual Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, *COLING-96*, Copenhagen, 5-9 August 1996.

Charest, S., Brunelle, E., Fontaine, J. (2010). Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multilexémiques, *Actes de TALN 2010,* Montréal, pp. 19-23 juillet 2010.

Chen, B., El-Bèze, M., Haddara, M., Kraif, O., Moreau de Montcheuil, G. (2005). Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale, *Actes de TALN-RECITAL 2005*, 6-10 juin 2005, Dourdan, vol. 1, pp. 415-420.

Chiao, Y.-C., Kraif, O., Laurent, D., Nguyen, T. M. H.,Semmar, N., Stuck, F., Véronis, J., Zaghouani, W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project, *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, Genova, May 2006.

Chuquet, H., Paillard, M. (2004). *Approche linguistique des problèmes de traduction anglais-français*, Collection OPHRYS TRADUCTION, Ophrys : Paris.

Church, K. W. (1993). Char align : A program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, *ACL-93*, Columbus Ohio, pp. 1-8.

Church, K. W., Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Machine Translation*, vol. 16, n. 1, pp. 22-29.

Corman, J. (2012). *Extraction d'expressions polylexicales sur corpus arboré*, Mémoire de master 2 sous la dir. d'Agnès Tutin et Olivier Kraif, Université Grenoble-Alpes. [URL : http://dumas.ccsd.cnrs.fr/docs/00/70/48/73/PDF/CORMAN_Julien_M2R.pdf, consulté en juin 2014].

Corréard, M.-H. (1998). Traduire avec un dictionnaire, traduire pour un dictionnaire. In Thierry Fontenelle, Philippe Hiligsmann, Archibald Michiels, André Moulin, Siegfried Theisse (eds) *Euralex'98 Proceedings*, Plenary Lectures, Vol.1, Liège, pp. 17-24. [URL: http://www.euralex.org/proceedings-toc/euralex_1998-1/]

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

Daille, B., Gaussier, E., Langé, J. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Actes, COLING'94*, p. 515–521

Danielsson, P., Ridings, D. (1997). Practical presentation of a "vanilla" aligner. Presented at the *TELRI Workshop on Alignment and Exploitation of Texts*. Institute Jožef Stefan, Ljubljana. [URL : http://nl.ijs.si/telri/Vanilla/doc/Ijubljana/, consulté en juin 2014].

Davis, M. W., Dunning T. E., Ogden W. C. (1995). Text Alignment in the Real World : Improving Alignments of Noisy Translations Using Common Lexical Features. In *Proceedings of EACL 95*, 8 p. [URL : http://www.crl.nmsu.edu, consulté en juin 2014].

Débili, F., Sammouda, E. (1992). Appariement des phrases de textes bilingues Français - Anglais et Français - Arabe. In *Proceedings of the 14th International Conference on Computational Linguistics*, *COLING-92*, Nantes, 23-28 août 1992, pp. 518-524.

Déjean, H., Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*.

Di Spaldro, J., Auger, P., Ladouceur, J. (2010) Le calque technoscientifique : un procédé néologique avantageux pour la terminologie française?, *Neologica*, no 4, 2010, pp. 163-184

Diab, M., Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora, in *Proc. of ACL-02*, Philadelphia.

Diab, M., Hacioglu K., Jurafsky D. (2007). *Arabic Computational Morphology: Knowledge based and Empirical Methods*, chapter 9, A. Soudi, A. van den Bosch et G. Neumann (Eds.), Springer, pp. 159–179.

Drouin, P., Doll, F. (2008) Quantifying Termhood Through Corpus Comparison, *Proceedings of Terminology and Knowledge Engineering (TKE-2008)*, pp. 191–206, Copenhagen Business School, Copenhagen.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, Morristown, NJ, vol. 19, n. 1, pp. 61-74.

Duchet, J.-L., Kraif, O., Torrellas Castillo, M. (2008) Corpus massifs et corpus bilingues alignés : leur impact sur la recherche linguistique. *Bulletin de la Société de Linguistique de Paris*, t. CIII, fasc. 1, pp. 129-150.

Evert, S. (2007). Corpora and collocations. in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Evert, S. & The OCWB Development Team (2010). *CQP Query Language Tutorial, The IMS Open Corpus Workbench (CWB),* 17 February 2010 [URL : http://cwb.sourceforge.net/, consulté en juin 2014].

Falaise, A., Tutin, A., Kraif, O. (2012). Une interface pour l'exploitation de corpus arborés par des non informaticiens : la plate-forme ScienQuest du projet Scientext, *TAL*, Volume 52, n° 3, pp. 241-246.

Fleury, S. (2009). Exploration du corpus Traductions alignées du discours d'investiture de B. Obama in André Salem, Serge Fleury (sous la dir. de) *Explorations textométriques, Volume 3 : corpus multilingues*, Université Paris 3 Sorbonne Nouvelle.

Fung, P. (2000). A statistical view on bilingual lexicon extraction - From parallel corpora to non-parallel corpora. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 11, p. 18.

Fung, P., Church, K. W. (1994). K-vec : A New approach for Aligning Parallel Texts. In *Proceedings of the 15th International Conference on Computational Linguistics*, *COLING-94*, Kyoto, pp. 1096-1102.

Gadamer, H.-G. (1960). *Vérité et Méthode*, Le Seuil, Paris.

Gale, W., and Church, K. (1991). "A Program for Aligning Sentences in Bilingual Corpora," *Association for Computational Linguistics,* pp. 177-184 [URL : http://aclweb.org/anthology//P/P91/P91-1023.pdf, consulté en juin 2014].

Gaussier, E., Langé, J.-M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *T.A.L.*, vol. 36, n. 1-2, pp. 133-155.

Goffin, R. (1994). L'eurolecte : oui, jargon communautaire : non, *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 39, n° 4, 1994, p. 636-642.

