Table des matières

0. Introduction

Le travail réalisé pendant mon stage a consisté en une série d'expériences en apprentissage dont le but était de mettre en évidence l'apport d'informations linguistiques et distributionnelles pour les performances de deux types de classifieurs supervisés : le *Séparateur à Vaste Marge* (SVM) et une variante de la méthode des *k* plus proches voisins (KNN). Cela nous a permis, deuxième objectif de ce travail, de comparer les performances de ces deux classifieurs.

Ces deux méthodes de classification sont utilisées pour la désambiguïsation sémantique automatique et supervisée des unités polysémiques, dans le cadre d'une application de traduction automatique.

L'originalité de notre travail, par rapport à l'état de l'art en désambiguïsation lexicale, réside dans l'utilisation de ressources linguistiques externes. Ces ressources s'avèrent particulièrement précieuses lorsque l'on veut augmenter la qualité de la désambiguïsation en prenant en compte certains phénomènes linguistiques tels que unités multi-mots. Or, c'est ce que nous nous proposons dans le présent travail, sans toutefois vouloir résoudre totalement le problème que pose ce type d'expressions au plan sémantique (dans le cadre de notre stage, s'entend). Nos prétentions s'arrêtent aux unités multi-mots continues et figées, avec un relevé encore manuel de leurs équivalents (composés ou non) dans une autre langue. L'objectif est d'éviter la traduction terme à terme des unités morphémiques qui entrent dans la composition de ces expressions, puisque ces unités, qui sont, par ailleurs, autonomes syntaxiquement et sémantiquement, forment « une unité de sens [...] dont la signification dépasse celle de ses éléments pris isolément » (Riegel, Pellat et Rioul 1994). À long terme, la tâche de détection des traductions de telles expressions devrait être en partie automatisée. Car, devant l'apparition croissante de nouvelles technologies, apportant avec elles leurs lots de mots nouveaux (mots qui sont souvent complexes, justement, dans les domaines techniques), l'idée d'une maintenance exclusivement manuelle de dictionnaires énumérant ces expressions de façon exhaustive est de moins en moins raisonnable.

La conception distributionnelle du sens, qui est à la base des méthodes actuelles de désambiguïsation sémantique, donne un rôle central à un autre type de ressource linguistique : le corpus. L'avenir de la TA est, pour certains, à l'utilisation de corpus à très grande échelle. Les expériences que nous décrivons dans le présent rapport ont été réalisées avec des corpus de taille relativement petite. Nous projetons, pour la suite, de les étendre à l'aide de méthodes d'apprentissage.

La première partie de ce rapport est un très bref état de l'art de la traduction automatique (TA désormais). Un état de l'art complet de ce domaine dépasse le cadre de ce document. La TA est, en effet, un domaine de recherche très ancien, qui recouvre non seulement toutes les applications du Traitement Automatique du Langage Naturel, mais aussi un bon nombre d'autres disciplines tels que l'Intelligence Artificielle, la philosophie du langage, la sémantique, etc. En deuxième partie, nous présentons les ressources linguistiques que nous avons utilisées et la chaîne de prétraitements que nous avons appliquée à notre corpus d'entraînement. La troisième partie est une description des trois classifieurs que nous avons évalués : une variante de la méthode des *k* plus proches voisins (KNN) que nous avons développée, la méthode KNN classique et une variante de la méthode du *Séparateur à Vaste Marge* (SVM), pour laquelle nous avons utilisé une librairie accessible librement en ligne. Les résultats de l'évaluation de ces detroisux classifieurs sont présentés en quatrième partie. Enfin, la cinquième partie est un bref état de l'art de l'alignement automatique de corpus, suivi d'une description de la méthode d'alignement mot à mot que nous avons développée pour les besoins de la méthode de désambiguïsation que nous avons développée.

1. Traduction automatique et désambiguïsation sémantique : état de l'art

L'état de l'art présenté dans cette section a pour sources principales [Léon à paraître] et [Wilks 2009]. Pour une histoire détaillée de la TA en France, voir [Léon 2002].

1.1. La TA depuis les années 90

Depuis les années 90, l'approche empirique et le traitement statistique de grands corpus dominent en TA. Le regain d'intérêt des linguistes pour le lexique, dans les années 80, a conduit à la construction de corpus de documents textuels de plus en plus grands.

Les systèmes de TA les plus récents sont des systèmes hybrides : traduction par règles (dans le cadre de l'approche transfert), traduction statistique (*Statistical Machine Translation,* SMT) et traduction guidée par l'exemple (*Example-Based Machine Translation,* EBMT). Ces deux dernières approches sont englobées dans la traduction basée sur des données (*Corpus-based Machine Translation*, CBMT).

Pour la traduction statistique, [Brown et al. 1990] ont mis au point cinq modèles de plus en plus sophistiqués de traduction mot à mot. Ces modèles sont basés sur trois modules probabilistes avec trois rôles différents :

- un module de traduction (*translation model*), qui doit trouver la traduction la plus probable en LC d'un mot donné en LS,
- un module de distortion (*distortion model*), qui se sert de connaissances sur l'ordre des mots dans les LS et LC pour aligner des positions dans la phrase en LS avec des positions dans la phrase en LC,
- et un module de fertilité (*fertility model*), qui détermine le nombre de mots en LC qui traduisent un mot en LS ou, inversement, le nombre de mots en LS qui traduit un mot en LC, mais aussi, les mots en LS qui ne produisent aucun mot dans la LC (dans la phrase à traduire).

La traduction guidée par l'exemple, ou « traduction par analogie », prend la phrase pour unité de traduction. Elle consiste à rechercher les meilleurs exemples de référence dans une base de données, représentée sous forme d'arbres de dépendance. Une nouvelle traduction est générée par analogie : seules les parties qui changent par rapport à un ensemble de traductions connues sont adaptées, modifiées ou substituées (voir [Somers 1999 et 2003] pour un survol de ces techniques).

1.2. La traduction automatique aujourd'hui

1.2.1. Diverses conceptions du sens

Des tests de compréhension ont révélé que si un locuteur humain lisait un texte en plaçant une feuille de carton avec une fenêtre laissant paraître quelques mots contigus, beaucoup d'ambiguïtés pourvaient être levées. Ainsi est née la conception distributionnelle du lexique, où sont reproduits

non plus les sens, mais les usages des mots. L'idée, en effet, a fait son chemin avec, entre autres, Firth (1957) (« *you shall know a word by the company it keeps* »), Sparck Jone (1964) (qui caractérise la synonymie en termes de contextes linguistiques similaires), etc. Auparavant, Meillet (1921), entre autres linguistes, défendait déjà l'idée d'une conception contextuelle du signifié selon laquelle une unité lexicale n'a pas de sens par elle-même mais seulement dans un contexte : « Le sens d'un mot ne se laisse définir que par une moyenne entre [ses] emplois linguistiques ». La même position quant au sens des mots était adoptée par [Wittgenstein 1953]. Ces hypothèses sont la base, notamment, des approches empiriques, qui acquièrent des connaissances statistiques sur les sens des mots à partir de grands corpus, en caractérisant les sens des mots en termes de cooccurrences lexicales.

Dans la conception traditionnelle de la désambiguïsation sémantique (*word sense disambiguation*), les sens des mots sont considérés comme relevant du domaine du lexique, indépendamment à la fois de leur contexte d'usage et de toute application. Les sens des mots sont donc explicitement énumérés et le rôle de la désambiguïsation sémantique consiste à sélectionner le sens le plus adéquat. Cette conception du sens a conduit à la construction d'énormes ressources linguistiques : dictionnaires électroniques monolingues et multilingues, thésauri, bases de données terminologiques spécialisées dans divers domaines techniques, corpus alignés, etc.

Cette conception du sens n'est pas la plus appropriée dans le cadre du TAL, en particulier en TA : les sens énumérés ne sont pas toujours représentatifs des sens des mots observés dans les données à traduire. D'autres conceptions du sens des mots ont été donc été proposées depuis. [Kilgarriff 1997], par exemple, a suggéré de définir les usages des mots par des « grappes d'usages » (*clusters of word usages*) extraites directement des corpus concernés par les diverses applications du TAL. Cette conception est d'autant plus cohérente que chaque application nécessite différents types de distinctions de sens, en terme de finesse, en particulier. [Schütze 1998] décrit une implémentation de cette conception du sens dans le cadre d'une application d'extraction d'informations, la tâche de désambiguïsation lexicale y est désignée comme « discrimination des sens » (*context group discrimination*). Dans le cadre de la TA, les systèmes basés sur le modèle interlangue recourrent à la désambiguïsation sémantique pour identifier la représentation intermédiaire la plus correcte des concepts exprimés en LS. La désambiguïsation sémantique y est donc vue comme une tâche monolingue, de classification ou de discrimination selon les expériences.

Les systèmes à transfert utilisent également la désambiguïsation sémantique pour construire une représentation adéquate du texte en LS qui doit ensuite être transformée vers la LC. Ces systèmes présentent l'avantage de pouvoir confronter directement le mot en LS avec la liste de ses traductions, traitant ainsi cette dernière comme l'inventaire de ses sens. Le système Systran est une bonne illustration de la manière dont la désambiguïsation sémantique est intégrée dans les systèmes traditionnels de TA : le sens du mot en langue source est tout d'abord identifié sur la base de son contexte (désambiguïsation monolingue), puis la traduction en LC est sélectionnée à l'aide de règles lexicales prenant en compte le contexte linguistique en LC. La désambiguïsation sémantique peut donc difficilement être isolée du processus de traduction.

Parmi les systèmes de TA actuels, c'est l'approche statistique qui domine. Les systèmes sont basés sur un modèle probabiliste qui détecte des liens de traduction entre expressions (*phrases*) en LS et en LC. Par exemple, le fait que l'expression '*we have'* en anglais apparaît souvent comme traduction de la séquence '*nous avons'* en français (Och 2002, Koehn et al. 2003). La traduction d'expressions à expressions, contrairement aux modèles d'IBM de traduction mot à mot, permet de relier un mot en LS avec plusieurs mots en LC ou, inversement, de relier plusieurs mots en LS avec un mot en LC.

1.2.2. Une nouvelle conception du sens en TA

Aujourd'hui, la TA s'est complètement affranchie de tout inventaire de sens prédéfinis. Le problème de la sélection lexicale dans les systèmes de TA est à présent assimilé à un problème de désambiguïsation sémantique, dans lequel les sens des mots sont leurs traductions, relevées dans le corpus d'entraînement du système de traduction. Et la désambiguïsation sémantique est effectuée par un algorithme d'apprentissage qui est entraîné sur les mêmes données que le système de traduction automatique (Och et Ney 2004, Koehn 2004, Chiang 2005, Cabezas et Resnik 2005, Vickrey et al. 2005, Carpuat et Wu 2007).

S'inscrivant dans le cadre de cette nouvelle approche, [Och & Ney 2001] ont proposé une méthode complexe dont la caractéristique principale réside dans l'utilisation des traductions disponibles en plusieurs langues pour désambiguïser les mots polysémiques. [Nomoto 2004], [Schwartz 2008], [Callison-Burch et al. 2008], [Leusch et al. 2009] et [Crego et al. 2009], entre autres, proposent des méthodes de plus en plus sophistiquées dans cette direction. On parle à présent de « traduction statistique multi-source » : l'utilisation de documents en plusieurs langues cibles permet de résoudre les problèmes d'ambiguïté sémantique posés par un mot dans une LS par les hypothèses de traduction de ce mot dans des langues où ces dernières ne sont pas ambiguës afin d'améliorer la sélection lexicale.

[Apidianaki 2009b] souligne les limites de cette approche, qui proviennent essentiellement du statut qui est accordé aux traductions sur le plan sémantique. Les traductions d'un mot sont, en effet, considérées comme autant de ses sens, sans aucune considération de la complexité des liens sémantiques qui relient les traductions entre elles : un sens peut être représenté par plusieurs traductions, et, inversement, une traduction peut participer à la représentation de plusieurs sens. D'autre part, puisque chaque traduction représente un sens distinct, la sélection d'une traduction différente de la référence est automatiquement comptée comme une erreur. Il conviendrait cependant, selon l'auteure, de « considérer la distance entre les sens lexicalisés par des [traductions] différentes », afin d'éviter que des choix, par le module de sélection lexicale, de traductions du même sens que la référence, soient comptabilisés parmi les erreurs. [Apidianaki 2008a] propose donc d'appliquer, préalablement à la sélection lexicale, une méthode de discrimination des sens entraînée sur le corpus du système de traduction automatique. Ainsi, l'inventaire des sens d'un mot ambigu qui sont représentés dans le corpus est construit automatiquement : les traductions du mot sont regroupées sur la base de la similarité distributionnelle des instances du mot qu'elles traduisent. La désambiguïsation sémantique devient alors une tâche de classification qui consiste à trouver, parmi les groupes de traductions découverts, celui qui représente le sens le plus proche d'une nouvelle instance du mot. La traduction du mot est ensuite réalisée par un module de sélection lexicale, dont le rôle est de sélectionner la traduction la plus appropriée parmi le groupe de traductions proposé.

1.3. Conclusion

Pour [Wilks 2009], malgré les progrès qu'a connus le domaine de la linguistique théorique et formelle, les systèmes de TA mis au point jusq'à ce jour sont encore « rudimentaires ». L'avenir de la TA est, selon lui, à la mise en place de projets à long terme, basés sur de véritables formalismes linguistiques et des ressources linguistiques à très grande échelle (dictionnaires et corpus). Les ressources linguistiques à grande échelle sont, pour lui, le futur de la TA (corpus monolingues et multilingues, dictionnaires, étiqueteurs syntaxiques, etc.). Les problèmes d'utilisabilité de ces ressources (formats, interfaces d'utilisation) sont en train d'être résolus, en particulier en Europe (NERC, ELRA, MULTEXT, PAROLE, EAGLES). L'avenir des ressources linguistiques est donc

ouvert et elles ne pourront plus être dissociées de la TA. Les techniques d'apprentissage récemment utilisées dans le cadre de la TA et qui viennent s'ajouter aux techniques statistiques nécessitent, en effet, des ressources annotées de très grande taille. La disponibilité de telles ressources est donc un besoin urgent.

La désambiguïsation sémantique, par ailleurs, a aujourd'hui atteint des performances plus qu'acceptables. La désambiguïsation lexicale est devenue un passage obligé de toutes les applications du TAL. Elle est considérée à l'unanimité comme le futur de la TA. [Wilks 2009] insiste particulièrement sur la nécessité de développer des outils et formalismes sémantiques de haut niveau qui nous permettraient de représenter le contenu sémantique et la structure rhétorique des textes en dépassant le cadre de la phrase. D'autre part, les ressources linguistiques annotées sémantiquement sont, à ce jour, encore très rares. Des efforts restent donc à faire dans cette direction.

Enfin, les fluctuations sur la conception du sens ont aboutit très récemment, dans le cadre de la TA, à une approche de la désambiguïsation sémantique encore toute jeune et qui reste à définir sur des points cruciaux, en particulier, la question de l'évaluation, où la notion d'erreur reste floue. La méthode de désambiguïsation que nous avons développée appartient à cette dernière génération. Elle est fortement inspirée de [Apidianaki 2008a et 2009a].

2. Les espaces sémantiques des mots

2.1. Les vocables étudiés

L'étude présentée ici porte sur une vingtaine de mots (*tableau 1* suivant), dont seize noms et quatre verbes. Le critère de choix de ces vocables est leur caractère fortement polysémique. Deux à quinze usages sont représentés pour chaque mot; et le nombre total de traductions en langue cible pour un mot donné peut varier entre deux à trente-huit.

Les noms
article, barrage, cadre, compte, conclusion, culture, matière, passage, produit, raison, rapport, réserve, société, traitement, vol
Les verbes
lever, monter, porter, saisir

- TABLEAU 1 – *Les 20 vocables polysémiques étudiés* -

Les tableaux de l'annexe 1 (sections *8.1.1. Les noms* et *8.1.2. Les verbes*) représentent les vocables utilisés (16 noms et 4 verbes), leurs différents usages représentés dans la version française du corpus Europarl (première colonne) illustrés par des exemples ou des synonymes (deuxième colonne). Les usages des mots sont tirés du [Larousse 2009]. La troisième colonne contient, pour chaque usage, la liste des traductions en langue anglaise relevés manuellement selon la méthode décrite à la section (*2.2.2*).

Le tableaux de l'annexe 2 (sections *8.2.1. Les noms* et *8.2.2. Les verbes*) contiennent des informations quantitaves sur les vocables étudiés : la taille des sous-corpus de chaque traduction en langue cible (colonne 2) et la taille des sous-corpus d'apprentissage (colonne 3) et de test (colonne 4) du vocable. La taille des sous-corpus s'exprime en nombre de segments alignés. On désigne par « segment aligné » un couple de segments : un segment en LS et le segment qui est sa traduction en LC. Un segment en LS ou en LC peut être composé de plus d'une phrase; et les segments en LS et en LC d'un couple donné ne contiennent pas nécessairement le même nombre de phrases. Un segment peut être composé d'une à trois phrases.