Gougenheim, G., Michea, R., Rivenc, P., Sauvageot, A., (1964). *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Didier, Paris.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., Zampa, V. (2007). Integrating learner corpora and natural language processing: a crucial step towards reconciling technological sophistication and pedagogical effectiveness, *Recall*, Vol. 19, N° 3, pp. 252-268.

Greimas, A. J., Courtès, J. (1993). *Sémiotique*, Paris, Hachette, Coll. HU linguistique.

Grundy, V. (1996). L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue, in H. Béjoint, P. Thoiron, Claude Boisson, *Les dictionnaires bilingues*, De Boeck Supérieur, pp. 127-149.

Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Texto!* [en ligne], décembre 2005, vol. X, n°4. URL : http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html, consulté en juin 2014.

Haddara, M., Kraif, O. (2005). Etude de contextes multilingues alignés en vue de la désambiguïsation sémantique, *Actes des 4èmes Journées de la Linguistique de Corpus*, Lorient, 15-17 septembre 2005

Hanks, P. (2004). Corpus Pattern Analysis, in G. Williams & S. Vessier (eds) *Proceedings of the 11th Euralex International Congress*, Université de Bretagne Sud, Lorient, pp. 87-98.

Harris, B. (1988). Are you Bi-Textual ? *Language Technology*, n° 7, pp. 41-41.

Heiden, S., Tournier, M. (1998). Lexicométrie textuelle, sens et stratégie discursive, actes *I Simposio Internacional de Análisis del Discurso*, Madrid.

Hjelmslev, L. (1971). *Essais linguistiques*, Paris :Editions de Minuit.

Hoey, M. (2005) : *Lexical Priming: A New Theory of Words and Language*, London : Routledge.

Hunston, S., Francis, G. (2000) *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*, Studies in Corpus Linguistics, John Benjamins.

Isabelle, P. (1992). La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. *META*, Outremont, PQ, XXXVII, 4, pp. 721-731.

Jakobson, R. (1963). Aspects linguistiques de la traduction. *Essais de linguistique générale*, Paris, Les éditions de Minuit, pp. 78-86.

Johns, T. (1986). Microconcord : a language learner's research tool, *System*, 14/2.

Johns, T. (1991). Should you be persuaded: two examples of data driven learning. *Classroom Concordancing*, *ELR Journal* (New Series), Tim Johns & Philip King (eds) vol. 4.1, n°16.

Kandel, L., Moles, A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision*, 19 :253–274

Kay, M., Röscheisen, M. (1993). Text-Translation Alignement. *Computational Linguistics*, Morristown, NJ, vol. 19, n. 1, pp. 121-142.

Kilgariff, A.,Tugwell, D. (2001) WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, *Proc ACL workshop on COLLOCATION Computational Extraction Analysis and Exploitation*, Toulouse July 2001.

König, E., Lezius, W. (2003). *The TIGER language - A Description Language for Syntax Graphs, Formal Definition. Technical report IMS*, Universität Stuttgart, Germany. [URL: http://www.wolfganglezius.de/lib/exe/fetch.php?media=cl:tigerlangform.pdf, consulté en juin 2014]

Kraif, O. (2001a). *Constitution et exploitation de bi-texte pour l'aide à la traduction*, Thèse de doctorat, Université de Nice.

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

Kraif, O. (2003a) Repérage de traduction et commutation interlingue : Intérêt et méthodes, *Actes de TALN 2003*, Batz-sur-Mer, 11-14 juin 2003, tome 2, pp. 127-138.

Kraif, O. (2003b) From Translational Data to Contrastive Knowledge: Using Bi-text for Bilingual Lexicons Extraction, *International Journal of Corpus Linguistics*, June 2003, vol. 8, iss. 1, John Benjamins, pp. 1-29(29).

Kraif, O. (2004). Propositions pour l'intégration d'outils TAL aux dispositifs informatisés d'apprentissage des langues, in Christian Degache (sous la dir. de), *Intercompréhension en langues romanes, LIDIL*, N° 28, Université Stendhal, Grenoble, pp. 153-165

Kraif, O. (2008a). Alignement multilingue pour l'étude contrastive : outils et applications, in Marie Hédiard (a cura di) *Linguistica dei corpora, Strumenti e applicazioni*, Edizioni dell'Università degli Studi di Cassino, pp. 83-99.

Kraif, O. (2008b). Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles,* Presses universitaires de Lyon, vol. 2, pp. 625-634

Kraif, O. (2011) Les concordances pour l'observation des corpus : utilité, outillage, utilisabilité, In Jean Chuquet (sous la dir. de) *Le langage et ses niveaux d'analyse*, Presses universitaires de Rennes (PUR), chap. 4, pp. 67-80.

Kraif, O., Chen, B. (2004). Combining clues for lexical level aligning using the Null hypothesis approach, *Proceedings of Coling 2004*, Geneva, August 2004, pp. 1261-1264.

Kraif, O., Diwersy, S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques, *Actes de la conférence TALN 2012*, Grenoble, pp. 399-406.

Kraif, O., & Diwersy, S. (2014). Explorating combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions. In P. Blumenthal, I. Novakova, & D. Siepmann (éd.), *Les émotions dans le discours - Emotions in Discourse.* Berlin, Allemagne: Peter Lang.

Kraif, O., El-Bèze, M., Meyer, R., Richard, C. (2006). Le corpus Carmel : un corpus multilingue de récits de voyage, *7th Conference on Teaching and Language Corpora: TaLC7*, Université Paris 7, Paris.

Kraif, O., Ponton C. (2007). Du bruit, du silence et des ambiguïtés : que faire du TAL pour l'apprentissage des langues ?, *Actes de TALN 2007*, Toulouse, 12-15 juin 2007, pp. 43-151.

Kraif, O., Tutin, A., Diwersy, S. (2014) Extraction de pivots complexes pour l'exploration de la combinatoire du lexique : une étude dans le champ des noms d'affect, *Actes du Congrès Mondial de Linguistique Française 2014*, 19-23 juillet 2014, Berlin.

Kraif, O., Tutin, A. (2006) Des corpus bilingues alignés annotés sémantiquement pour l'aide à la rédaction: application aux collocations de la langue scientifique générale, Aide à la rédaction - *Apports du Traitement Automatique des Langues, Journée d'étude l'ATALA*, Paris [URL : http://perso.limsi.fr/amax/recherche/atala06/, consulté en juin 2014].

Kraif, O., Tutin, A. (2011). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. In Natalie Kübler (Ed.) *Corpora, Language, Teaching, and Resources: From Theory to Practice. Selected papers from TaLC7, the 7th Conference of Teaching and Language Corpora*. coll. Etudes contrastives. Bruxelles: Peter Lang.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* 29 (1): 1–27.

Lamy, M-N., Mortensen, H. J. K.(2012). Using concordance programs in the Modern Foreign Languages classroom. Module 2.4 in Davies G. (ed.) *Information and Communications Technology for Language Teachers (ICT4LT)*, Slough, Thames Valley University [URL : http://www.ict4lt.org/en/en_mod2-4.htm, consulté en juin 2014].

Landure, C. (2011). Data-Driven Learning : apprendre et enseigner à contre-courant, *Mélanges CRAPEL, numéro spécial : Pratiques d'accompagnement(s) des apprenants en présentiel et à distance*, n° 32, pp. 163-178

Langé, J.-M., Gaussier, É. (1995). Alignement de corpus multilingues au niveau des phrases. *T.A.L.*, vol. 36, n. 1-2, pp. 67-79.

Langlais, P., El-Bèze, M. (1997). Alignement de corpus bilingues : algorithmes et évaluation. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 191-197.

Langlais, P., Simard, M., Véronis, J., Armstrong, S., Bonhomme P., Débili, F., Isabelle, P., Souissi, E., Théron, P. (1998). ARCADE: A co-operative research project on bilingual text alignment. In *Proceedings of First International Conference on Language Resources and Evaluation* (*LREC*), *Granada*, *Spain*, *28-30 May 1998*, pp. 289-292.

Laplace, C. (1994) *Théorie du langage et théorie de la traduction*, Paris : Didier érudition.

Le Serrec, A., L'Homme, M.-C., Drouin, P., Kraif, O. (2010). Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment, *Terminology* 16(1), pp. 77-106.

Loiseau, M., Antoniadis, G., & Ponton, C. (2010). Pratiques enseignantes et « contexte pédagogique » dans le cadre de l'indexation pédagogique de textes, in *Actes du Congrès Mondial de Linguistique Française*, La Nouvelle-Orléans, Etats-Unis, pp. 12-15 Juillet 2010. [URL : http://www.linguistiquefrancaise.org/index.php?option=com_article&access=doi&doi=10.1051/cmlf/2010233&Itemid=129, consulté en juin 2014]

Longrée, D., Mellet, S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours, Dans D. Legallois et A. Tutin (sous la dir. de) *Vers une extension du domaine de la phraséologie, Revue Langages*, n° 189, 2013/1, Armand Colin : Paris, pp. 65- 79.

Mahimon, M.-D. (1999) *Identification des équivalences traductionnelles sur un corpus Français / Anglais*, Mémoire de DEA, sous la dir. de Jean Véronis, Université de Provence Aix-Marseille 1, Aix-en-Provence.

Mallak, I. (2011). *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information*. Thèse de doctorat à l'Université Toulouse III - Paul Sabatier.

McEnery, A. M., Oakes, M. P. (1995). Sentence and word alignment in the CRATER project : methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop*, Dublin.

Melamed, I. D. (1997). A Word-to-Word Model of Translational Equivalence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, 7-12 July 1997, pp. 490-497.

Melamed, I. D. (1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, 1-2 August 1997, pp. 97-108 [URL : http://www.cis.upenn.edu/~melamed/home.html, consulté en juin 2014].

Melamed, I. D. (1998). Empirical Methods for MT Lexicon Development. In *Proceedings of AMTA-1998*, 13 p. [URL : http://www.cis.upenn.edu/~melamed/home.html, consulté en juin 2014].

Melnikova, E., Novakova, I., Kraif, O. (2009) Quels corpus pour l'analyse contrastive ? L'exemple des constructions verbo-nominales de sentiment en français et en russe. *Actes des 6èmes Journées de la Linguistique de Corpus* (disp. à l'adresse : http://www.licorn-ubs.com/jlc6/ACTES/Melnikova_etal_JLC09.pdf).

Miao, J., Salem, A. (2009). Comparaisons textométriques de traductions franco-chinoises in André Salem, Serge Fleury (sous la dir. de) *Explorations textométriques, Volume 3 : corpus multilingues*, Université Paris 3 Sorbonne Nouvelle

Moreau de Montcheuil, G., Chen B., El-Bèze, M., Kraif, O. (2004). Using a Word Sense Disambiguation system for translation disambiguation: the LIA-LIDILEM team experiment, *in Proceedings of Senseval3 Workshop*, Barcelona, june 2004, pp. 175-178.

Morin, E. Daille, B. (2011). Bilingual Terminology Mining from Comparable Corpora. In S. Sharoff, R. Rapp, P. Zweigenbaum, P. Fung, *BUCC: Building and Using Comparable Corpora*, Springer.

Morin, E., Daille, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN-RECITAL 2012)*. Long paper. pages 141-154, Grenoble

Morin, E., Dufour-Kowalski, S. Daille, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables, *Actes de TALN 2004*, Fès, 19–21 avril 2004. [URL : http://www.atala.org/taln_archives/TALN/TALN-2004/taln-2004-long-013.pdf (consulté en juin 2014]

Motoc, D. (2002). Traduction et création. De la re-création du texte littéraire traduit à la créativité du processus traducteur, in *Actes de l'Arches*, Tome 4 [URL : http://www.arches.ro/revue/no04/no4art07.htm, consulté en juin 2014]

Och, F.-J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

Ploux, S. (2007). Enrichir automatiquement des dictionnaires électroniques de synonymes et de traduction : une application du modèle d'appariement multilingue des Atlas sémantiques *Actes des 2èmes journées d'animation scientifique régionales « Élaborer des dictionnaires en contexte multilingue »*, Tunis.