2.2. Construction des corpus d'évaluation et de test

2.2.1. Le corpus utilisé

Les corpus d'apprentissage et de test sont extraits du corpus Europarl [Koehn 2003], un corpus multilingue aligné au niveau des phrases. Ce corpus a été constitué à partir d'actes du Parlement Européen rédigés dans onze langues européennes (allemand, anglais, danois, espagnol, finnois, français, grec, italien, néerlandais, portugais, suédois), entre mars 1996 et septembre 2003. La taille du corpus dans sa version alignée est, pour chaque couple de langues, de l'ordre d'un million de phrases, soit trente millions de mots en moyenne pour chacune des deux langues. Notre étude porte, pour l'instant, sur la traduction du français, langue source (LS désormais) vers l'anglais, langue cible (LC).

2.2.2. La méthode de construction des corpus

a. Constitution d'un lexique bilingue

a.1. Construction manuelle

On a construit un lexique bilingue en relevant, manuellement, pour chaque mot, la liste de ses traductions en LC dans le corpus aligné. (voir le *tableau 8.1.1* de l'*annexe 8.1* : les vocables et leurs équivalents de traduction). Une entrée du lexique contient un vocable ambigu, l'une de ses formes en LS et la traduction de cette forme en LC (simples ou composées) dans le corpus. Par exemple, le vocable *rapport* est représenté par les entrées suivantes dans le lexique :

rapport#rapport#report
rapport#rapport#relation
rapport#par rapport à#regarding
rapport#par rapport à#with regard to

a.2. Construction automatique

La construction du lexique bilingue a, dans un premier temps, été réalisée manuellement et pour un nombre limité de mots, afin d'expérimenter notre méthode de désambiguïsation sémantique. Mais cette méthode ne peut servir pour notre objectif final : construire un outil de désambiguïsation sémantique à très large couverture. La construction manuelle d'un tel lexique est, en effet, un travail trop coûteux en temps.

En outre, la construction automatique du lexique bilingue suppose la disponibilité d'un corpus aligné mot à mot, de façon bidirectionnelle : une version alignée de la LS vers la LC et une version LC-LS. L'ensemble des traductions d'un mot serait alors l'intersection des deux ensembles de mots avec lesquels il serait aligné dans chacune des deux versions.

Un tel corpus n'étant pas disponible, nous avons décidé d'aligner mot à mot les versions française et anglaise (nos langues de travail) du corpus EuroParl. La méthode d'alignement que nous avons développée est décrite à la section (*6.3*) plus bas. L'expérience de notre méthode de désambiguïsation avec un lexique bilingue construite automatiquement n'est pas encore réalisée à ce jour.

b. Les sous-corpus des mots

Pour chaque mot, pour chacune de ses traductions en LC présentes dans le lexique bilingue, on extrait d'EuroParl un sous-corpus de segments alignés. Pour simplifier, on nommera « sous-corpus source » et « sous-corpus cible », respectivement, la version en LS et celle en LC du sous-corpus d'un mot. Par exemple, pour le mot *article*, quatre traductions en anglais ont été relevées : *article*, *rule*, *clause* et *press report*. Pour *press report*, on a donc extrait un nombre donné de segments en français dans lesquels apparaît le mot *article* et qui sont alignés avec un segment en anglais qui traduit *article* par *press report*. On procède de même pour les autres équivalents du mot article (*article*, *rule et clause*), et pour le reste des vocables sélectionnés.

À partir des sous-corpus source et cible, on construit les espaces sémantiques source et cible du mot, qu'on représente par les différents types de contextes décrits à la section 3.3. Dans les espaces

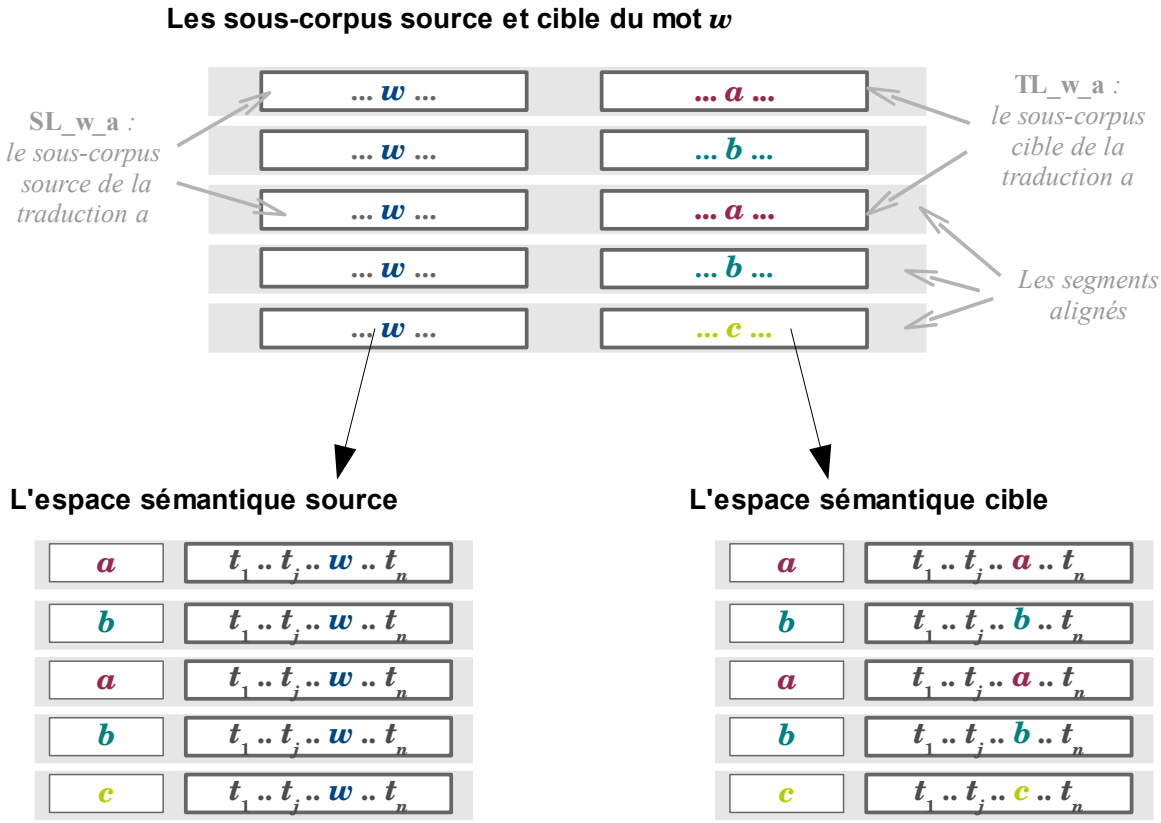
sémantiques, les composantes des vecteurs sont les cooccurrents du mot ambigu.

De chaque sous-corpus représentant une traduction, on réserve 20% des segments pour les tests, et on utilise les 80% restants pour l'entraînement.

b. Les espaces sémantiques des mots

La méthode de désambiguïsation sémantique que nous tentons de mettre en place sera une sous-tâche d'une application de traduction automatique. Par conséquent, l'espace sémantique d'un mot ambigu est construit avec des informations issues des deux langues étudiées. Les classes sont les différentes traductions du mot ambigu que nous avons relevées dans le lexique bilingue; et les composantes des vecteurs sont les mots des segments en LS (dans l'espace sémantique source) ou en LC (dans l'espace sémantique cible).

Le schéma suivant est tiré de Apidianaki 2008b]. Il représente une illustration des sous-corpus source et cible d'un mot polysémique w, à partir desquels on a construit les espaces sémantiques source et cible de ce mot. Dans les espaces sémantiques, chaque traduction $\mathbf{c}_i$ (a, b et c , ici) est elle-même représentée par un sous-corpus, noté **SL_w_c**$_i$.



- **Dessin 1** – *Construction des espaces sémantiques source et cible d'un mot* w, *pour lequel on a relevé trois traductions,* a, b *et* c -

2.3. Représentation des espaces sémantiques

Dans cette section, nous décrivons les différents types d'informations linguistiques et distributionnelles utilisées pour l'entraînement des classifieurs que nous avons testés. Notre expérience consiste à affiner progressivement les informations linguistiques utilisées. L'objectif est d'évaluer l'impact de chaque type d'information linguistique sur les résultats des classifieurs et, par là, le potentiel sémantique des différents contextes relativement au sens du mot.

2.3.1. Les trois types de contextes pris en compte

Les espaces sémantiques des mots peuvent être représentés par trois types de contextes :

- le **contexte thématique** est composé de l'ensemble des cooccurrents du mot,
- le **contexte positionnel** est composé des voisins gauche et droit du mot dans chacun des segments où il occurre,
- et le **contexte syntaxique** est composé des cooccurrents syntaxiques du mot : les cooccurrents avec lesquels il entretient une relation syntaxique directe (sujet, complément, complément du nom, etc.).

Ces trois types de contextes sont, pour la suite, dénotés par la variable ***ctxt_type***, dont les modalités sont ***ctxt*** (le contexte thématique), ***pos*** (le contexte positionnel) et ***synt*** (le contexte syntaxique). Dans la suite, on désignera par « contexte positionnel » ou « contexte syntaxique » les vecteurs contenant à la fois :

- les composantes du contexte positionnel ou syntaxique définis précédemment
- et les composantes du contexte thématique qui ne sont pas inclues dans les composantes précédentes.

Les deux types de composantes (voisins ou cooccurrents syntaxiques d'une part, et cooccurrents thématiques d'autre part) sont distingués par le poids qui leur est assigné (voir *section 3.3* suivante) et, éventuellement, par leur forme (mot ou lemme).

composantes	exemples
le contexte thématique	
les cooccurrents	*Je, voudrais, intervenir, en, application, de, l, article, 133, du, règlement*
le contexte positionnel	
voisins gauche et droit + autres cooccurrents	*L, 133* *Je, voudrais, intervenir, en, application, de, article, du, règlement*
le contexte syntaxique	
Cooccurrents syntaxiques + autres cooccurrents	*133, règlement, de, l* *Je, voudrais, intervenir, en, application, article, du*

- TABLEAU 2 – *Les trois types de contextes des mots polysémiques* -

2.3.2. Pondération des composantes

Pour la pondération des différents types de contextes représentés dans un vecteur de l'espace sémantique, on a définit comme poids de base (*poids_base*) 1, et comme ratio (*ratio*) 2 (différents ratios ont été testés, c'est celui-ci qui nous a permis d'obtenir les meilleures performances). Les **poids absolus** sont attribués aux composantes selon la règle suivante :

- si aucun contexte n'est marqué (contexte thématique seul) :
  $p(cooccurrents) = poids\_base$
  $p(mot) = ratio * poids\_base$

- si le contexte positionnel est marqué :
  $p(voisins) = 2 * poids\_base$
  $p(mot) = ratio * p(voisins)$

- si le contexte syntaxique est marqué :
  $p(cooccurrents\_syntaxiques) = ratio * poids\_base$
  $p(mot) = ratio * p(cooccurrents\_syntaxiques)$

Pour la suite, on notera $poids\_absolu_j$ le poids absolu d'un trait t_j et on appellera **traits forts** les traits de poids supérieur au poids de base.

2.3.3. Les informations linguistiques utilisées

Le mode de représentation des trois types de contextes présentés précédemment peut varier selon les modalités des quatre variables illustrées dans le tableau XX suivant.

variables	modalités	exemples
forme des traits *forme*	mots lemmes	*cet, article, ne, va, pas, non, plus, dans, l, intérêt, du, consommateur* *ce, article, ne, aller, pas, non, plus, dans, le, intérêt, du, consommateur*
type des traits *infos*	*forme* *forme#étiq_gramm*	*cet, article, ne, va, pas, non, plus, dans, l, intérêt, du, consommateur* *cet*#pro:dem, *article*#nom, *ne*#adv, *va*#ver:infi, *pas*#adv, ...
mots composés *comp*	non marqués marqués	*Pénuries, d, essence, barrages, routiers, protestations, ...* *Pénuries, d, essence, barrages routiers, protestations, ...*
filtrage local (N, V, Adj) *filtrage_local*	non oui	*cet, article, ne, va, pas, non, plus, dans, l, intérêt, du, consommateur* *article, intérêt, consommateur*

- TABLEAU 3 – *Les différents types d'informations linguistiques utilisés pour la représentation des trois types de contextes* -

Le marquage des mots composés a été réalisé à l'aide du logiciel Unitex, un outil de traitement de corpus développé, entre autres, au laboratoire d'informatique de l'Institut Gaspard Monge (université de Paris-Est Marne-la-Vallée). Pour la lemmatisation et l'étiquetage morpho-syntaxique, nous avons utilisé l'étiqueteur TreeTagger, développé à l'université de Stuttgart par Helmut Schmid (Schmid 1995) et qui permet d'étiqueter des textes dans plusieurs langues.

Un mode de représentation de l'espace sémantique d'un mot est ensuite désigné par une combinaison des différentes variables définies.

- **Exemples :**

Un contexte d'usage du mot :
Je voudrais intervenir en application de l'article 133 du règlement.

COMBINAISON 1 : illustration du contexte thématique

Définition :
ctxt_type = *ctxt*
infos = *forme#etiquette_grammaticale*
comp = marqués
contexte thématique :
forme = *mots*
filtrage_local = *oui*

Explication :
L'espace sémantique du mot est représenté par son contexte thématique.
Les composantes des vecteurs sont de type *forme*, les mots composés sont marqués.
Le contexte thématique est composé de mots (formes fléchies) qui sont filtrés (mots, verbes et adjectifs uniquement).

Le mot ambigu : *article*
Les composantes du contexte thématique : *je, voudrais, intervenir, en, application, de, l, article, 133, du, règlement*

Les composantes du vecteur dans l'espace sémantique :
voudrais#ver-cond:1, *intervenir#ver-infi*:1, *application#nom*:1, ***article#nom*:2**, *règlement#nom*:1

COMBINAISON 2 : illustration du contexte syntaxique

Définition :
ctxt_type = *synt*
infos = *forme#etiquette_grammaticale*
comp = marqués
contexte syntaxique :
forme = *mots*
filtrage_local = *non*
contexte thématique :
forme = *lemmes*
filtrage_local = *oui*

Explication :
L'espace sémantique du mot est représenté par son contexte syntaxique.
Les composantes des vecteurs sont de type *forme#étiquette_grammaticale*, les mots composés sont marqués. Les cooccurrents syntaxiques sont des formes fléchies et ne sont pas filtrés. Le reste des cooccurrents (le contexte thématique) est composé de lemmes qui sont filtrés (noms, verbes et adjectifs uniquement).

Le mot ambigu : *article*
Les composantes du contexte syntaxique : *de, l, 133, règlement*
Les composantes du contexte thématique : *je, vouloir, intervenir, en, application, du*

Les composantes du vecteur dans l'espace sémantique :
vouloir#ver-cond:1, *intervenir#ver-infi*:1, *application#nom*:1, ***de#prepded*:2**, ***l#detart*:2**, ***article#nom*:4**, ***133#card*:2**, ***règlement#nom*:2**

COMBINAISON 3 : illustration du contexte positionnel

Définition :
ctxt_type = *pos*
infos = *forme*
comp = marqués
contexte positionnel :
forme = *mots*
filtrage_local = *non*
contexte thématique :
forme = *lemmes*
filtrage_local = *non*

Explication :
L'espace sémantique du mot est représenté par son contexte positionnel.
Les composantes des vecteurs sont des formes simpels (*forme*), les mots composés sont marqués. Les voisins du mot sont des formes fléchies et ne sont pas filtrés. Le reste des cooccurrents (le contexte positionnel) est composé de lemmes qui ne sont pas filtrés.

Le mot ambigu : *article*
Les composantes du contexte positionnel : *l, 133*
Les composantes du contexte thématique : *je, vouloir, intervenir, en, application, de, du, règlement*

Les composantes du vecteur dans l'espace sémantique :
je#proper:1, *vouloir#ver-cond*:1, *intervenir#ver-infi*:1, *en#prep*:1, *application#nom*:1, *de#prepded*:1, ***l#detart*:2**, ***article#nom*:4**, ***133#card*:2**, *du#prep*:1, *règlement#nom*:2

3. Présentation des classifieurs

Les différents types de contextes présentés à la section précédente (*3*) ont été évalués à l'aide de deux méthodes de classification supervisée :

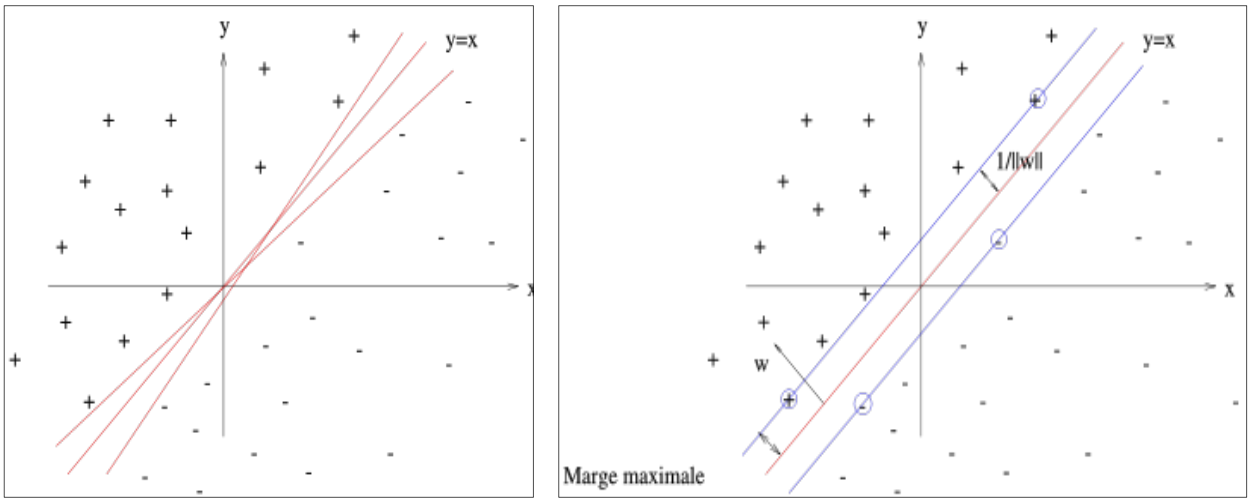
- le **Séparateur à Vaste Marge** (SVM) : pour lequel nous avons utilisé la librairie LibSVM disponible librement en ligne

- et **la méthode des *k* plus proches voisins** (KNN) : dont nous avons développée une variante en nous inspirant de la méthode de désambiguïsation sémantique pour la traduction automatique présentée dans [Apidianaki 2009a].

3.1. SVM : le Séparateur à Vaste Marge

Pour tester la méthode SVM, nous avons utilisé la librairie *LIBSVM*, librement disponible sur l'Internet. Les SVM sont une classe d'algorithmes appartenant, avec la méthode des *k* plus proches voisins, à la catégorie des méthodes d'apprentissage à partir de cas (ou « supervisé)).

La méthode SVM est une méthode de classification linéaire qui repose sur l'hypothèse que, étant donné un espace approprié, il existe un classificateur linéaire (appelé hyperplan) permettant de distinguer les deux classes de l'espace (+/-). Le but de cette méthode est d'apprendre, à partir d'un ensemble d'exemples d'apprentissage (apprentissage supervisé), une fonction qui prédit les classes pour de nouveaux objets. Plus concrètement, il s'agit de trouver l'**hyperplan optimal**, qui sépare les données et maximise la distance entre les deux classes.



- **Dessin 2** – *L'espace d'apprentissage d'une méthode SVM : en (a), trois hyperplans valides, en (b), l'hyperplan optimal et les vecteurs supports (entourés) qui ont permis de déterminer la marge maximale entre les deux classes.* -

L'hyperplan optimal est celui, parmi tous les hyperplans valides, qui réalise la **marge** maximale entre les points des deux classes (celui qui passe au milieu des points des deux classes). C'est la

raison pour laquelle on parle de séparateur *à vaste marge*. Les points les plus proches de la frontière entre les deux classes et qui sont utilisés pour la détermination de l'hyperplan optimal sont appelés **vecteurs supports** (entourés dans le dessin 2.b).

L'hyperplan optimal est celui qui permettra au mieux de classer les nouveaux exemples. La classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal.

Face à un cas non linéairement séparable (c'est-à-dire la plupart des problèmes réels), les méthodes SVM recourrent à une fonction noyau pour effectuer une transformation non linéaire des données. Le résultat de cette transformation, appelé ***espace de re-description***, est un espace de dimension plus grande.

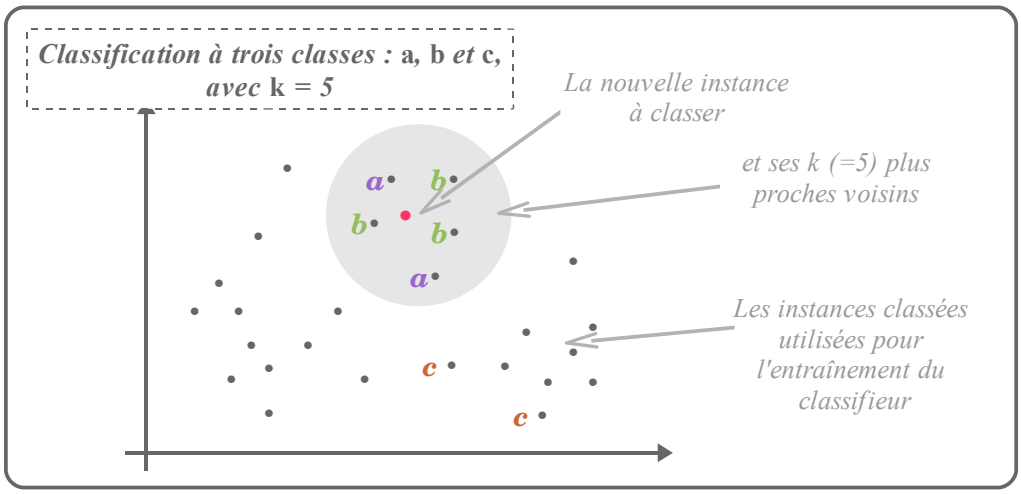
3.2. KNN : les *k* plus proches voisins

La méthode des *k* plus proches voisins est une méthode d'apprentissage par analogie (ou apprentissage à partir de cas) dont le principe de base consiste à rechercher, parmi un ensemble d'exemples préalablement classés, les *k* (avec *k*=1 ou plus) exemples les plus similaires au nouvel objet (mot, document, requête) que l'on veut classer, puis à inférer à partir des classes de ces *k* exemples la classe de cet objet. Cette méthode n'inclut donc pas une phase réelle d'apprentissage, d'où sa simplicité. Trois paramètres doivent donc être déterminés pour la construction d'un modèle d'inférence basé sur cette méthode :

- **la valeur de *k*** : le nombre d'exemples déjà classés à prendre en compte pour calculer la classe de la nouvelle instance,

- **une fonction de distance** : pour rechercher les *k* exemples les plus similaires à la nouvelle instance, le choix de cette métrique étant déterminé par le mode de représentation des données utilisé (vecteurs, graphes, etc.),

- **une fonction de prédiction** : pour décider de la classe à attribuer à cette nouvelle instance, étant données les classes des *k* exemples les plus similaires trouvés.



- **Dessin 4** – *Projection de l'espace sémantique d'un mot, avec une nouvelle instance à classer et ses (k=) 5 plus proches voisins* -

3.2.1. La méthode classique des *KNN*

Pour mesurer l'avantage de la classification en deux étapes (regroupement des traductions puis classification), nous avons comparé les résultats de la méthode [Apidianaki 2009] (*section 3.2.2*) avec les résultats d'une variante de la méthode classique des plus proches voisins, que nous décrivons ici. Dans la définition de cette variante des *KNN*, nous avons réutilisé deux formules de la méthode [Apidianaki 2009] :

1. la formule de calcul des poids des traits relativement aux classes
2. et l'utilisation de la moyenne des scores des classes comme seuil minimal pour l'assignation d'une classe à une nouvelle instance de mot à classer.

a. Recherche des *k* plus proches voisins

ENTRÉE : l'ensemble d'entraînement et une nouvelle instance du mot à classer.

Pour la recherche des k plus proches voisins, nous avons utilisé le **coefficient de Jacquard**. Étant données deux instances du mot, a (une instance de l'ensemble d'entraînement) et b (la nouvelle instance à classer), représentées par les vecteurs u et v respectivement, la similarité entre a et b selon le coefficient de Jacquard est définie comme suit :

$$\text{sim}(a, b) = |\text{intersection}(u,v)| \ / \ |\text{union}(u,v)|$$

avec :

$\text{intersection}(u,v)$ le vecteur des composantes qui sont à la fois dans u et dans v, et $\text{union}(u,v)$ le vecteur des composantes qui sont dans u ou dans v.

SORTIE : les k instances de l'ensemble d'entraînement qui sont les plus proches de la nouvelle instance à classer.

b. Calcul des scores des classes étant données les classes des *k* plus proches voisins

ENTRÉE : les k instances de l'ensemble d'entraînement qui sont les plus proches de la nouvelle instance à classer.

À chaque classe c_i représentée dans les k plus proches voisins trouvés est associé un vecteur $v(c_i)$ qui contient l'union des composantes des vecteurs des plus proches voisins qui lui sont associés.

Puis le score de la classe c_i est égal à la somme des poids relatifs des composantes du vecteur $v(c_i)$. (le calcul des poids relatifs des traits par rapport aux classes est décrit à la *section 3.2.2.B.a* plus bas)

$$\text{score}(c_i) = \sum_{j=1..|v(ci)|} w_{ij}$$

SORTIE : les k classes des k plus proches voisins et leur score.

c. Recherche de la(es) classe(s) de la nouvelle instant étant données les scores des classes représentées parmi les plus proches voisins

ENTRÉE : les k classes des k plus proches voisins et leur score.

La(es) classe(s) assignée(s) à b, la nouvelle instance à classer, sont celles qui ont obtenu un score supérieur ou égal à la moyenne des scores des k classes des k plus proches voisins.

$$\text{moy\_scores} = \Sigma_k \text{ score}(c_i) / k$$
$$\text{Classes}(b) = \{ c_i \mid score(c_i) >= \text{moy\_scores} \}$$

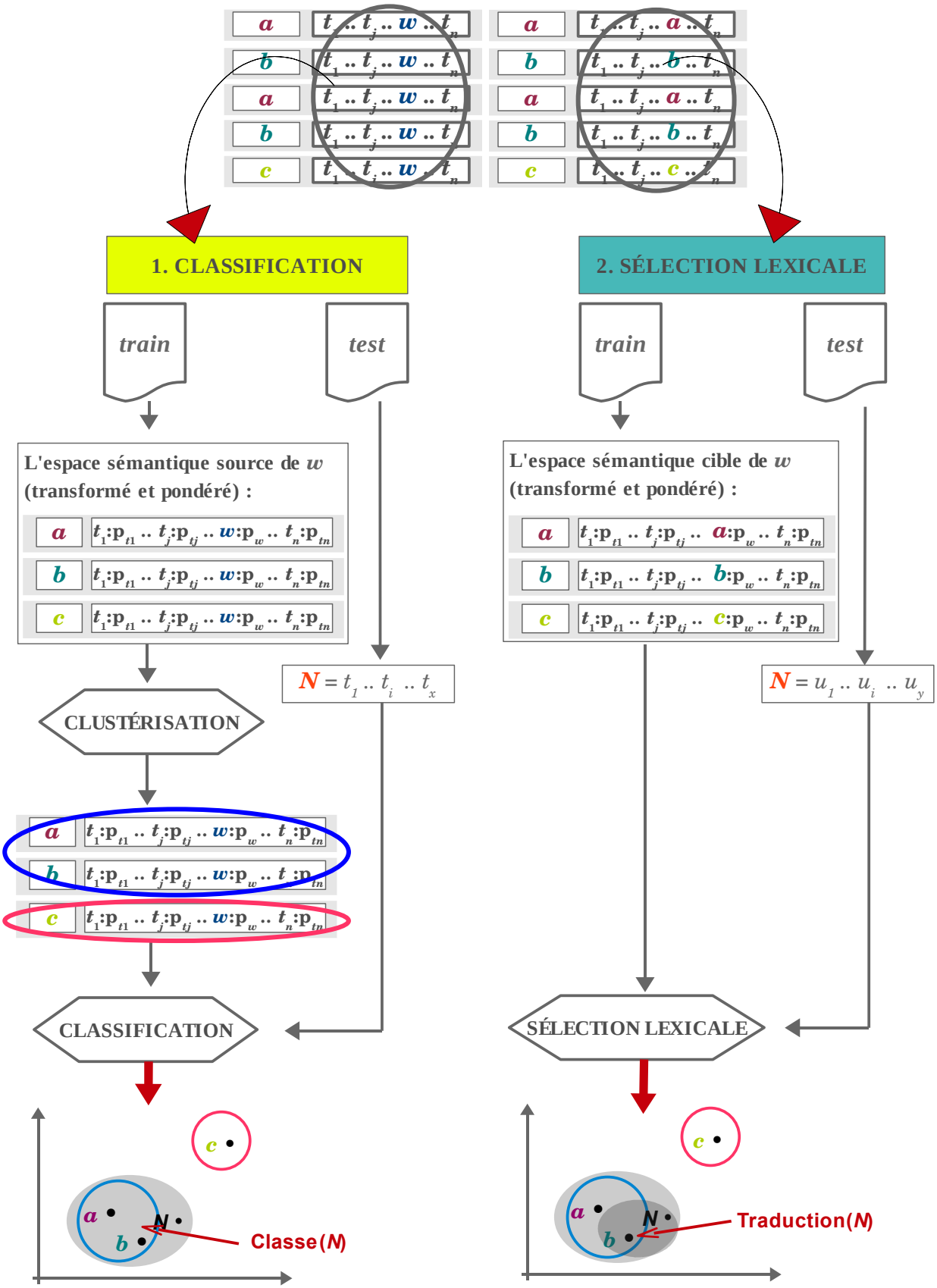
3.2.2. La méthode [Apidianaki 2009]

La méthode que nous avons développée et évaluée est fortement inspirée de la méthode décrite dans [Apidianaki 2008a, 2009a]. Cette méthode procède en deux étapes.

La première étape sert à discriminer les différents usages d'un mot, en regroupant ses traductions les plus similaires sémantiquement à l'aide du méthode de *clustering*. À la fin de cette première étape, la classification d'une nouvelle instance du mot consiste à lui assigner le groupe de traductions représentant son usage. Là réside l'avantage principal de cette approche. En effet, comme le souligne son auteure, la correspondance entre les sens d'un mot en LS et ses traductions en LC n'est pas biunivoque : plusieurs traductions peuvent être utilisées pour traduire un même sens du mot en LS. De fait, comme le remarque [Martinet 1970], « à chaque langue correspond une organisation particulière des données de l'expérience. Apprendre une autre langue, ce n'est pas mettre de nouvelles étiquettes sur des objets connus, mais s'habituer à analyser autrement ce qui fait l'objet de la communication ». L'idée de l'originalité des langues les unes par rapport aux autres reposant sur la spécificité du découpage de leur espace sémantique est également défendue par la sémantique différentielle de Rastier. Ce dernier admet, en effet, « l'irréductibilité des langues les unes aux autres et la spécificité de leurs sémantiques, dont témoignent en premier lieu leurs lexiques » (Rastier 2001).

La seconde étape permet de déterminer, parmi les traductions représentant l'usage choisi pour la nouvelle instance, celle qui est la plus probable, en fonction de son contexte en langue cible. Il se peut que la traduction choisie ne soit pas la traduction de référence. Le tout est de savoir comment décider s'il s'agit d'une erreur de classification du système : sur quels critères peut-on se baser, lors de l'évaluation de la méthode, pour définir ce qui doit être compté comme une erreur. Il s'agit là d'un problème récurrent dans le domaine de l'évaluation des techniques et applications du TAL, en général, et qui reste sans réponse véritable à ce jour.

A. Schéma global de la méthode développée



- **Dessin 5** – *Schéma global de la méthode développée* -

B. Description de la méthode développée

a. Transformation de l'espace sémantique source du mot

Entrée : la part du sous-corpus source du mot destinée à l'apprentissage (les vecteurs $v_t(c)$)

Dans un premier temps, chaque classe (traduction du mot) c est représentée par un vecteur de traits, noté $v(c)$, qui est l'union des vecteurs d'instances qui lui sont associés dans l'espace sémantique.
Puis, on calcule les poids des traits des vecteurs source $v(c)$, relativement à chacune des classes auxquels ils sont associés dans l'espace sémantique. Un trait est associé à une classe c_i dans l'espace sémantique s'il apparaît dans le vecteur $v(c_i)$, qui la représente. Les formules utilisées permettent de mesurer la pertinence des traits dans la représentation des classes. Dans les formules suivantes, t est un trait, c est une classe, *nb-classes* est le nombre de classes représentées dans l'espace sémantique (le nombre de traductions du mot relevées) et *nb-traits* est le nombre total de traits de l'espace sémantique. $C(c_i, t_j)$ est le nombre de vecteurs d'instances associés à la classe c_i et qui contiennent le trait t_j dans l'espace sémantique.