Ploux, S., Ji. H. (2003). A model for matching semantic maps between languages (French/English, English/French). *Computational Linguistics*, vol. 29, no. 2, p. 155–178.

Pouliquen, B., Steinberger, R. (2007). Acquisition and Use of Multilingual Name Dictionaries. pp. 1-10. *Proceedings of the Workshop Acquisition and Management of*

*Multilingual Lexicons (AMML'2007)* held at RANLP'2007. Borovets, Bulgaria, 26 September 2007.

Pouliquen, B., Steinberger, R., Ignat, C., Temnikova, I., Widiger, A., Zaghouani, A., Žižka J. (2005). Multilingual person name recognition and transliteration. *CORELA - Numéros thématiques | Le traitement lexicographique des noms propres*. Publié en ligne le 02 décembre 2005.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In Actes, ACL'99, p. 519–526

Rastier, F. (1990). La triade sémiotique, le trivium et la sémantique linguistique. *Nouveaux Actes sémiotiques*, Paris.

Rastier, F. (2006). La traduction : interprétation et genèse du sens , dans Marianne Lederer et Fortunato Israël, éds. *Le sens en traduction*, Paris, Minard, 2006]

Read, J. (2000). *Assessing vocabulary*. Cambridge, Cambridge University Press

Rézeau, J. (2007). L'apport du concordancier à l'analyse et à la remédiation des erreurs des apprenants dans les forums de discussion en ligne, *Alsic*, Vol. 10, n° 2 | 2007, document alsic_v10_04-pra1, mis en ligne le 15 décembre 2007 [URL : http://alsic.revues.org/561, Consulté le 02 juillet 2014. ; DOI : 10.4000/alsic.561]

Rosch, E. (1975). Cognitive Representations of Semantic Categories, *Journal of Experimental Psychology*: General, Vol.104, No.3, (September 1975), pp. 192–233.

Ruhlen, M. (1994). *On the Origin of Languages: Studies in Linguistic Taxonomy.* Stanford: Stanford University Press.

Saâdane, H., Semmar, N. (2012). Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. *Actes TALN 2012,* 127-140.

Sagot, B., Fišer, D. (2008). Building a Free FrenchWordNet fromMultilingual Resources. *Proceeding of Ontolex*, Marrakech, Maroc.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Seretan, V., Nerima, L., Wehrli, E. (2003). Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. *Proceedings of the Fourth International Conference on Recent Advances in NLP*, (RANLP-2003), 424–431.

Shei, C.C., Pain, H. (2000). An ESL Writer's Collocational Aid. *Computer Assisted Language Learning (CALL)*. 13(2): 167-182.

Simard, M. (1998). The BAF : A Corpus of English-French Bitext. *First International Conference on Language Resources and Evaluation*, Granada, Espagne, pp. 489-494.

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

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

Simard, M., Plamondon, P. (1996). Bilingual Sentence Alignment : balancing robustness and accuracy. In *Proceedings of AMTA-96*, Montréal, Canada, pp. 135-144.

Simondon, G. (1989). *Du mode d'existence des objets techniques. L'invention philosophique.* Aubier, Paris, 3ème edition. Première édition : 1958.

Sinclair, J. (1991). *Corpus, concordance, collocation,* Oxford University Press.

Sinclair, J. (1996). EAGLES, Preliminary recommendations on Corpus Typology, EAG--TCWG--CTYP/P, Version of May, 1996 [URL : http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html, consulté en juin 2014]

Sinclair, J. (1996). The search for units of meaning. *Textus online only*. 9, N. 1, 1000-1032.

Sinclair, J. (2004). *Trust The Text , Language, Corpus and Discourse* , Routledge (Taylor and Francis).

St.John, E. (2001). A Case For Using A Parallel Corpus And Concordancer For Beginners Of A Foreign Language, *Language Learning & Technology*, Vol. 5, No. 3, September 2001, pp. 185-203.

Stefanowitsch, A., Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2: 209-243.

Stubbs, M. (2009) Memorial Article: John Sinclair (1933–2007) The Search for Units of Meaning, Sinclair on Empirical Semantics, *Applied Linguistics*, 30 (1) : 115-137.

Tapanainen, P., Järvinen, T. (1997). A non-projective dependency parser, In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, pp. 64-74.

Teubert, W. (1996) Comparable or Parallel Corpora? *International Journal of Lexicography,* 9 (3): 238-264.

Teubert, W. (2005). My version of corpus linguistics, *International Journal of Corpus Linguistics*, vol. 10-1, pp. 1-13.

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

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work,* Amsterdam and Philadelphia: John Benjamins.

Torellas Castillo, M. (2009). *Les interférences linguistiques dans les textes en espagnol des institutions de l'Union Européenne : étude fondée sur le corpus bilingue massif aligné de l'acquis communautaire*. Thèse de doctorat, sous la dir. de J.L. Duchet, Université de Poitiers.

Tufis, D., Ion, R., Ide, N. (2004). Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, Geneva, 2004.

Tutin, A., Novakova, I., Grossmann, F., Cavalla, C. (2006) : Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires. *Langue française*, 150, 32-49.

Tutin, A. (2007). Autour Du Lexique et de La Phraséologie Des Écrits Scientifiques. *Revue Française de Linguistique Appliquée Lexique et écrits scientifique* (XII(2)) : 5–14.

Tutin, A. (2008). For an extended definition of lexical collocations. *Proceedings Of Euralex*. Université Pompeu Fabra, Barcelone, 15-19 juillet 2008.