* ***calcul du poids global d'un trait :***

$$gw_j = 1 - ((\Sigma_{i=1..\,nb\text{-}classes}\ p_{ij} \cdot \log(p_{ij})) / nrels) \qquad (1)$$

avec :

$$p_{ij} = C(c_i, t_j) / |v(c_i)| \qquad (1.1)$$

et *nrels* = le nombre total de classes auxquelles t_j est associé (1.2)

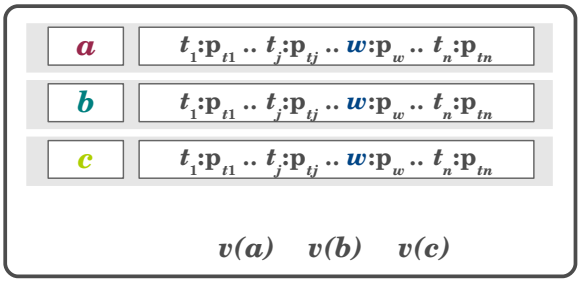
* ***calcul du poids local d'un trait, relativement à une classe c_i :***

$$lw_{ij} = \log C(c_i, t_j) \qquad (2)$$

* ***calcul du poids total d'un trait, relativement à une classe c_i :***

$$w_{ij} = poids\_absolu_j * gw_j * lw_{ij} \qquad (3)$$

Sortie : l'espace sémantique source du mot, avec les instances du mot regroupées par traduction et les traits pondérés



- **Dessin 6** – *L'espace sémantique source du mot* w, *transformé et pondéré* -

b. La matrice de similarité entre les traductions du mot

Entrée : l'espace sémantique source du mot

La matrice de similarité contient les mesures de similarité calculées entre toutes les paires possibles de classes. La formule utilisée est le **coefficient de Jacquard pondéré** (*WJ*) présenté dans [Grefenstette 1994]. Pour deux classes c_m et c_n, la formule est la suivante :

$$WJ(c_m, c_n) = \Sigma_{j=1..nb\text{-}traits} \min(w_{mj}, w_{nj}) / \Sigma_{j=1..nb\text{-}traits} \max(w_{mj}, w_{nj}) \quad (4)$$

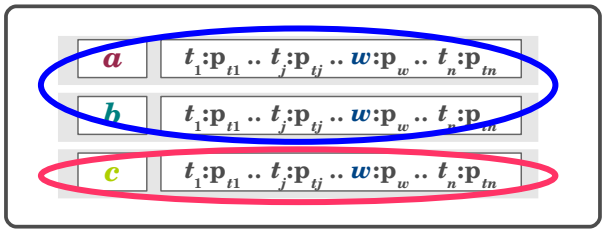
Sortie : une matrice associant à chaque paire de traductions du mot leur similarité sur la base des contextes en LS, $v(c)$, auxquels elles sont associées.

c. Regroupement des traductions du mot

Entrée : la matrice de similarité entre les traductions du mot

Les traductions du mot sont regroupées en cliques qui sont sensées représenter ses différents emplois dans la LS. Chaque groupe de traductions est, ensuite, représenté par un vecteur $v(G)$ qui est l'intersection des vecteurs $v(c)$ de ses membres. Les groupes de traductions ainsi constitués peuvent se chevaucher, puisque des acceptions différentes du mot dans la LS peuvent être traduites par une même unité dans la LC.

Sortie : les groupes G de traductions distributionnellement similaires et les vecteurs $v(G)$ des traits qui leurs sont associés



- **Dessin 7** – *L'espace sémantique source clustérisé du mot* **w** -

d. Classification

Entrée : les groupes G de traductions calculés précédemment et les vecteurs $v(G)$ des traits qui leurs sont associés

On se trouve, ici, dans la dernière partie de la première étape du système. Il s'agit, à présent, d'assigner à une nouvelle instance N du mot, le groupe de traductions le plus correct, c'est-à-dire celui dont la similarité avec le contexte de N est la plus forte selon la formule (5) (plus bas). Pour cela, le contexte en langue source de la nouvelle instance est, éventuellement, pré-traité, de la même manière que le sous-corpus d'apprentissage du mot, et représenté sous la forme d'un vecteur V composé de traits pondérés.

**** Calcul de la similarité de N avec chaque groupe G_k de traductions :***

- soit $intersection(V, G_k)$ l'intersection des traits de V d'une part, et des traits de $v(G_k)$ d'autre part,
- $sim(N, G_k)$ est la similarité entre N et G_k selon la formule suivante (avec $|G_k|$ le nombre de traductions que contient le groupe G_k) :

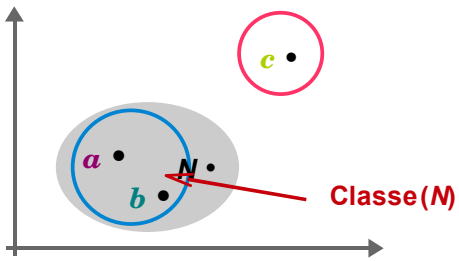
$$sim(N, G_k) = \Sigma_{j=1..|intersection(V,Gk)|,\, i=1..|Gk|}\, w_{ij} \,/\, (|G_k| * |intersection(V, G_k)|) \qquad (5)$$

**** Choix du groupe de traductions le plus correct :***

Le groupe de traductions le plus probable pour la nouvelle N est celui qui maximise $sim(N, G_k)$:

$$\begin{aligned} Classe(N) &= \arg\max sim(N, G_k) \\ &= \{ G_k \mid \forall\, d \neq k, sim(N, G_d) < sim(N, G_k) \} \end{aligned}$$

Sortie : la nouvelle instance du mot et le groupe de traductions G qui lui a été assigné

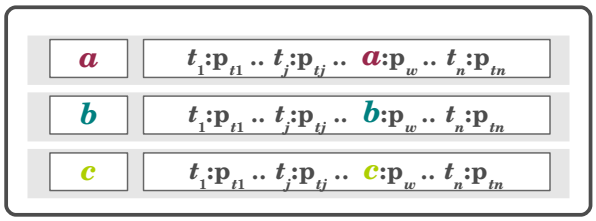


- Dessin 8 – *Classification d'une nouvelle instance N du mot w -*

e. Transformation de l'espace sémantique cible du mot

Entrée : la part du sous-corpus cible du mot destinée à l'apprentissage (les vecteurs $v_u(c)$)

De la même manière que pour les traits des vecteurs source, on calule les poids relatifs des traits des vecteurs cible $v_u(c)$, puis, chaque classe est représentée par un vecteur $v(c)$ qui est l'union des vecteurs $v_u(c)$ qui lui sont associés dans l'espace sémantique du mot.
Sortie : l'espace sémantique cible du mot



- Dessin 9 – *L'espace sémantique cible du mot* w, *transformé et pondéré -*

f. Sélection lexicale

Entrée : la nouvelle instance du mot et le groupe de traductions G qui lui a été assigné (étape 4)

Le contexte en langue cible de la nouvelle instance N est pré-traité de la même manière que le sous-corpus cible d'apprentissage du mot, et représenté sous la forme d'un vecteur V composé de traits pondérés. Puis, on calcule le score de similarité entre chaque traduction c_i et la nouvelle instance N. La traduction du mot est celle qui obtient le score le plus élevé.

* ***Pour chaque traduction c_i du groupe :***

- soit $intersection(V, c_i)$ l'intersection des traits de V et $v(c_i)$,
- soit $weight(N, c_i)$, la somme des poids relatifs des traits de $intersection(V, c_i)$,
- $sim(N, c_i)$ est la similarité entre N et c_i selon la formule suivante :

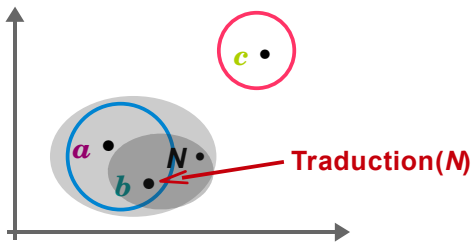
$$sim(N, c_i) = weight(N, c_i) / \Sigma_{n=1..|G|}\ weight(N, c_n) \qquad (5)$$

* ***Détermination de la traduction la plus probable du mot :***

La traduction la plus probable pour la nouvelle instance est celle qui maximise $sim(N, c_i)$:

$$Traduction(N) = \arg\max sim(N, c_i)$$
$$= \{ c_i \mid \forall\, n \neq i, sim(N, c_n) < sim(N, c_i) \}$$

Sortie : la nouvelle instance du mot et sa traduction c_i



- **Dessin 10** – *Application de la méthode de sélection lexicale pour trouver la traduction de la nouvelle instance* N *du mot* w -

3.2.3. Conclusion

Telle quelle, cette méthode, qui traduit un mot sur la base de ses contextes en lange source et en langue cible, est une méthode expérimentale. Pour passer au stade supérieur et créer un véritable système de désambiguïsation par les traductions, qui soit capable de traduire un texte complet en langue source, il conviendrait d'élargir le lexique au moins à tous les mots du texte concerné. Ceci est notre prochain objectif. Pour cela, il nous avions besoin d'un corpus aligné mot à mot, la construction manuelle d'un lexique bilingue à très large couverture étant trop coûteuse. Nous avons donc décidé d'aligner mot à mot les versions française et anglaise du corpus Europarl. Puis nous procéderons à la construction automatique du lexique et la méthode de sélection lexicale sera ajustée aux informations, plus restreintes, dont elle disposera.

4. Évaluation des performances des classifieurs

4.1. Les métriques d'évaluation

Les métriques d'évaluation utilisées sont les suivantes :

rappel = nombre de prédictions correctes / nombre d'instances de test

précision = nombre de prédictions correctes / nombre de prédictions

*f-score = (2 * (rappel * précision)) / (rappel + précision)*

Une prédiction est correcte si le groupe de traductions choisi par l'algorithme de classification contient la traduction de référence. C'est la notion de *f-score enrichi* définie dans [Apidianaki 2009].

4.2. Les paramètres d'évaluation

Pour l'évaluation de nos deux classifieurs, nous avons testé toutes les combinaisons possibles des variables définies aux sections (3.2) et (3.4) concernant les données d'entraînement.

Les paramètres d'évaluation de nos deux classifieurs sont les variables définies aux section (3.2) et (3.4), dont nous avons testé toutes les combinaisons possibles. Nous y avons ajouté un paramètre concernant le type des traits pris en compte pour calculer la similarité entre classes (représenté par la variable ***sim***). Les classes sémantiquement similaires sont regroupées sur la base des traits qui leur sont associés dans l'espace sémantique. Pour calculer la similarité entre deux classes, on utilise :

- soit la totalité des traits contenus dans les deux vecteurs $v(c)$ qui les représentent (donc tous les types de contextes qui y sont représentés),
- soit uniquement les traits forts (ceux dont le poids absolu est supérieur au poids de base), qui correspondent soit aux voisins soit aux cooccurrents syntaxiques.

Dans les deux cas, tous les traits de l'espace sémantique sont pris en compte lors du calcul des poids relatifs des traits.

Le tableau suivant résume les paramètres d'évaluation et les codes correspondant à leurs modalités utilisés dans la suite de cette évaluation. Le filtrage local et la lemmatisation (variables *filtrage_local* et *forme*) sont représentés dans les modalités des paramètres *mode_ctxt*, *mode_pos* et *mode_synt*.

PARAMÈTRE	SIGNIFICATION	MODALITÉS	
		CODE	VALEUR
ctxt_type	Le type de contexte du mot ambigu utilisé pour représenter son espace sémantique.	1 2 3	ctxt pos synt
infos	Le type des composantes de l'espace sémantique.	1 2	forme forme#étiquette_grammaticale
comp	Le marquage ou non des mots composés.	1 2	non oui
mode_ctxt ***mode_pos*** ***mode_synt***	Le mode de représentation des composantes des contextes thématique, positionnel et syntaxique.	1 2 3 4	mots mots filtrés lemmes lemmes filtrés
sim	Les composantes prises en compte pour calculer la similarité entre traductions.	1 2	toutes les composantes uniquement les composantes du contexte thématique ou positionnel (selon *ctxt_type*)

- TABLEAU 5 – *Les paramètres d'évaluation des classifieurs* -

4.3. Évaluation de la méthode développée

4.3.1. Évaluation quantitative

a. Les meilleures performances atteintes

a.1. Par le module de désambiguïsation

Le meilleur score atteint par notre méthode s'élève à **90,5%**. Nous l'avons obtenu avec les combinaisons de contextes suivantes (les valeurs du tableau correspondent aux valeurs des paramètres présentés au *tableau 2* plus haut) :

COMBINAISONS	C1	C2	C3	C4	C5	C6
ctxt_type	3	3	3	3	2	2
infos	2	2	2	2	2	2
comp	1	1	1	1	1	1
mode_pos	0	0	0	0	1	2
mode_synt	1	2	3	4	0	0
mode_ctxt	3	3	3	3	3	3
sim	2	2	2	2	2	2
SCORES	90,4	90,3	**90,5**	90,4	90,4	90,3

- TABLEAU 6 – *Les combinaisons optimales pour la désambiguïsation sémantique* -

On remarque que les meilleures performances sont réalisées quand :

- les traits sont du type *forme#étiquette_grammaticale*;
- les mots composés sont marqués;
- le contexte syntaxique du premier ordre ou le contexte positionnel sont composés de mots ou de lemmes, avec ou sans filtrage local;
- le contexte thématique est composé de lemmes et non filtré localement.
- et seuls les traits forts, cooccurrents du premier ordre ou voisins gauche et droit, sont pris en compte pour le regroupement des classes sémantiquement similaires;

On peut en conclure que la méthode de « filtrage par les groupes syntaxiques » proposée dans [Besançon et Rajman 2002] peut être efficace dans notre cas. Dans tous les cas, la pertinence de l'idée d'un traitement différentiel des contextes du mot est prouvée.

a.2. Par le module de sélection lexicale

Le module de sélection lexicale a effectué 87,4% d'assignations correctes.

b. Évaluation des paramètres

b.1. Les tendances globales des paramètres

Le *tableau 7* suivant représente la force d'association entre les variables deux à deux. Il permet de dégager la tendance globale de chaque variable et la variation de son comportement par rapport à toutes les autres. Une cellule du tableau contient le meilleur score atteint par l'ensemble des combinaisons de contextes réalisant la modalité en entrée de ligne et la modalité en entrée de colonne des deux variables concernées.

D'un point de vue général, en restant dans le cadre de notre expérience dans certains cas, ce tableau nous conduit à plusieurs conclusions.

- *sim*. En mesurant les similarités sémantiques entre les traductions d'un mot sur la base de leur contexte syntaxique ou de leur contexte positionnel, on obtient des résultats meilleurs qu'avec le contexte thématique. On en déduit que ces deux types de contextes ont un potentiel sémantique plus important que le contexte thématique par rapport au sens du mot.

- *infos*. L'insertion d'informations morpho-syntaxiques dans les traits de l'espace sémantique, et donc la prise en compte de ces informations dans la mesure des similarités entre les traductions d'un mot, apporte plus de précision.

Ces deux premiers points appuient la thèse de François Rastier (Rastier 2009) selon laquelle « la morphosyntaxe définit [...] des zones structurelles de localité, telles que les unités lexicales qui sont placées sous un même noeud syntaxique entretiennent des interactions sémantiques privilégiées ». Ainsi, les unités lexicales « ont une définition relationnelle [...] dans l'ordre syntagmatique », et il existe « une sémantique propre à la combinaison des morphèmes ». La sémantique interprétative de Rastier stipule, en effet, que « les phénomènes morphosyntaxiques [...], au cours de la compréhension, conditionnent les opérations interprétatives ».

- *comp.* En décomposant les expressions multi-termes en plusieurs unités autonomes, on introduit des informations erronées dans l'espace sémantique des mots. De fait, les sous-unités d'un mot composé ne sont que « des parties d'un seul mot » et « la signification du composé dans son ensemble est [...] divergente de l'étymologie précise de ses éléments » (Sapir 2001).

- *forme.* La lemmatisation des mots du contexte permet de découvrir des similarités sémantiques plus fortes entre mots, elle empêche la discrimination de sens trop fins qui peuvent n'être pas pertinentes. On remarquera, ici, la variation des effets de la lemmatisation en corrélation avec la taille des sous-corpus des mots. Lorsque nous avons réalisé la même expérience avec des corpus de taille plus restreinte (5 à 30 exemples par traduction), nous avons, en effet, conclu que la lemmatisation faisait perdre de l'information. On en arrive à une conclusion générale concernant ce paramètre : plus la taille des corpus diminue, plus la lemmatisation est à proscrire. Cela peut s'expliquer par le fait qu'avec un corpus de petite taille, autrement dit, une quantité d'informations plus restreinte, aucune information linguistique disponible ne doit être ignorée. Autrement dit, la description du contexte doit être plus précise. En l'occurrence, les informations morphologiques sur les mots doivent être conservées. Contrairement à cela, un grand corpus permet, ou plutôt exige, des généralisations sur certains types d'informations. Plus concrètement, les unités lexicales regroupées en un même lemme ont, forcément, une fréquence de cooccurrence avec le mot étudié qui est plus élevée, dans un corpus de grande taille. Et ces différentes unités ne cooccurrent jamais dans un même contexte syntaxique ou positionnel. Or, en désambiguïsation sémantique, la discrimination des usages d'un mot polysémique est réalisée sur la base des fréquences de cooccurrence entre ses cooccurrents. Donc, les unités lexicales qui ne cooccurrent jamais dans ses contextes sont considérées comme participant à des usages différents. L'impact de la lemmatisation est illustré par la *figure 11* suivante.