Tutin, A. (2010). *Dans cet article, nous souhaitons montrer que…*Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines, *Lidil*, 41 | 2010, 15-40.

Tutin, A. (2010). *Sens et combinatoire lexicale : de la langue au discours,* Synthèse d'HDR, Université Stendhal Grenoble 3.

Véronis, J. (2000). From the Rosetta Stone to the information society : A survey of parallel text processing. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 1, 24 p.

Véronis, J., Hamon, O., Ayache, C., Belmouhoub, R., Kraif, O., Laurent, D., Nguyen, T., Semmar, N., Stuck. F, Zaghouani, W. (2008). La campagne d'évaluation ARCADE 2, in Stephane Chaudiron, Khalid Choukry (sous la dir. de) *L'évaluation des technologies de traitement de la langue*, Hermès, Lavoisier, Paris, pp. 47-69.

Vinay, J.-P., Darbelnet, J. (1958) *Stylistique comparée du français et de l'anglais*, Paris : Didier.

Vossen, P. (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. *Computational Linguistics*, Volume 25, Number 4.

Wang Lixun (2001) Exploring Parallel Concordancing In English And Chinese , *Language Learning & Technology* , September 2001, Vol. 5, Num. 3 , pp. 174-184

Wing-Kai, H., Kunihiko, S., Wing-Kin, S. (2003) Breaking a time-and-space barrier in constructing full-text indices, *Proceedings of Foundations of Computer Science*, 44th Annual IEEE Symposium on Computer Science.

Wittgenstein, L. (1953, 1958). *Philosophical Investigations*, translated by G.E.M. Anscombe, Basil Blackwell, New York, Macmillan]

Wittgenstein, L. (1961) *Tractatus Logico-Philosophicus*, *suivi de investigations philosophiques*, trad. de Pierre Klossowski, Paris, Gallimard.

# Annexe

## Annexe - 1. Activités de bi-concordance proposée par Joseph Rézeau

[URL : http://www.uhb.fr/campus/joseph.rezeau/concord.htm, consulté en mai 2007]

### Traduction de ON en anglais

### Exercice 1 : Repérage

La *grammaire anglaise de l'étudiant* de Berland-Delépine donne les traductions suivantes :

a) la voix passive (traduction la plus courante)
b) les pronoms *we*, *you*, *they*
c) *people* (pluriel) ou *somebody* (singulier) pour un sujet inconnu ou non précisé
d) le pronom indéfini *one*, dans un style soigné … ton un peu sentencieux
e) *there is* + nom à sens verbal *(there was a knock at the door*

Classez les traductions de **on** des 20 exemples ci-dessous, tirés au hasard du ***Petit Prince*** dans ces 5 catégories.

La voix passive est-elle la plus courante ?

Combien de traductions non classables dans les 5 catégories avez-vous trouvées ?

Conclusions

| | |
|---|---|
| *1. **On** en avale une par semaine et l'**on** n'éprouve plus le besoin de boire.* | You need only swallow one pill a week, and you would feel no need of anything to drink. |

| | |
|---|---|
| *2. Il faut s'astreindre régulièrement à arracher les baobabs dès qu'**on** les distingue d'avec les rosiers auxquels il ressemblent beaucoup quand ils sont très jeunes.* | You must see to it that you pull up regularly all the baobabs, at the very first moment when they can be distinguished from the rose-bushes which they resemble so closely in their earliest youth. |
| *3. **On** épargne cinquante-trois minutes par semaine.* | With these pills, you save fifty-three minutes in every week. |
| *4. Quand **on** veut faire de l'esprit, il arrive que l'**on** mente un peu.* | When one wishes to play the wit, he sometimes wanders a little from the truth. |
| *5. Quand le mystère est trop impressionnant, **on** n'ose pas désobéir.* | When a mystery is too overpowering, one dare not disobey. |
| *6. Ils répètent ce qu'**on** leur dit…* | They repeat whatever one says to them… |
| *7. Donc, quand la moralité de l'explorateur paraît bonne, **on** fait une enquête sur sa découverte.* | Then, when the moral character of the explorer is shown to be good, an inquiry is ordered into his discovery." |
| *8. **On** note d'abord au crayon les récits des explorateurs.* | The recitals of explorers are put down first in pencil. |
| *9. **On** attend, pour noter à l'encre, que l'explorateur ait fourni des preuves.* | One waits until the explorer has furnished proofs, before putting them down in ink. |
| *10. **On** s'assoit sur une dune de sable. **On** ne voit rien.* | One sits down on a desert sand dune, sees nothing, hears nothing. |
| *11. **On** risque de pleurer un peu si l'**on** s'est laissé apprivoiser…* | One runs the risk of weeping a little, if one lets himself be tamed… |
| *12. C'est dur de se remettre au dessin, à mon âge, quand **on** n'a jamais fait d'autres tentatives que celle d'un boa fermé et celle d'un boa ouvert, à l'âge de six ans!* | It is hard to take up drawing again at my age, when I have never made any pictures except those of the boa constrictor from the outside and the boa constrictor from the inside, since I was six. |
| *13. **On** disait dans le livre : "Les serpents boas avalent leur proie tout entière, sans la mâcher.* | In the book it said: "Boa constrictors swallow their prey whole, without chewing it. |
| *14. C'est très utile, si l'**on** est égaré pendant la nuit.* | If one gets lost in the night, such knowledge is valuable. |
| *15. S'il s'agit d'une brindille de radis ou de rosier, **on** peut la laisser pousser comme elle veut.* | If it is only a sprout of radish or the sprig of a rose-bush, one would let it grow wherever it might wish. |
| *16. Voici mon secret. Il est très simple: **on** ne voit bien qu'avec le coeur.* | "And now here is my secret, a very simple secret: It is only with the heart that one can see rightly; |
| *17. -Tu sais…quand **on** est tellement triste **on** aime les couchers de soleil…* | "You know - one loves the sunset, when one is so sad…" |
| *18. Tantôt je me dis: "**On** est distrait une fois ou l'autre, et ça suffit!* | But at another time I say to myself: "At some moment or other one is absent-minded, and that is enough! |
| *19. Quand **on** a terminé sa toilette du* | "When you've finished your own toilet in the |

| | |
|---|---|
| *matin, il faut faire soigneusement la toilette de la planète.* | morning, then it is time to attend to the toilet of your planet, just so, with the greatest care. |
| *20. Or un baobab, si l'**on** s'y-prend trop tard, **on** ne peut jamais plus s'en débarrasser.* | A baobab is something you will never, never be able to get rid of if you attend to it too late. |