- *filtrage_local.* Le filtrage local concerne surtout le contexte thématique. Son application dégrade les performances de la désambiguïsation.

		sim		*infos*		*comp*		*forme*	
		1 < 2		1 < 2		1 < 2		1 < 2	
infos	**1** ∧ **2**	81.6	81.6						
		81.6	**90.5**						
comp	**1** ∧ **2**	81.3	81.8	81.6	81.8				
		81.3	**90.5**	81.4	**90.5**				
forme	**1** ∧ **2**	81.6	82.7	80.8	82.7	80.9	82.7		
		81.3	**90.5**	81.7	**90.5**	81.8	**90.5**		
filtrage_local	**1** ∧ **2**	80	84.6	75.9	84.9	76	84.5	75.8	84.5
		81.6	**90.5**	81.6	**90.5**	85.7	**90.5**	82.7	**90.5**

- **Tableau 7** - *La force d'association entre les variables deux à deux, représentée par la meilleure performance atteinte avec les combinaisons réalisant les modalités, en entrée de ligne et en entrée de colonne, des variables concernées.*

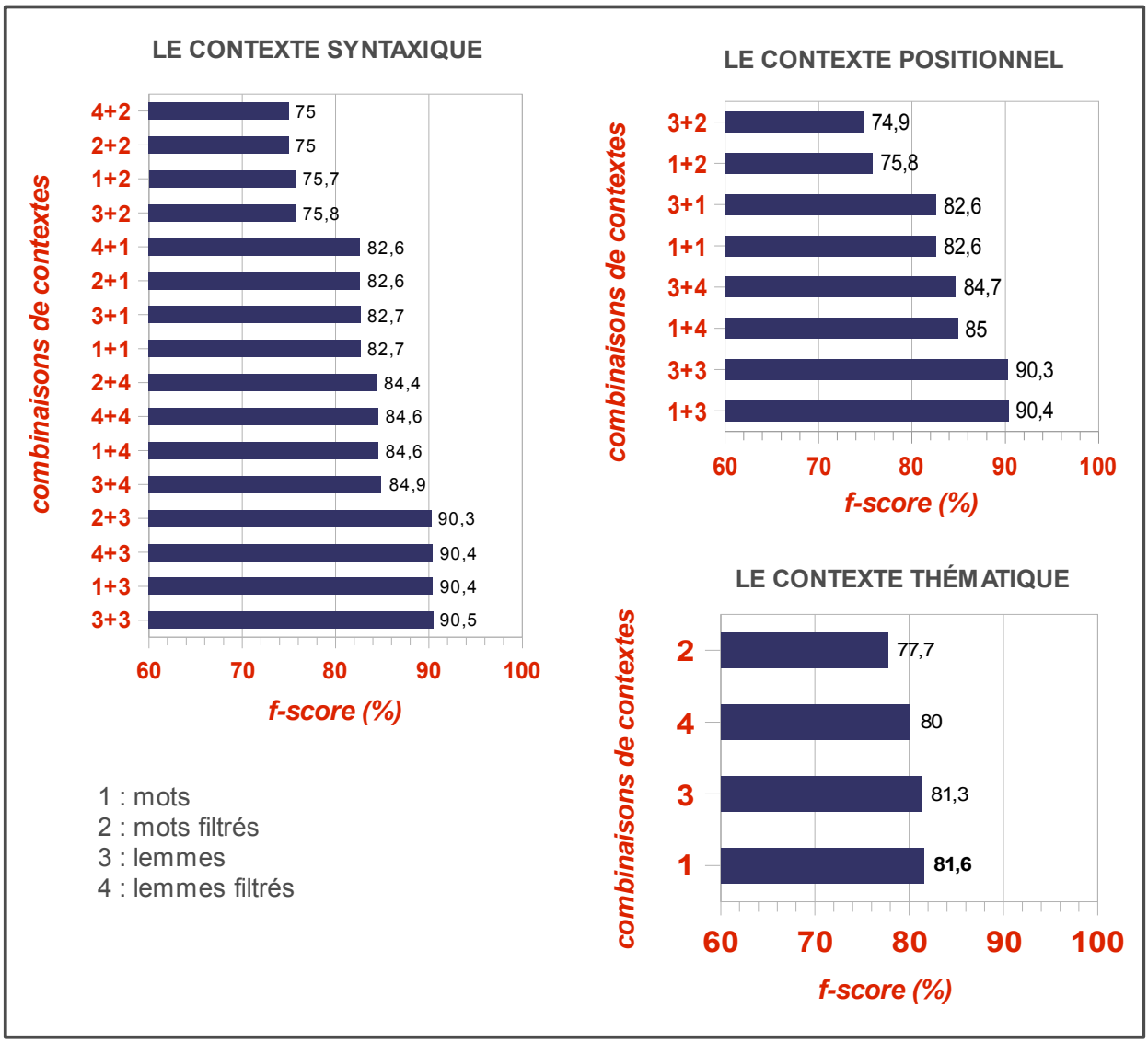
b.2. Analyse de l'évolution des performances

Les diagrammes de la *figure 11* suivante représentent l'évolution des performances de notre méthode relativement aux variables *forme* et *filtrage_local*. Les légendes des axes des ordonnées sont le mode de représentation des contextes du mot : *mode_pos+mode_ctxt* (pour le contexte positionnel), *mode_synt+mode_ctxt* (pour le contexte syntaxique) et *mode_ctxt* (pour le contexte thématique). Ces diagrammes prouvent que la corrélation entre les performances du système de désambiguïsation d'une part, et ces deux variables, d'autre part, est consistente. En effet, dans les trois diagrammes, on remarque, dans les étiquettes de l'axe des ordonnées, un premier tri selon la forme des éléments du contexte thématique, puis un second tri selon le filtrage local de ce même contexte, et enfin un troisième tri : selon la forme des voisins du mot pour le contexte positionnel et selon le filtrage local pour le contexte syntaxique.

- En ce qui concerne la variable *forme*, dans les contextes positionnel et syntaxique, les performances augmentent avec la lemmatisation du contexte thématique : les étiquettes de l'axe des ordonnées sont ordonnées en *mots* puis *lemmes*.

- En ce qui concerne le filtrage local, son impact est secondaire par rapport à la forme dans les contextes syntaxique et positionnel. On observe, dans ces deux contextes, l'évolution *mots filtrés, mots, lemmes filtrés et lemmes*. En revanche, dans le contexte purement thématique, l'évolution des performances suit exclusivement l'ordre des modalités de la variable *filtrage_local*, aucun ordre n'étant observé sur la forme.
- Pour le contexte de type positionnel, on remarque un troisième ordre de tri selon la forme des voisins du mot, et qui est l'inverse de l'ordre des formes observé pour le contexte thématique de ce même type de contexte : *lemmes* puis *mots*. On en déduit que la lemmatisation est plus intéressante lorsqu'elle est appliquée au contexte thématique. Une fois encore (comme pour la lemmatisation), cela peut s'expliquer par la taille de ces deux contextes : l'ensemble des voisins immédiats d'un mot, dans un corpus, est relativement petit, par rapport à l'ensemble de ses cooccurrents.

- Pour le contexte syntaxique, le troisième critère de tri porte sur le filtrage local des cooccurrents syntaxiques. Pour le même motif, peut-être, que pour la lemmatisation des voisins du mots.

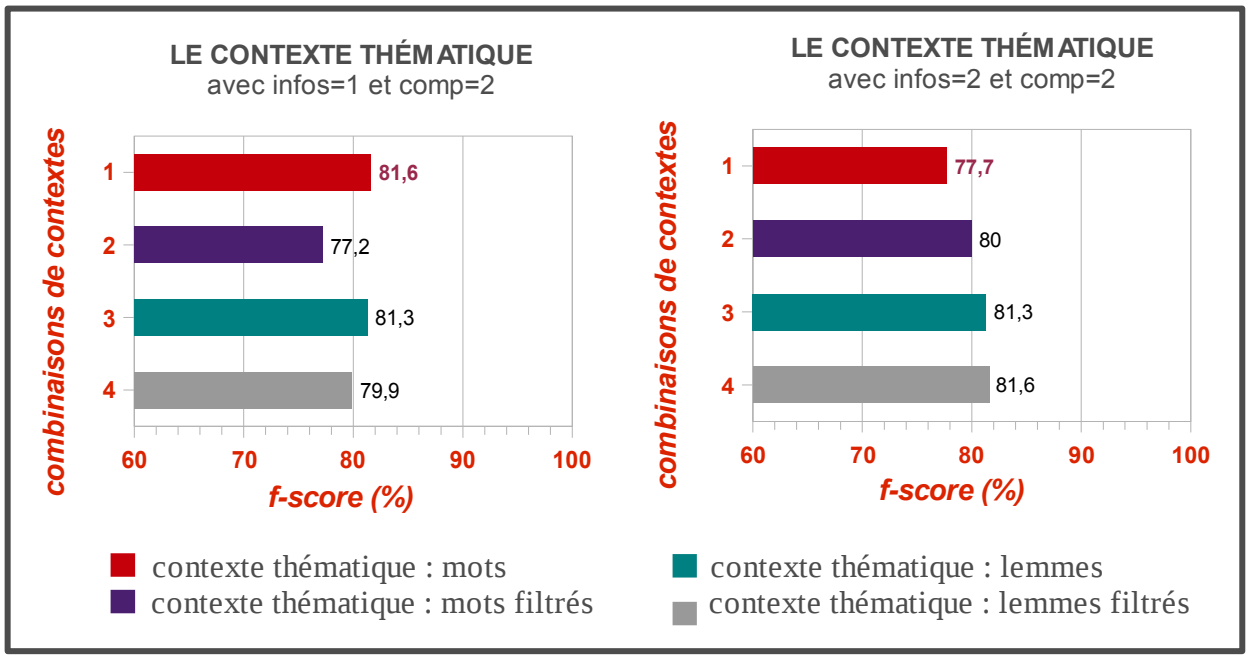


- **Dessin 11** – *Évolution du score global du module de désambiguïsation en fonction des variables* forme *et* filtrage_local -

c. Évaluation des différents types de contextes

c.1. Le contexte thématique

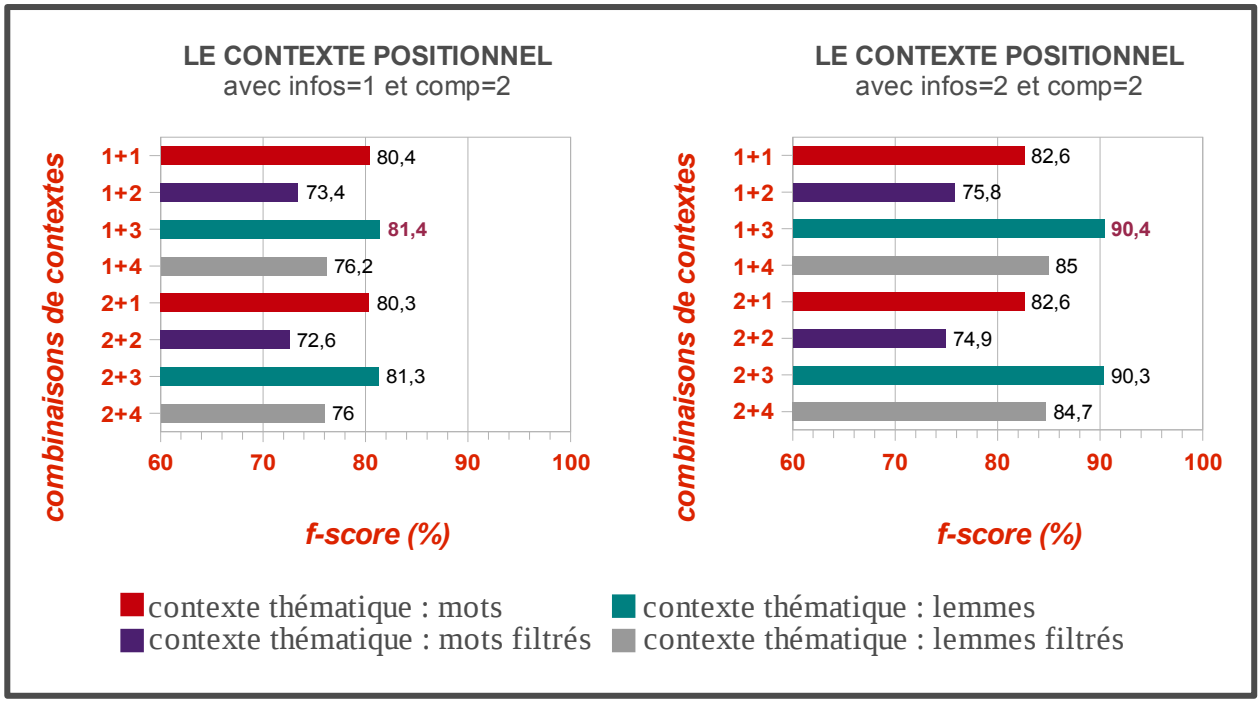
La meilleure performance atteinte avec le contexte thématique simple s'élève à 81,6%. Elle est obtenue avec des contextes composés de mots (variable *forme*), représentés par des traits de type 1 ou 2 (variable *infos*), et sans marquage des mots composés. Le type des traits, leur forme (variable *forme*) et le marquage des mots composés affectent peu les résultats : pratiquement le même score est atteint sans mots composés et avec des lemmes (81,3%), ainsi qu'avec marquage des mots composés, avec des lemmes et des traits de type 1 ou 2. L'évolution des performances, en ordre croissant, est : *mots filtrés*, *lemmes filtrés*, *lemmes* et *mots*.



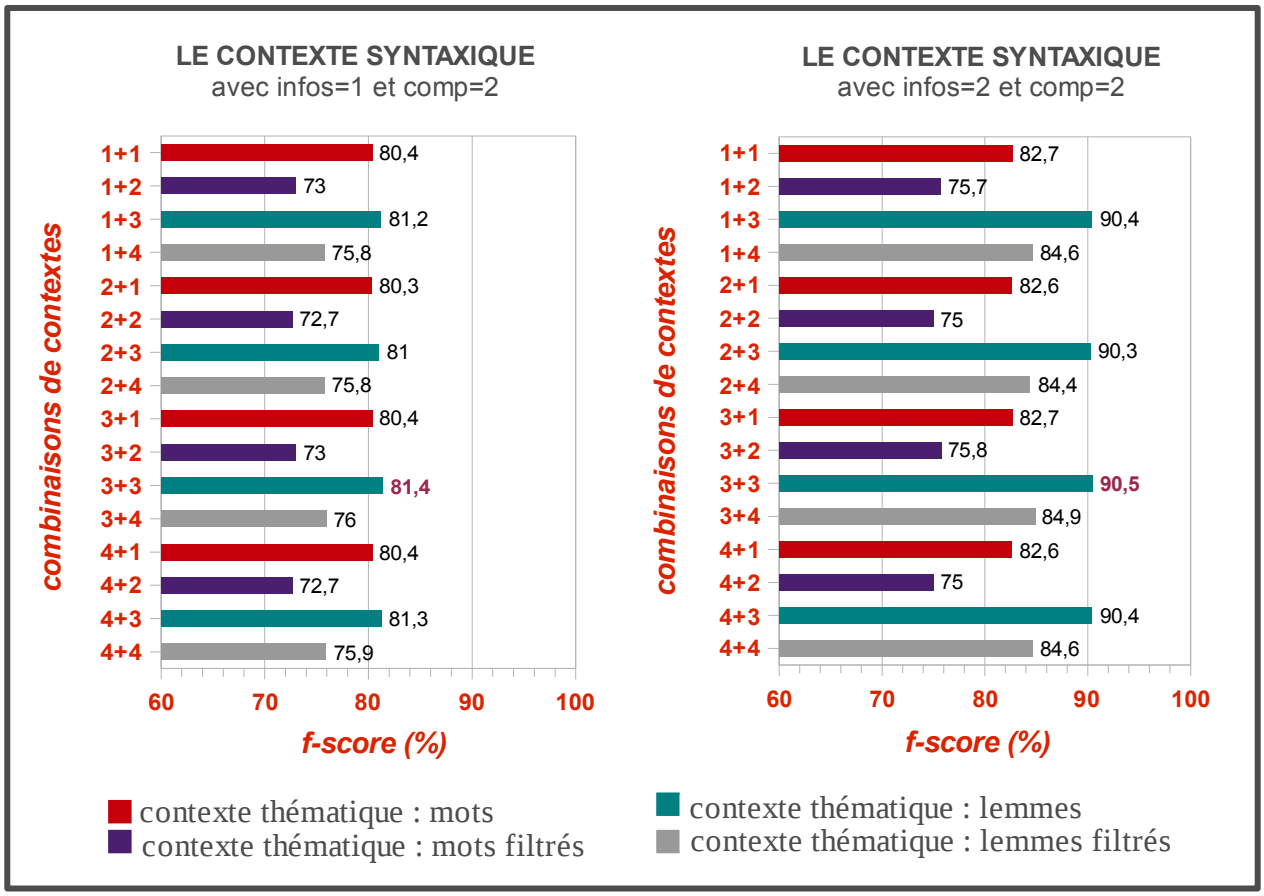
- Dessin 12 – *Évaluation du contexte thématique* -

c.2. Les contextes positionnel et syntaxique

La meilleure performance atteinte avec les contextes syntaxique et positionnel s'élève à 90,5%, c' est aussi la meilleure performance obtenue par le système. Elle est obtenue avec des traits de type 2, avec marquage des mots composés, les voisins du mot ou ses cooccurrents syntaxiques représentés par leur forme fléchie ou par leur lemme indifféremment, et le contexte thématique toujours lemmatisé.