## Exercice 2 : Complétez les traductions de *on* (en vous aidant de vos constatations de l'exercice 1)

| | |
|---|---|
| *1. Quand **on** veut un mouton, c'est la preuve qu'**on** existe"* | "If _____ wants a sheep, that is a proof that _____ exists." |
| *2. Je désire que l'**on** prenne mes malheurs au sérieux.* | I like my misfortunes _____ seriously. |
| *3. Il ne répondait jamais aux questions, mais, quand **on** rougit, ça signifie "oui", n'est-ce pas?* | He never answered questions - but when _____ flushes does that not mean 'Yes'? |
| *4. Car **on** peut être, à la fois, fidèle et paresseux.* | For it is _____ for a _____ to be faithful and lazy at the same time. |
| *5. Car je n'aime pas qu'**on** lise mon livre à la légère.* | For I do not want _____ to read my book carelessly. |
| *6. Mais s'il s'agit d'une mauvaise plante, il faut arracher la plante aussitôt, dès qu'**on** a su la reconnaître.* | But when it is a bad plant, _____ must destroy it as soon as possible, the very first instant that _____ recognizes it. |
| *7. **On** voit sur la Terre toutes sortes de choses…* | "On the Earth _____ sees all sorts of things." |
| *8. **On** pourrait entasser l'humanité sur le moindre petit îlot du Pacifique.* | All humanity could be _____ up on a small Pacific islet. |
| *9. -**On** n'est jamais content là où l'**on** est, dit l'aiguilleur.* | "No _____ is ever satisfied where _____ is," said the switchman. |
| *10. -**On** ne sait pas, lui dit le roi.* | "_____ do not know that," the king said to him. |
| *11. -**On** ne sait jamais, dit le géographe.* | "_____ never knows," said the geographer. |
| *12. -**On** ne connaît que les choses que l'**on** apprivoise, dit le renard.* | "_____ only understands the things that one tames," said the fox. |
| *13. **On** est un peu seul dans le désert…* | "_____ is a little lonely in the desert…" |
| *14. -**On** est seul aussi chez les hommes, dit le serpent.* | "_____ is also lonely _____ men," the snake said. |
| *15. -Les étoiles sont belles, à cause d'une fleur que l'**on** ne voit pas…* | "The stars are beautiful, because of a flower that cannot _____." |
| *16. Ils perdent du temps pour une* | "They waste their time over a rag doll and it becomes |

| | |
|---|---|
| *poupée de chiffons, et elle devient très importante, et si **on** la leur enlève, ils pleurent…* | very important to them; and if _____ takes it away from them, they cry… |
| *17. -Et quand tu seras consolé (**on** se console toujours) tu seras content de m'avoir connu.* | "And when your sorrow is comforted (_____ soothes all sorrows) you will be content that you have known me. |
| *18. -Droit devant soi **on** ne peut pas aller bien loin…* | "Straight ahead of _____, _____ can go very far…" |
| *19. -Comment peut-**on** posséder les étoiles?* | "How is it possible for _____ to own the stars ?" |
| *20. C'est pour saluer quand **on** m'acclame.(dit le roi)* | "It is to raise in salute when _____ acclaim me. |

---

## FOR + Groupe Nominal + TO-INFINITIF

[URL : http://www.uhb.fr/campus/joseph.rezeau/concord.htm, consulté en mai 2007]

### Exercice 1 : Repérages

a) Dans la plupart des citations anglaises suivantes, le GN qui suit **for** est sujet de l'infinitif qui suit.

Surlignez en jaune les citations dans lesquelles le GN **n'est pas** sujet de l'infinitif.

b) Dans les citations françaises, soulignez les différentes traductions des expressions comportant FOR + GN + INFINITIF. Lorsquelles sont traduites par un verbe, à quel temps est-il en général ?_____

| | |
|---|---|
| *1. Here he sat down, his back to the bank, waiting **for sleep to come*** | Il s'assit ensuite, le dos au talus, et attendit le sommeil, |
| *2. "But now I must tell you something. If you want to go right away the best thing is **for you to go sick**.* | "Mais, maintenant, je vais vous expliquer: au cas où vous voudriez partir tout de suite, le mieux serait que vous vous fassiez porter malade. |
| *3. As he waited **for night to come**, Giovanni stayed and watched the northern steppe.* | En attendant la tombée de la nuit, Giovanni resta à regarder la plaine septentrionale. |
| *4. They were waiting **for the dark to attack**.* | ils attendaient l'obscurité pour attaquer. |
| *5. The horse, had detected the presence of men in the direction of the Fort and was now waiting **for them to bring** it some forage.* | le cheval, demeuré seul, était allé à la recherche du salut, il avait senti la présence de l'homme du côté du fort et attendait maintenant qu'on lui apportât de l'avoine. |