- Dessin 13 – *Évaluation du contexte positionnel* -



- **Dessin 14** – *Évaluation du contexte syntaxique* -

4.3.2. Évaluation qualitative

USAGES SELON LES DICTIONNAIRES	USAGES DÉCOUVERTS
Conclusion	
1 : {*to conclude; conclusion; completion; outcome*} 2 : {*conclusion*} 3 : {*findings; outcome*} 4 : {*in conclusion*}	1 : {*in conclusion*} 2 : {*outcome, conclude, conclusion*} 3 : {*completion*} 4 : {*outcome, finding*}
Conseil	
1 : {*advice*} 2 : {*board; council*} 3 : {*consultancy*}	1 : {*board, advice*} 2 : {*council, board*} 3 : {*consultancy*}
Matière	
1 : {*in relation to; as regards; in terms of; in the area of*} 2 : {*raw materials*} 3 : {*matter; issue*}	1 : {*in term of, as regard, in relation to*} 2 : {*raw material*} 3 : {*issue, in the area of, matter*}

USAGES SELON LES DICTIONNAIRES	USAGES DÉCOUVERTS
Réserve	
1 : {*reserve; provision; funds*} 2 : {*reservation*} 3 : {*unreservedly; unqualified; totally; without any reservations*}	1 : {*totally, unreservedly*} 2 : {*totally, unqualified*} 3 : {*without any reservation*} 4 : {*provision, fund, reservation, reserve*}
Porter	
1 : {*carry*} 2 : {*wear; carry*} 3 : {*bear on*} 4 : {*relate to*} 5 : {*increase*}	1 : {*relate to*} 2 : {*wear, increase, carry*} 3 : {*bear on*}

- Tableau 8 – *Évaluation qualitative des classes découvertes pour les noms par la méthode KNN -*

On remarque que les classes de traductions découvertes ne correspondent pas souvent avec les groupes de classes représentant les sens du mot selon le dictionnaire traditionnel. Cela peut être dû, par exemple, à des usages métaphoriques des traductions du mot, et donc des contextes similaires à ceux de traductions relevant d'un sens différent. [Apidianaki 2008] propose, comme solution, d'utiliser les cooccurrents du second ordre du mot (les cooccurrents des cooccurrents directs, du premier ordre).

4.3.3. Conclusion

Nous avons expérimenté toutes les combinaisons possibles des modalités des variables que nous avons définies, dans toutes les combinaisons possibles de nos trois types de contextes. S'il est une conclusion que l'on peut en tirer, c'est qu' « il n'est [...] pas de 'bon' modèle dans l'absolu » (Habert et al. 1997). La performance optimale de notre méthode n'est-elle pas obtenue avec six combinaisons possibles sur les modalités des variables, et avec deux types de contextes (2 combinaisons avec le contexte positionnel et 4 avec le contexte syntaxique). Les effets des différents types d'informations linguistiques sont donc incertains, dépendants d'une part des autres types d'informations linguistiques avec lesquels ils sont combinés, et, d'autre part, du type d'application dans laquelle s'inscrit la tâche de désambiguïsation sémantique. On a vu, par exemple, que la lemmatisation n'améliore pas systématiquement les performances des outils de désambiguïsation sémantique. Comme le soulignent [Habert et al. 1997], « il n'existe que des modèles opératoires qui sont utiles à l'utilisateur final dans le cadre d'une application ». En conséquence, l'évaluation de la désambiguïsation sémantique ne peut se faire que dans le cadre d'une application donnée, en fonction du but de cette application.

4.4. Évaluation comparative

4.4.1. Avec la méthode de référence

Les performances de notre module de désambiguïsation sont comparables à celles atteintes par [Apidianaki 2009a], qui s'élèvent à **76,99%**, pour 150 mots ambigus. Dans cette expérience, le corpus utilisé est la version anglais-grec du corpus INTERA (Gavrilidou et al. 2004). Chaque

version du corpus est lemmatisée et étiquetée morpho-syntaxiquement et un filtrage est appliqué, à la suite duquel seuls les lemmes des noms, verbes et adjectifs du contexte sont conservés. Le lexique bilingue des mots ambigus est généré automatiquement à partir d'un corpus aligné automatiquement au niveau des unités lexicales. Avec les mêmes prétraitements de notre corpus, les performances de notre classifieur s'élèvent à 79,96%. La différence de performances peut s'expliquer par le fait que les paires de langues traitées relèvent d'un niveau de difficulté qui n'est pas comparable : la traduction de l'anglais vers le grec, deux langues d'alphabets différents, est une tâche sûrement plus complexe.

4.4.2. Avec la méthode *KNN* classique

a. Les meilleures performances atteintes

Le meilleur score atteint avec la méthode *KNN* classique s'élève à **88,6%**. Nous l'avons obtenu avec la combinaisons de contextes suivante (les valeurs du tableau correspondent aux valeurs des paramètres présentés au *tableau 2* plus haut) :

COMBINAISONS	C1	C2	C3	C4	C5
ctxt_type	2	2	2	2	3
infos	2	2	2	2	2
comp	1	1	1	1	1
mode_pos	1	1	2	2	0
mode_synt	0	0	0	0	1
mode_ctxt	1	3	1	3	1
sim	1	1	1	1	1
SCORES	88,6	88	88,3	88,2	88

- TABLEAU 9 – *Les combinaisons optimales pour la désambiguïsation sémantique par la méthode* KNN *classique* -

Les meilleures performances sont donc réalisées quand :

- les traits sont du type *forme#étiquette_grammaticale*;
- les mots composés sont marqués;
- soit le contexte positionnel est composé de mots, filtrés localement ou non, et le contexte thématique est composé de mots ou de lemmes non filtrés localement (les combinaisons 1 à 4 du *tableau 9*);
  soit le contexte syntaxique et le contexte thématique sont composés de mots non filtrés localement (la combinaison 5 dans le *tableau 9*);
- et tous les traits sont pris en compte pour la recherche des plus proches voisins d'un mot.

Lorsque la recherche des plus proches voisins se fait sur la base des traits forts uniquement, le score global diminue. La méthode de « filtrage par les groupes syntaxiques » n'est donc pas efficace dans le cadre de cette méthode.

b. Évaluation des paramètres

b.1. Les tendances globales des paramètres

Les graphiques de la *figure 15* suivante nous mènent aux mêmes conclusions concernant les tendances globales de quatre des cinq paramètres évalués : *infos*, *comp*, *forme* et *filtrage_local*.

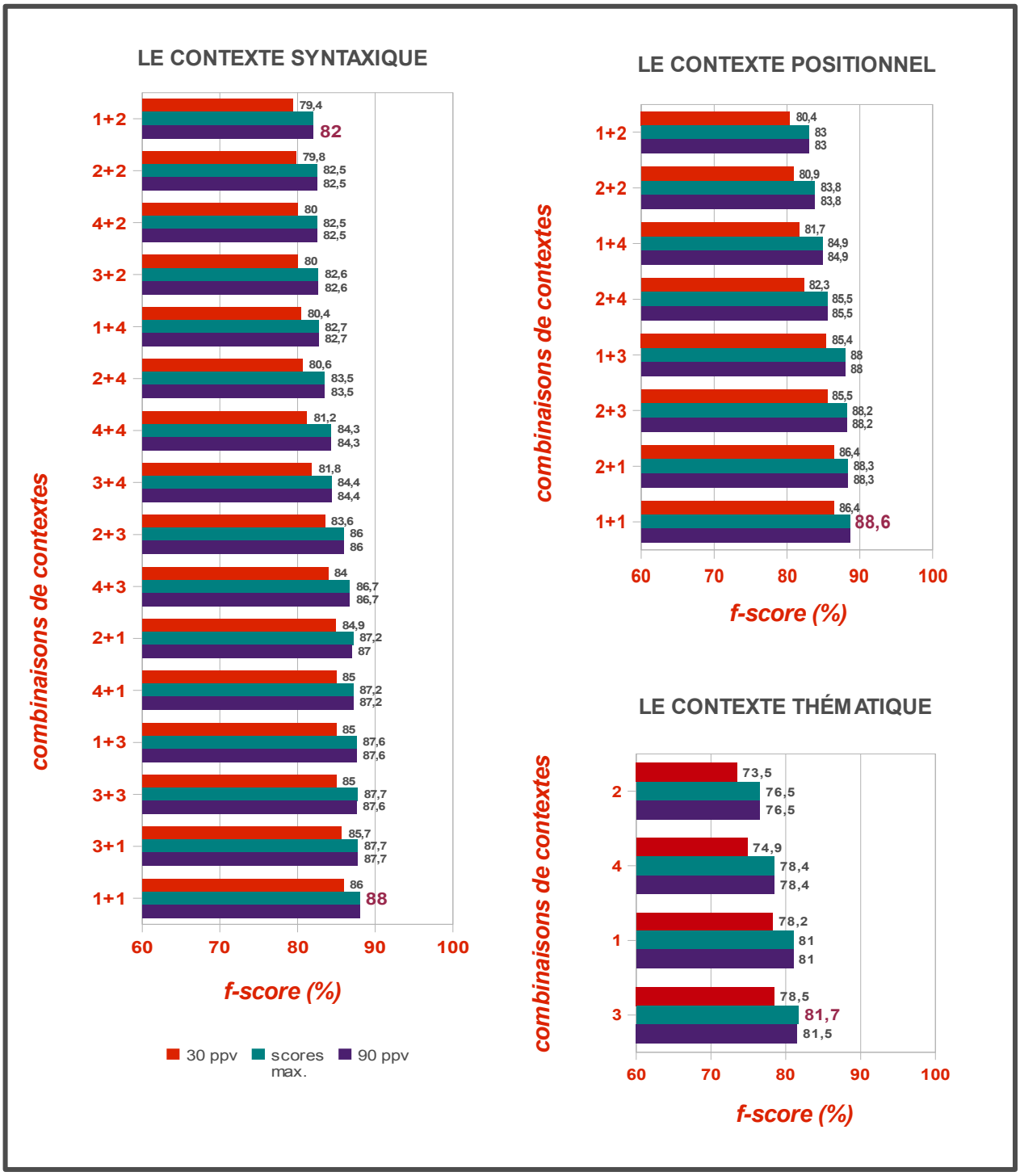
Pour le cinquième paramètre, représenté par la variable *sim*, les résultats que nous avons obtenu avec la méthode *KNN* classique sont à l'opposé de ceux obtenus avec la méthode [Apidianaki 2009]. En effet, les performances sont plus faibles lorsque la recherche des plus proches voisins d'un mot est réalisée uniquement sur la base des traits forts de son contexte.

b.2. Analyse de l'évolution des performances

Comme pour la méthode que nous avons développée, les performances de la méthode *KNN* classique évoluent principalement en fonction de la forme du contexte thématique. Les meilleures performances sont atteintes lorsque ce dernier est représenté par tous les mots ou les lemmes qui le composent.

En ce qui concerne le contexte positionnel et le contexte syntaxique, les meilleurs performances sont atteintes lorsque les traits forts (cooccurrents syntaxiques ou voisins gauche et droit) sont des mots non filtrés localement. En ce qui concerne le contexte thématique, les performances sont meilleurs lorsque le contexte est représenté par des lemmes, non filtrés localement. Dans tous les cas, donc, le filtrage local dégrade les performances.

Enfin, l'assignation d'un poids supérieur aux traits qui représentent les contextes positionnel et syntaxique (par rapport aux traits du contexte thématique) permet d'obtenir des performances supérieures à celles obtenues avec un contexte exclusivement thématique : 88,6% (pour le contexte positionnel) et 88% (pour le contexte syntaxique) contre seulement 81,7% (pour le contexte exclusivement thématique). Au regard de cette analyse, l'idée d'un traitement différentiel des contextes du mot reste pertinente.



- **Dessin 15** – *Évolution du score global en fonction des variables* forme *et* filtrage_local -

4.4.2. Avec la méthode *SVM*

a. Les meilleures performances atteintes

La méthode SVM a été évaluée sur la combinaison de contextes qui nous a permis d'obtenir les meilleures performances avec la méthode [Apidianaki 2009], à savoir :

- les traits sont du type *forme#étiquette_grammaticale*;
- les mots composés sont marqués;
- le contexte syntaxique du premier ordre ou le contexte positionnel sont composés de mots ou de lemmes, avec ou sans filtrage local;
- le contexte thématique est composé de lemmes et non filtré localement.
- et seuls les traits forts, cooccurrents du premier ordre ou voisins gauche et droit, sont pris en compte pour le regroupement des classes sémantiquement similaires;

Le score global atteint avec la méthode *SVM* s'élève à **70%**. La différence des performances des deux méthodes (*KNN* et *SVM*) peut s'expliquer par le fait que la méthode *SVM* nécessite un nombre très important d'exemples d'entraînement.

5. L'alignement de notre corpus

5.1. Introduction : les corpus parallèles

Un corpus parallèle est un ensemble de textes dans lequel les unités textuelles d'un document en LS sont mises en correspondance avec leur(s) traduction(s) dans une ou plusieurs LC.

Les corpus parallèles (ou *bi-textes*/*multi-textes*, selon Harris 1988) sont systématiquement utilisés par les chercheurs en traitement automatique du langage naturel (TAL) depuis les années 1980. À la fin des années 1950, quelques expériences en traduction automatique à l'aide de corpus parallèles ont été réalisées. Mais les capacités matérielles limitées des ordinateurs de l'époque, en matière de stockage des données et de puissance de calcul, empêchaient ce type de ressource linguistique de manifester pleinement l'intérêt de leur utilisation.

La première tentative d'alignement automatique (ou *appariement*) de corpus fut menée, en 1987, par Martin Kay et Martin Röscheisen, de Xerox (Kay et Röscheisen 1988). Depuis, les méthodes dans ce domaine ont fait foison. Et les applications du TAL qui utilisent des connaissances linguistiques acquises à partir de corpus parallèles sont nombreuses : la construction de dictionnaires bilingues et multilingues, l'extraction de terminologies spécifiques, l'extraction de connaissances pour la recherche d'informations, et, plus que toutes, la traduction automatique. Les corpus parallèles, sont, de ce fait, un matériau de plus en plus recherché, et de plus en plus disponible, dans des langues toujours plus variées.

Trois niveaux d'alignement ont été testés, à ce jour, correspondant à trois types d'unités textuelles : la phrase (on parle, en anglais, de *sentence alignment*), le syntagme (*clause ou phrase alignment*) et le mot (*word to word alignment*). Pour une brève description des méthodes d'appariemment dans ces trois niveaux, voir [Véronis 2000].

5.2. L'alignement de mots : un bref état de l'art

Dorénavant, l'appariement de corpus étant une tâche intermédiaire de notre travail, on s'intéressera uniquement à l'alignement de mots. On remarquera d'ailleurs que, pour la même raison, l'évaluation de notré méthode d'alignement n'est, à ce jour, pas encore réalisée.

L'alignement mot à mot se sert de corpus parallèles de phrases pour déterminer les liens de traduction entre mots simples et/ou complexes des deux langues. La découverte des liens de traduction entre mots se fait de deux manières différentes en fonction de l'approche adoptée. On distingue, en effet, parmi les méthodes d'alignement de mots qui ont été proposées, l'approche heuristique et l'approche statistique.

5.2.1. les méthodes heuristiques

a. Présentation générale

Les méthodes heuristiques d'appariement de textes utilisent des informations linguistiques qu'elles tirent exclusivement du texte dans lequel apparaissent les mots à aligner, à savoir la phrase. Ces

méthodes procèdent en deux étapes pour l'alignement des mots d'un couple (*s*,*c*) de phrases alignées, *s* étant une phrase en langue source et *c* sa correspondante en langue cible.

Lors de la première étape, des ancres sont posées au sein des segments du couple (*s*,*c*). Ces ancres sont soit des mots spécifiques listés dans un lexique bilingue, soit des cognats. Un cognat, noté (s_i,c_j), où s_i est un mot de *s* et c_j, un mot de *c,* est un couple mots apparentés : s_i et c_j entretiennent un lien de ressemblance graphique. Ce sont principalement des symboles tels que les dates, les nombres ou les pourcentages, des noms propres, et des mots à l'orthographe relativement proche.

La seconde étape consiste à compléter l'alignement de la phrase en projetant les liens de traduction des ancres (ou *couples amorces*) sur les autres mots de la phrase.

b. Les heuristiques utilisées pour le marquage des ancres

La localisation des ancres peut être effectuée de deux manières différentes. Pour la première, un lexique bilingue de « mots-ancres » est utilisé. Les ancres sont simplement retrouvées et marquées, comme dans [Debili et Sammouda 1992]. Le lexique est parfois utilisé en combinaison avec diverses mesures statistiques faisant intervenir entre autres la longueur des phrases des deux langues (Haruno et Yamazaki 1997; Johansson, Ebeling et Hofland, 1996). La seconde méthode, utilisée par [Simard, Foster et Isabelle 1992], [Church 1993] et [McEnery et Oakes 1995], se base exclusivement sur des informations lexicales pour détecter des cognats dans chaque couple de phrases alignées. Cette méthode est la plus utilisée dans les méthodes heuristiques.