| | |
|---|---|
| *6. As he read the officers stared at him, looking **for something to show** itself in his face.* | Les officiers ne le quittaient pas du regard pendant qu'il lisait, cherchant à deviner sur son visage quelque chose. |
| *7. They got there first and there's nothing left **for us to do** here-but we would look remarkably silly."* | Ils sont arrivés les premiers et nous, nous n'avons plus rien à faire ici, mais nous aurions bonne mine si nous partions!" |
| *8. But I wouldn't count too much on that. It only needs another two years to pass -only two years-and it would be too much of an effort **for you to go back**."* | Laissez seulement passer deux années encore, rien que deux années suffisent, et vous en aller vous coûtera un trop gros effort. |
| *9. He advanced into the courtyard and looked about him with apparent anxiety, searching **for someone to tell something to**.* | Il s'était avancé dans la cour et regardait autour de lui presque avec anxiété, en quête de quelqu'un à qui dire quelque chose. |
| *10. But there was no longer Simeoni's telescope **for him to see them with**.* | Mais la longue-vue de Simeoni, qui permettait de les voir, n'était plus disponible. |
| *11. The days turn into months and the months into years and soon it is time **for Aurora to return** to her parents.* | Les jours, les mois, les années passent et Aurore qui va bientôt avoir seize ans doit être rendue à ses parents |
| *12. All heads of a household were entitled to receive 160 acres in return **for the right to live** on the land for five consecutive years.* | Chaque chef de famille peut se voir attribuer 160 acres de terre à condition d'avoir résidé sur le domaine pendant cinq années consécutives. |

## Exercice 2 Complétez les citations anglaises

| | |
|---|---|
| *1. Grown-ups never understand anything by themselves, and it is tiresome **for** _____ **to** ____ **always and forever** _____ **things** to them.* | Les grandes personnes ne comprennent jamais rien toutes seules, et **c'est fatigant, pour les enfants, de toujours et toujours leur donner des explications**. |
| *2. For it is possible **for** _____ **to** _____ faithful and lazy at the same time.* | Car **on peut être**, à la fois, fidèle et paresseux. |
| *3. All the stars will pour out fresh water **for** _____ **to** _____..."* | Toutes les étoiles **me verseront à boire**… |
| *4. From there the desert stretches to the rocky cone of the New Redoubt, even and compact enough **for** _____ **to** _____ freely.* | De là jusqu'au cône rocheux de la Nouvelle Redoute, le désert s'étend uniforme et compact, **comme pour permettre à l'artillerie d'avancer sans encombre**. |
| *5. then he sat in his office and could _____ wait **for** _____ **to** _____ so that he might throw himself into an easy chair or on to his bed.* | Assis ensuite dans son bureau, **il lui tardait de voir arriver le soir** pour pouvoir se jeter dans un fauteuil ou sur son lit. |
| *6. Tronk ... pointed out sharply that it was _____ **for** _____ **to** | Tronk, … démontra sèchement à Lazzari qu'il était impossible **que son cheval se** |

_____ ***run away*** *- to get into the northern valley it would have had to jump the walls of the Fort or cross the mountains.*

**fût échappé**: pour passer dans la vallée du Nord, il eût fallu que l'animal traversât les remparts du fort ou franchît les montagnes.

*7. "How is it* _____***for*** _____ ***to*** _____ *the stars ?"*

-Comment peut-on posséder les étoiles?

# Annexe - 2. Composition des corpus comparables DE-Source et FR-Source

## DE-Source

| Oeuvre originale | Année de trad. | Titre traduction | Traducteur | Taille DE | Taille FR |
|---|---|---|---|---|---|
| Bernhard, Thomas (1985) Alte Meister | 1988 | Maîtres anciens | Gilberte Lambrichs | 70080 | 85084 |
| Dönhoff, Marion Gräfin von (1988) Kindheit in Ostpreußen | 1988 | Une enfance en prusse orientale | Colette Kowalski | 56677 | 78045 |
| Göhre, Frank (1993) St.-Pauli-Nacht | 1996 | La nuit de St. -Pauli | Patrick Kermann | 37777 | 45390 |
| Jelinek, Elfriede (1983) Die Klavierspielerin | 1983 | La pianiste | Y. Litaize, M. Hoffmann | 40420 | 40150 |
| Kirchhoff, Bodo (1991) Infanta | 1991 | Infanta | Bernard Lortholary | 106510 | 121805 |
| Martin R. Dean (1997) Die Ballade von Billie und Joe | 1997 | La ballade de Billie et Joe | Sibylle Müller | 165972 | 198975 |
| Rosendorfer, Herbert (1977) Stephanie und das vorige Leben | 1991 | Stéphanie et la vie antérieure | Françoise Saint-Onge | 25234 | 29619 |
| Rosendorfer, Herbert (1991) Die Wiederentdeckung des Gehens beim Wandern. Harzreise. | 1991 | La meilleure façon de marcher. Voyage dans le Harz | Maryse Julien, Robert Jacob | 19697 | 24981 |
| Roth, Josef (1990) Orte | 1990 | Croquis de voyage | Jean Ruffet | 41047 | 49447 |
| Schmitter, Elke (2000) Frau Sartoris | 2000 | Madame Sartoris | Anne Weber | 42882 | 51521 |
| Süskind, Patrick (1987) Die Taube | 1987 | Le pigeon | Bernard Lortholary | 22485 | 27324 |
| Suter, Martin (2000) Die dunkle Seite des Mondes | 2000 | La face cachée de la lune | Olivier Mannoni | 78428 | 96273 |
| Vanderbeke, Birgit (1999) Ich sehe, was du nicht siehst | 1999 | Devine ce que je vois | Anne Weber | 31245 | 38566 |
| Violet, Bettina (1996/1993) Das wilde Löwenkind | 1993 | Le sauvage enfant-Lion | E. Neiter, G. Mange | 18515 | 21944 |
| | | | Total | 756969 | 909124 |

## FR-Source

| Oeuvre originale | Année de trad. | Titre traduction | Traducteur | Taille DE | Taille FR |
|---|---|---|---|---|---|
| Aubert, Brigitte (2001) Descentes d'organes | 2002 | Nachtlokal | Mitglieder des Kollektivs Druck-Reif | 58837 | 63576 |
| Boissard, Janine (1998) Marie-Tempête | 2002 | Der Ruf des Meeres | Weidmann, Angelika | 129940 | 135428 |
| Buron, Nicole de (1998) Chéri, tu m'écoutes ? Alors répète ce que je viens de dire… | 1999 | Liebling, hörst du mir zu? | Riek, Walther | 70744 | 73591 |
| Cauwelaert, Didier van (1997) La vie interdite | 2002 | Auf Seelenspitzen | Heinemann, Doris | 101972 | 107580 |
| Châtelet, Noëlle (1996) La dame en bleu | 1997 | Die Dame in Blau | Wittmann, Uli | 23182 | 23351 |
| Decoin, Didier (1994) Docile | 1996 | Die schöne Buchhändlerin | Reitz, Barbara | 99132 | 107430 |
| Dorin, Françoise (1997) Les vendanges tardives | 1998 | Späte Früchtchen | Filius-Jehne, Christiane & Schoelzel, Christiane | 77748 | 79425 |
| Dormann, Geneviève (1993) La petite main | 1995 | Die Gespielin | Kuhn, Irene | 91622 | 92201 |
| Dorner, Françoise (2006) La douceur assassine | 2007 | Die letzte Liebe des Monsieur Armand | Gersch, Christel | 28333 | 30140 |
| Echenoz, Jean (1999) Je m'en vais | 2002 | Ich gehe jetzt | Schmidt-Henkel, Hinrich | 52893 | 57559 |
| Nothomb, Amélie (1999) Stupeur et tremblements | 2000 | Mit Staunen und Zittern | Krege, Wolfgang | 30880 | 31143 |
| Nothomb, Amélie (2002) Robert des noms propres | 2003 | Im Namen des Lexikons | Krege, Wolfgang | 32676 | 34123 |
| Pouy, Jean-Bernard (1986) La pêche aux anges | 1988 | Geld für kleine Engel | Bahr, Elke | 39690 | 40915 |
| Vargas, Fred (2002) Coule la Seine | 2007 | Die schwarzen Wasser der Seine | Schock, Julia & Scheffel, Tobias | 28540 | 29134 |
| | | | Total | 866189 | 905596 |

## Annexe - 3. Types de noms apparaissant dans diverses constructions

### *cacher + DetPoss + N* sans négation :

**Classe Emo - :**

actes, activité, argent, arme, barriques, beauté, béquilles, bras, butin, cadavre, camion, cancer, Candidat, cartes, chemise, cheveux, cicatrices, cocards, compromissions, comptes, consommation, corps, cou, couleur, crime, décisions, découvertes, défauts, démission, dents, dépendance, déroute, difficultés, documents, drame, droits, DVD, échec, économies, enfants, épouse, faciès, faute, femme, feuilles, figure, fils, flétrissure, foi, forfait, fortune, fragilité, fragilités, fric, fusils, génie, gouttière, grossesse, grossesses, homosexualité, identité, ignorance, implication, inaptitude, incompétence, intention, jambes, jeu, liaisons, liens, lunettes, machine, mains, maisons, maîtresse, marchandises, marijuana, marteaux, métier, micro, misère, mode, mouchoir, nature, notes, nudité, orientation, origine, partie, passé, patronymes, paupières, pauvres, pensée, performances, petit-fils, pieds, poitrine, positions, présence, procédures, produit, profondeur, promise, racines, responsabilité, revenus, romantisme, secrets, seins, séropositivité, sexe, sommeil, statut, surnom, téléphone, tête, textes, trésor, vide, visage, volonté, vulnérabilité

**Classe Emo + :**

ambition, amertume, amours, angoisse, appétit, blessures, bonheur, chagrin, colère, déception, dépit, désespoir, désir, détresse, dissensions, embarras, émoi, émotion, enthousiasme, envie, état, étonnement, exaspération, faiblesses, fierté, goûts, honte, inquiétude, intentions, intérêt, joie, larmes, lassitude, mépris, ressentiment, révolte, satisfaction, sentiments, tourments, triomphe, trouble

### *Ne pas cacher + DetPoss + N*

**Classe Emo - :**

âge, allégeance, appartenance, approche, arrières-pensées, but, chevelure, cible, concordance, convictions, critiques, démesure, désaccord, dette, difficultés, divisions,

engagement, faiblesse, fantasmes, foi, grossesse, homosexualité, identité, idéologie, idylle, influences, intention, interrogations, jeu, objectifs, opinion, opposition, orientations, orientations, passé, patriotisme, proximité, relation, reproches, rêve, rupture, sensibilité, souhait, soutien, valeurs, visage, vocation, volontarisme, volonté, vote, yeux

**Classe Emo + :**

admiration, affection, affinités, affliction, agacement, allégresse, ambition, amertume, amitié, amour, angoisse, animosité, appétit, appréhension, attrait, aversion, bienveillance, bonheur, colère, consternation, contentement, crainte, craintes, cupidité, curiosité, déception, découragement, dédain, défiance, dégoût, dépit, désappointement, désarroi, désenchantement, désir, détestation, doutes, embarras, émerveillement, émotion, ennui, enthousiasme, envie, espoir, états d'âme, étonnement, exaspération, excitation, fascination, fatigue, fidélité, fierté, frustration, fureur, gêne, goût, hâte, hésitations, hostilité, humeur, impatience, impressions, inclination, incompréhension, incrédulité, inquiétude, intérêt, irritation, joie, jubilation, larmes, lassitude, mal-être, malaise, mécontentement, méfiance, mépris, morosité, nostalgie, obsession, optimisme, passion, perplexité, pessimisme, plaisir, prédilection, préférence, préoccupation, rancœur, ras-le-bol, regrets, réserves, réticence, satisfaction, saudade, scepticisme, semi-déception, sentiments, soulagement, stupéfaction, stupeur, surprise, sympathie, tendresse, tentation, tiédeur, tristesse, trouble, vertige