Outre celle de chaînes complètement identiques (*inflation*, *inflation*), les cognats se sont vus assigner diverses définitions dans la littérature sur l'alignement de textes. Pour [Simard, Foster et Isabelle 1992], ce sont les mots qui commencent par un même préfixe de quatre lettres (*financier*, *financial*). [McEnery et Oakes 1995] y ont ajouté les mots dont la similarité selon la formule de Dice (appliquée aux lettres qui les composent) est supérieure à un seuil donné. D'autres heuristiques sont également utilisées pour comparer deux chaînes candidates :

- elles contiennent un nombre en commun (*G8*, *G8*),
- elles ont un nombre minimal, ou une proportion minimale, de sous-chaînes communes (*gouvernement*, *government* : ***go***-*ou*-*uv*-***ve*-*er*-*rn***-*ne*-*em*-***me*-*en*-*nt*** ↔ ***go***-*ov*-***ve*-*er*-*rn***-*nm*-***me*-*en*-*nt***).
- etc.

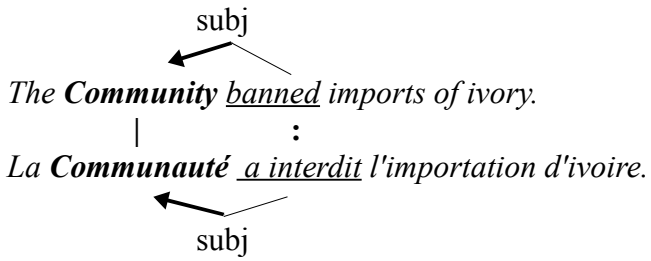
c. Les règles de projection

Les règles utilisées pour la projection des liens de traduction des ancres se basent, en général, sur les différents types de liens que les mots de la phrase peuvent entretenir avec les ancres (proximité dans l'espace, cooccurrence simple, cooccurrence syntaxique).

Le lien le plus couramment utilisé pour la projection est la cooccurrence syntaxique. [Debili et Zribi 1996] ont utilisé une règle de propagation des liens, qu'ils ont désignée comme règle de *raisonnement par analogie*. [Ozdowska 2004] préfère parler de règle *de propagation des liens d'appariement suivant les relations de dépendance syntaxique*.

Selon cette règle, les mots de la phrase sont alignés sur la base des liens de cooccurrence syntaxique, préalablement mis en évidence, qu'ils entretiennent avec les ancres. Dans l'exemple suivant, tiré de [Ozdowska 2008], le couple amorce est (*Communauté*, *Community*). En partant de

ce couple, on peut apparier (*a interdit*, *banned*), tous deux en relation de sujet avec les éléments du couple amorce.



L'hypothèse sous-jacente des méthodes d'appariemment dirigées par la syntaxe, formulée par [Debili et Zribi 1996], est que *« les liaisons paradigmatiques peuvent aider à déterminer les relations syntagmatiques, et inversement »*.

Les limites de la techinque d'appariement par propagation des liens de traduction proviennent de ce que sa faisabilité est conditionnée par la disponibilité d'un corpus aligné au niveau des phrases, d'outils d'analyse linguistique, de préférence de performance similaire pour les deux langues concernées, et, surtout, par l'existence d'une proportion minimale de couples amorces au sein de chaque couple de phrases du corpus aligné. Si cette dernière condition, en particulier, n'est pas satisfaite, alors l'appariement est tout simplement irréalisable.

d. Conclusion

La méthode des ancres ne peut fonctionner pour des paires de langues éloignées qu'à condition que la quantité de symboles (dates, noms propores, etc.) par couple de phrases soit suffisant. Pour des paires de langues d'alphabets différents, la méthode n'est pas applicable. [Wu 2004] démontre enfin que les statistiques sur la corrélation entre les longueurs des phrases de divers couples de langues ne constituent pas toujours une information pertinente pour l'alignement de textes.

Depuis, des combinaisons plus complexes d'informations, incluant des statistiques, sur les cooccurrences des mots en LS et en LC ou sur leur position dans la phrase, par exemple, ont été proposées, en particulier par [Langlais et El-Beze 1997].

5.2.2. Les méthodes statistiques

a. L'état de l'art actuel

En 1949, Warren Weaver suggérait d'appliquer les techniques statistiques, alors émergentes, de la théorie de la Communication, au problème de la traduction automatique (Weaver 1955). Pour diverses raisons philosophiques et théoriques, cette suggestion eut un faible écho au sein de la communauté des chercheurs en traduction automatique et les rares expériences recourrant à ces techniques furent vite abandonnées. Puis, au début des années 90, [Brown, Lai et Mercer 1991] et [Gale et Church 1991] développèrent deux méthodes d'alignement de corpus, au niveau des phrases, intégrant des calculs statistiques basés sur la longueur des phrases (en termes de lettres ou de mots) dans les deux langues étudiées, ainsi que des techniques de programmation dynamique (l'algorithme EM, Espérance et Maximisation). En s'inspirant de ces travaux, le groupe de traduction statistique d'IBM à Yorktown (Brown, Della Pietra, Della Pietra et Mercer) développa cinq modèles purement mathématiques et de plus en plus sophistiqués (voir [Brown et al. 1993] pour une description

détaillée) qui procèdent par induction sur les cooccurrences entres les mots des deux langues étudiées. Ces cinq modèles sont aujourd'hui l'état de l'art en matière de traduction automatique statistique et d'alignement de mots et les systèmes les plus récemment développés pour l'alignement de textes en sont fortement inspirés.

b. La place des ressources linguistiques

Depuis [Véronis 2000], les expériences en appariemment de mots à l'aide de modèles statistiques tendent à limiter les ressources linguistiques exogènes utilisées à des outils d'analyse linguistique, tels que les étiqueteurs morpho-syntaxiques, les lemmatiseurs, et les analyseurs syntaxiques. Ces derniers, en particulier, sont de plus en plus souvent intégrés dans la chaîne de traitement des systèmes ainsi développés, pour la langue source uniquement (Fox 2002; Lin et Cherry 2003), ou pour les deux langues de travail (Wu 2000). Ils interviennent, en général, après les calculs syntaxiques, et pour confirmer ou corriger les appariements que ces derniers ont produits.

Dans leur conclusion, [Brown et al. 1993] précisent qu'il n'était pas dans leur intention « d'ignorer la linguistique, ni de la remplacer », et qu'au contraire, celle-ci, combinée aux mathématiques, devrait nous permettre de construire des systèmes de traitement du langage encore plus performants. Le recours, par les méthodes statistiques d'alignement, à des ressources linguistiques exogènes, reste toutefois relativement plus restreint que celles utilisées par les méthodes heuristiques. Pour cette raison, entre autres, les systèmes statistiques sont moins coûteux. Outre leur simplicité, et donc, leur facilité d'implémentation, ils sont plus facilement transposables à des couples de langues différents. Les techniques statistiques d'appariement sont, par ailleurs, capables d'accomplir l'alignement de textes dans des langues pour lesquelles peu d'outils et de ressources linguistiques sont disponibles. Ces critères étaient particulièrement attrayants pour nous, l'appariement de corpus étant une tâche intermédiaire dans notre travail. Nous avons donc adopté cette approche.

5.3. Notre approche

La méthode que nous avons développée pour l'alignement des mots de notre corpus relève de l'approche statistique décrite plus haut (section 5.2.2). Cette méthode se base sur les cooccurrences entre les mots en langue source d'une part, et les mots en langue cible d'autre part, pour découvrir des liens de traduction entre mots au niveau de la phrase. Il s'agit, en fait, de la méthode décrite dans [Gaussier, Hull et Aït-Mokhtar 2000], à quelques adaptations près. Cette méthode est principalement inspirée de deux références : [Melamed 1998] et [Hiemstra 1996]. Il s'agit d'une méthode d'alignement mot à mot : un mot en langue source est aligné avec un et un seul mot en langue cible.

5.3.1. Préliminaires : les tables de contingence

a. Notations

On considère deux variables qualitatives observées simultanément sur n observations. Ces deux variables sont S, les mots du corpus en langue source, et C, les mots du corpus en langue cible. La variable S possède u modalités, notées s_1, ..., s_i, ..., s_u. La variable C possède v modalités, notées c_1, ..., c_j, ..., c_v.

Nos observations sont les cooccurrences entre les mots en langue source et les mots en langue cible, autrement dit, le nombre d'occurrences des différents couples (s_i, c_j) dans le corpus. Un mot en langue source s_i et un mot en langue cible c_j cooccurrent lorsqu'il existe un couple (s,c) de phrases alignées, et un couple (s_i, c_j) de mots, tels que s_i apparaît dans s et c_j apparaît dans c. Par exemple, dans le couple de phrases (*les votes sont clos*, *the vote is closed*), on observe, les couples de cooccurrents suivants : (*les*, *the*), (*les*, *vote*), (*les*, *is*), (*les, closed*), (*votes*, *the*), (*votes*, *vote*), (*votes*, *is*), (*votes*, *closed*), ..., (*clos*, *closed*),

Une **table de contingence** est un tableau à double entrée dans lequel on dispose les modalités de S en lignes et les modalités de C en colonnes. Ce tableau est donc de dimension s*c.

Chaque ligne et chaque colonne correspond à un sous-échantillon particulier :

- La ligne d'indice i est la répartition sur c_1, ..., c_j, ..., c_v des observations pour lesquelles la modalité s prend la valeur s_i : pour nous, c'est la répartition des fréquences de cooccurrence d'un mot s_i en langue source avec chacun des mots c_j en langue cible. La ligne d'indice i est donc la distribution empirique de la modalité s_i par rapport à chacune des modalités c_j.

- La colonne d'indice j est la répartition sur s_1, ..., s_i, ..., s_u des observations pour lesquelles la modalité c prend la valeur s_j : la fréquence de cooccurrence d'un mot c_j en langue cible avec chacun des mots s_i en langue source. Et la colonne d'indice j est la distribution empirique de la modalité c_j par rapport à chacune des modalités s_i.

Une table de contingence se présente sous la forme suivante :

	c_1	...	c_j	...	c_v	sommes
s_1	n_{11}		n_{1j}		n_{1v}	$n_{1\bullet}$
...						...
s_i	n_{i1}		n_{ij}		n_{iv}	$n_{i\bullet}$
...						...
s_u	n_{u1}		n_{uj}		n_{uv}	$n_{u\bullet}$
sommes	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet v}$	$n_{\bullet\bullet}$

b. Effectifs

- n_{ij} est l'**effectif conjoint** des modalités s_i et c_j.
  C'est le nombre d'observations réalisant simultanément les modalités s_i de S et c_j de C, c'est-à-dire la fréquence de cooccurrence entre s_i et c_j, ou, tout simplement, le nombre d'occurrences du couple (s_i, c_j).Les effectifs conjoints d'une table de contingence sont obligatoirement des valeurs brutes (pas de pourcentages ou autres fréquences relatives).

On appelle **effectifs marginaux** les quantités $n_{i\bullet}$ et $n_{\bullet j}$, définis comme suit :

$$n_{i\bullet} = \sum_{j=1}^{v} n_{ij} \quad \text{et} \quad n_{\bullet j} = \sum_{i=1}^{u} n_{ij}$$

- $n_{i\bullet}$ est l'**effectif marginal** de s_i.
  C'est la somme des valeurs de la ligne d'indice i : le nombre d'observations pour lesquelles la modalité s prend la valeur s_i. Autrement dit, c'est le nombre de couples (s_i, c_j) contenant un s_i donné.

- $n_{\bullet j}$ est l'**effectif marginal** de c_j.
  C'est la somme des valeurs de la colonne d'indice j : le nombre d'observations pour lesquelles la modalité c prend la valeur c_j. Autrement dit, le nombre de couples (s_i, c_j) contenant un c_j donné.

- $n_{\bullet\bullet} = n$
  On note ainsi la somme des effectifs marginaux des lignes ou des colonnes.

  $$n_{\bullet\bullet} = \sum_{i=1}^{u} n_{i\bullet} = \sum_{j=1}^{v} n_{\bullet j}$$

c. Fréquences relatives

- $f_{j|i}$ est la **fréquence conditionnelle** de la modalité c_j de C relativement à la modalité s_i de S. C'est la probabilité que la c_j cooccurre avec le s_i . $f_{i|j}$ est la fréquence de la modalité s_i de S conditionnelle à la modalité c_j de C. On a :

  $$f_{j|i} = n_{ij} / n_{i\bullet} \quad \text{et} \quad f_{i|j} = n_{ij} / n_{\bullet j}$$

- c_{ij} est la **fréquence conjointe** des modalités s_i de S et c_j de C. C'est la probabilité de cooccurrence d'un mot s_i avec un mot c_j.

  $$c_{ij} = n_{ij} / n_{\bullet\bullet}$$

d. Profils

- En divisant l'effectif conjoint n_{ij}, pour un i donné et pour j allant de 1 à v, par l'effectif marginal $n_{i\bullet}$, on obtient l'ensemble des fréquences conditionnelles de la variable C relativement à la modalité s_i de S (soit les fréquences conditionnelles de chacun des mots c_j en langue cible relativement à un mot s_i donné en langue source). Cet ensemble de fréquences forme le **i-ème profil-ligne**, il est défini comme suit :

  $$\{n_{i1} / n_{i\bullet}, \ldots, n_{ij} / n_{i\bullet}, \ldots, n_{iv} / n_{i\bullet}\}$$

- De même, en divisant l'effectif conjoint n_{ij}, pour un j donné et pour i allant de 1 à u, par l'effectif marginal $n_{\bullet j}$, on obtient l'ensemble des fréquences conditionnelles, $f_{i|j}$, de la variable S relativement à la modalité c_j de C (soit les fréquences conditionnelles de chacun des mots s_i en langue source relativement à un mot c_j donné en langue cible). C'est le **j-ème profil-colonne**, défini par :

  $$\{n_{1j} / n_{\bullet j}, \ldots, n_{ij} / n_{\bullet j}, \ldots, n_{uj} / n_{\bullet j}\}$$

5.3.2. Les algorithmes EM

L'algorithme EM est une technique très populaire d'estimation des paramètres de modèles probabilistes afin de retrouver le paramètre qui réalise le maximum de vraisemblance. L'intérêt pour cet algorithme est dû à sa simplicité d'implémentation et à ses bonnes propriétés de convergence dès les premières itérations, entre autres. À l'inverse, l'une de ses faiblesses reconnues est sa convergence très lente dans les dernières itérations. Une grande part des travaux sur les méthodes EM est donc concentrée sur les multiples façons d'augmenter la vitesse de convergence des algorithmes EM (nous ne traiterons pas ce point, ici).

a. Description générale

a.1. Préliminaires

On considère un corpus de segments $s = s_1, ..., s_i, ..., s_p$ en langue source et de segments $c = c_1 , ..., c_j , ..., c_q$ en langue cible, alignés. On dispose également des fréquences de cooccurrence (fréquences observées, dorénavant) de chaque mot s_i avec chaque mot c_j dans le corpus : le nombre de fois où le mot si apparaît dans un segment s qui est aligné avec un segment c contenant c_j.

On suppose, par ailleurs, que chaque mot c_j d'un segment c est la traduction de l'un des mots s_i appartenant au segment s avec lequel il est aligné (dans cette expérience, un mot en langue source peut être aligné avec au plus un mot en langue source). On dit qu'un mot s_i est la source d'un mot c_j lorsque c_j est la traduction de s_i.

La donnée inconnue est donc, pour un couple (s, c), et pour chaque mot c_j appartenant à c, l'identité du mot s_i appartenant à s qui en est la source. On se sert d'EM pour déterminer, pour chaque couple (s_i, c_j) issu d'un couple (s, c) donné, et à partir de la fréquence observée de ce couple, la vraisemblance maximale pour que s_i soit la source de c_j dans ce couple (s, c).

Dans la suite, on appelle **matrice locale, notée n**, la matrice représentant les fréquences conditionnelles n_{ij} des couples (s_i, c_j) issus du n-ème couple de segments alignés (s, c). Les entrées des lignes de n sont les mots s_i du segment s, celles de ses colonnes sont les mots c_j de c, ses effectifs conjoints sont les probabilités d'alignement de chaque s_i avec chaque c_j. Les matrices locales sont construites à l'étape E. Et on appelle **matrice globale, notée M** (à ne pas confondre avec l'étape M d'EM, même si c'est à cette étape qu'elle est construite), la matrice représentant le maximum de vraisemblance des couples (s_i, c_j), noté $\mathbf{M}^{p}_{[ij]}$, dans le corpus. Les entrées des lignes de M les mots en langue source du corpus aligné (S), les entrées de ses colonnes sont les mots en langue cible (C).

a.2. Déroulement de l'algorithme

Le point de départ de l'algorithme est une matrice globale : une table de contingence F contenant les fréquences de cooccurrence, ou **fréquences observées**, notées f_{ij}, entre les mots s_i de S et les mots c_j de C. À chaque itération p, EM maximise la vraisemblance des couples (s_i, c_j), en estimant leur espérance au niveau local, en E, puis en maximisant la vraisemblance de leur espérance au niveau global, en M.

- À l'étape E, on calcule l'espérance de $F_p(n_{ij} \mid s_i, c)$, qu'on notera plus simplement n_{ij}, la fréquence conditionnelle des couples (s_i, c_j) pour chaque couple (s,c) dans lequel ils cooccurrent (ce sont les **fréquences estimées**). n_{ij} est donc la probabilité qu'un mot c_j d'un segment c soit la traduction d'un mot s_i d'un segment s, étant donnés s_i et l'ensemble (c) de ses traductions possibles.

- Puis, à l'étape M, on maximise la vraisemblance de l'espérance trouvée à l'étape E, au niveau global.

- On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi jusqu'à la satisfaction du critère de terminaison de l'algorithme.

On note n^p et $\mathbf{M}^p$, respectivement, les matrices locales et la matrice globale issues de l'itération p.

a.3. Terminaison de l'algorithme

L'une des propriétés de l'algorithme est de faire croître la vraisemblance à chaque itération : $\mathbf{M}^p{}_{[ij]} >= \mathbf{M}^{p+1}{}_{[ij]}$. La convergence d'un algorithme EM est définie par [Dempster et al. 1977] comme le fait que la vraisemblance ne croît plus, ou se met à croître de manière non significative. L'une des causes du succès de la méthode EM réside dans sa vitesse de convergence, une propriété théorique élémentaire des techniques d'apprentissage au sens de [Vapnik 1999]. Dans le cas où la vraisemblance est bornée, l'algorithme EM converge très vite dès les premières itérations. L'algorithme converge vers un point stationnaire de la vraisemblance, qui n'est pas nécessairement le maximum global.

Le but, pour nous, est d'atteindre le maximum local de vraisemblance, c'est-à-dire la meilleure probabilité distributionnelle possible pour chaque couple (s_i, c_j) issu d'un segment aligné(s,c). Il convient, donc, de s'assurer que l'algorithme itère autant de fois qu'il le faut pour atteindre son point stationnaire.

Pour cela, à la fin de chaque itération p, on évalue la variance maximale de la vraisemblance par rapport à l'itération p-1, soit la différence maximale entre les valeurs cellule à cellule (*maximum cell deviation*) des matrices globales $\mathbf{M}^{p-1}$ et $\mathbf{M}^p$. La variance maximale doit être inférieure à un seuil de convergence τ fixé. C'est le critère de convergence défini par [Dempster el al. 1977]. Pour approcher au plus le point stationnaire de la vraisemblance, on fixe une valeur très faible pour τ (par exemple 0.001). La valeur de τ affecte entre autres le nombre d'itérations effectuées : plus sa valeur est faible, et plus l'algorithme itère.

Le choix des valeurs initiales est un autre facteur pouvant conditionner le résultat de l'apprentissage, ainsi que le comportement des valeurs quant à leur convergence. [Hiemstra 1996] présente une section complète sur les multiples manières d'initialiser l'algorithme et en propose une nouvelle, que nous avons utilisée (voir la *section 6.3.3*).

b. L'algorithme IPFP

L'algorithme IPFP (*Iterative Proportional Fitting Procedure*), qui appartient à la classe des algorithmes de type EM, aurait été décrit pour la première fois par [Kruithof 1937] (une description

très complète de l'algorithme et de ses fondements mathématiques se trouve dans [Bishop & al. 1975]. Dans cet algorithme, chaque itération p est réalisée en deux étapes au cours desquelles l'algorithme recalcule les profils-lignes, puis les profils-colonnes des matrices locales. Ces deux étapes sont notées, $(p, 1)$, étape 1, et $(p, 2)$, étape 2 de l'itération p. Les matrices locales issues de chacune de ces étapes sont notées, respectivement, $n^{(p, 1)}$ et $n^{(p, 2)}$. La fréquence conditionnelle du couple (s_i, c_j) est notée $n_{ij}^{(p, 1)}$ puis $n_{ij}^{(p, 2)}$, avec :

$$n_{ij}^{(p,1)} = n_{ij}^{(p-1,\,2)} * (I_{i\bullet} \,/\, n_{i\bullet}^{(p-1,\,2)})$$
$$n_{ij}^{(p,\,2)} = n_{ij}^{(p,\,1)} * (I_{\bullet j} \,/\, n_{\bullet j}^{(p,\,1)})$$

La matrice I est la matrice initiale décrite par la suite (à la *section 6.3.3.b*). Lors de la première itération, $n_{ij}^{(p-1,\,2)}$ est I_{ij}.

Calcul des profils-lignes et des profils-colonnes

Les matrices locales des exemples sont celles du couple de segments alignés (s,c) suivant :

s	*Les votes sont clos.*
c	*The vote is closed.*

Soient I la matrice globale initiale de l'algorithme et la matrice locale $n^{(p-1,\,2)}$:

$n^{(p-1,\,2)}$	*the*	*vote*	*is*	*closed*	**sommes**
les					
votes		***A***			***B***
sont					
clos					
sommes					

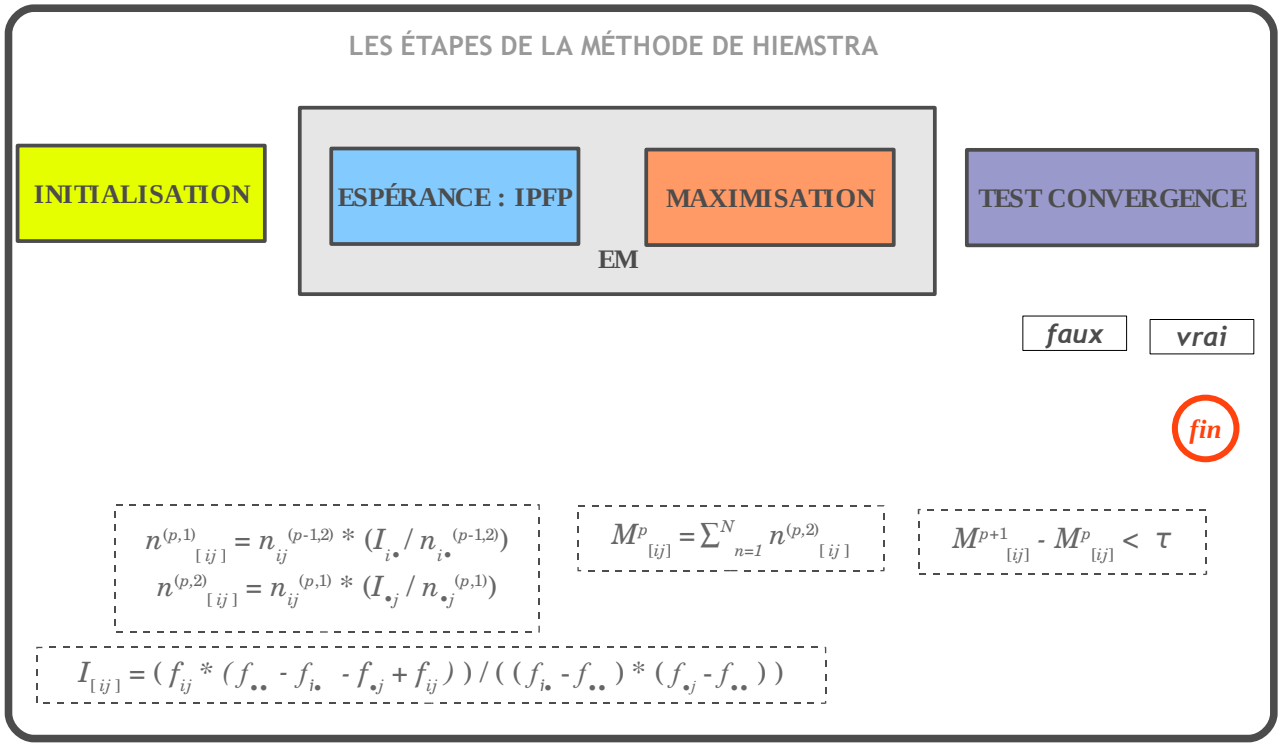
- $n_{ij}^{(p,\,1)} = C = A * (I_{i\bullet} \,/\, B)$

$n^{(p,\,1)}$	*the*	*vote*	*is*	*closed*	**sommes**
les					
votes		***C***			
sont					
clos					
sommes		***D***			

- et $n_{ij}^{(p,2)} = G = C * (I_{\bullet j} / D)$

$n^{(p,2)}$	*the*	*vote*	*is*	*closed*	**sommes**
les					
votes		***G***			
sont					
clos					
sommes					

5.3.3. La méthode de Hiemstra



- **Dessin 15** – *Les étapes de la méthode d'alignement de Hiemstra* -

a. Le point de départ de l'algorithme : la matrice *F* des fréquences observées

Le point de départ de l'algorithme est, comme nous l'avons dit précédemment, la matrice globale ***F*** contenant les **observées**, notées f_{ij}, entre les mots s_i de S et les mots c_j de C. Pour un mot s_i en LS et un mot c_j en LC, la valeur de f_{ij} est la fréquence d'occurrence du couple de mots (s_i, c_j), ou le nombre de fois où le mot c_j apparaît dans une phrase c en LC qui est alignée avec une phrase s en langue source contenant le mot s_i.

F	...	the	...	vote	...	is	closed	
...								
Les		$f_{les,the}$		$f_{les,vote}$		$f_{les,is}$	$f_{les,closed}$	
...								
votes		$f_{votes,the}$		$f_{votes,vote}$		$f_{votes,is}$	$f_{votes,closed}$	
...								
sont		$f_{sont,the}$		$f_{sont,vote}$		$f_{sont,is}$	$f_{sont,closed}$	
...								
clos		$f_{clos,the}$		$f_{clos,vote}$		$f_{clos,is}$	$f_{clos,closed}$	$f_{i\bullet}$
...								
...								
							$f_{\bullet j}$	$f_{\bullet\bullet}$

b. Initialisation de l'algorithme : la matrice globale *I*

Pour l'estimation des fréquences globales initiales (la matrice ***I***) de l'algorithme, Hiemstra propose la formule suivante :

$$I_{ij} = (f_{ij} * (f_{\bullet\bullet} - f_{i\bullet} - f_{\bullet j} + f_{ij})) / ((f_{i\bullet} - f_{\bullet\bullet}) * (f_{\bullet j} - f_{\bullet\bullet}))$$

I	...	the	...	vote	...	is	closed	
...								
Les		$I_{les,the}$		$I_{les,vote}$		$I_{les,is}$	$I_{les,closed}$	
...								
votes		$I_{votes,the}$		$I_{votes,vote}$		$I_{votes,is}$	$I_{votes,closed}$	
...								
sont		$I_{sont,the}$		$I_{sont,vote}$		$I_{sont,is}$	$I_{sont,closed}$	
...								
clos		$I_{clos,the}$		$I_{clos,vote}$		$I_{clos,is}$	$I_{clos,closed}$	$I_{i\bullet}$
...								
...								
							$I_{\bullet j}$	$I_{\bullet\bullet}$

Selon les tests effectués par Hiemstra sur différentes manières de définir les fréquences initiales, cette formule est la seule qui permet d'obtenir un alignement de qualité à la fois pour les mots pleins et pour les mots fonctionnels. Les autres formules testées ne sont satisfaisantes que pour les mots pleins. [Melamed 1998], lui, aligne séparément ces deux classes de mots, avec des paramètres différents.

c. Application de l'algorithme EM pour l'estimation des paramètres du modèle

Dans la méthode de Hiemstra, l'étape E de l'algorithme EM est remplacée par une itération de l'algorithme IPFP. L'étape M consiste ensuite à construire une nouvelle matrice globale dans laquelle la fréquence du couple (s_i, c_j) est la somme de ses fréquences dans les matrices locales $n^{(p, 2)}$ où il est représenté :

$$\mathbf{M}^{p}_{[ij]} = \sum^{N}_{n=1} n_{ij}^{(p, 2)}$$

d. Vérification du test de convergence

On calcule la différence cellule à cellule entre les matrices globales $\mathbf{M}^{p-1}$ et $\mathbf{M}^{p}$ (lors de la première itération, $\mathbf{M}^{p-1}$ est la matrice globale initiale I). Soit τ le seuil de convergence fixé :

- si la différence maximale est inférieure à τ, alors l'algorithme termine, et renvoie la matrice globale $\mathbf{M}^{p}$, construite lors de la dernière itération;
- sinon, il effectue une nouvelle itération d'IPFP.

5.3.4. Calcul des probabilités bidirectionnelles de traduction

À partir des valeurs globales des paramètres obtenus à l'aide de l'algorithme EM (représentées par la matrice $\mathbf{M}^{p}$ renvoyée), on peut calculer les probabilités de traduction des couples (s_i, c_j), de la langue source vers la langue cible, notées P_{ij}, et de la langue cible vers la langue source, P_{ji}.

$$P_{ij} = \mathbf{M}^{p}_{ij} \;/\; \mathbf{M}^{p}_{i\bullet} \qquad \text{et} \qquad P_{ji} = \mathbf{M}^{p}_{ij} \;/\; \mathbf{M}^{p}_{\bullet j}$$

5.3.5. L'algorithme *Competitive Linking*

L'algorithme *Competitive Linking* d'alignement proposé par [Melamed 1998] implémente l'heuristique selon laquelle un mot en LS ne peut être aligné qu'avec un seul mot en LC, et vice versa. Cet algorithme se base sur les probabilités de traduction.

Pour chaque couple de phrases (s,c) :

1. On calcule des scores d'alignement des couples de mots, au niveau global. Le score du couple (s_i, c_j) est :

   $$A_{ij} = P_{ij} * P_{ji}$$

2. La liste des couples de mots représentant (s,c) est triée par ordre décroissant des scores d'alignement.
3. Le couple de mots (s_i, c_j) ayant le score le plus élevé est aligné
4. et tous les couples de mots contenant s_i ou c_j sont supprimés de la liste
5. Les étapes (2) et (3) sont répétées jusqu'à ce que tous les termes soient alignés.

5.3.6. Les étapes de notre méthode d'alignement

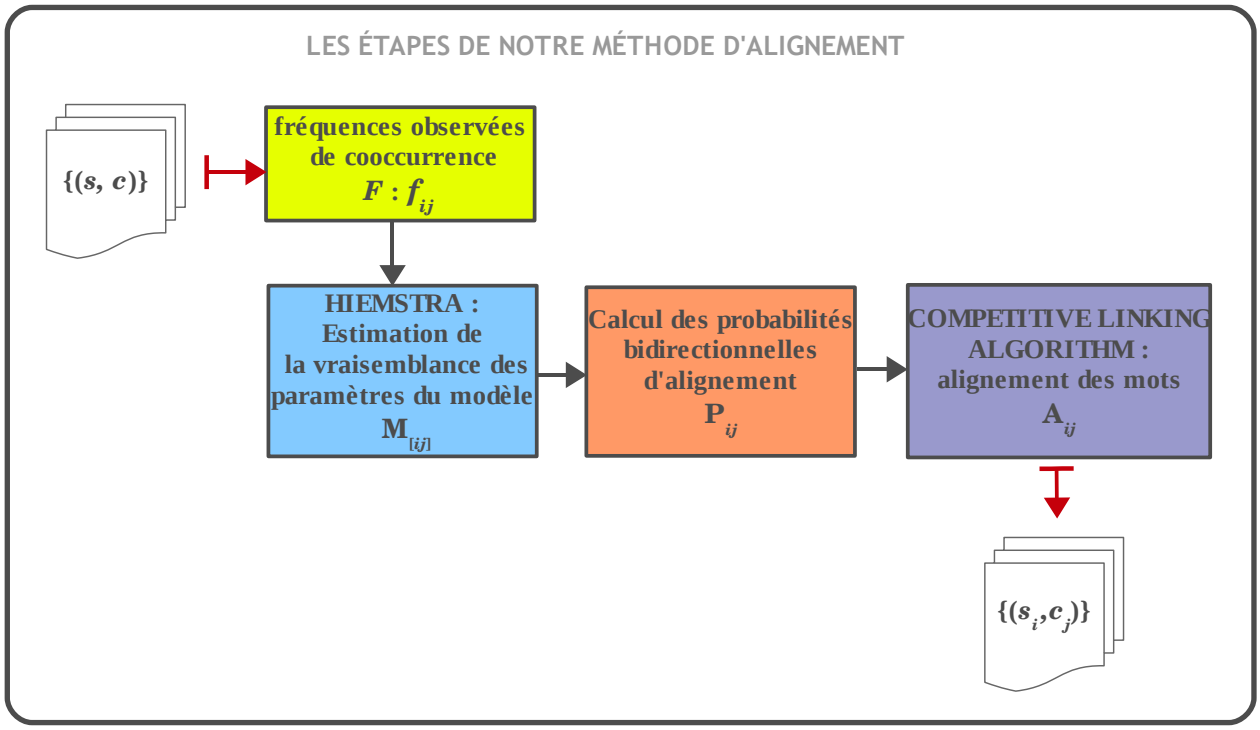
La figure 16 suivante décrit les trois étapes de notre méthode d'alignement :

1. la méthode de [Hiemstra 1996] est utilisée pour estimer le maximum de vraisemblance des paramètres du modèle. La méthode d'initialisation de l'algorithme EM selon [Hiemstra 1996] nous permet d'aligner tous les mots du corpus (les mots pleins et les mots fonctionnels) en une passe avec un résultat de même qualité pour les deux catégories.

2. on calcule les probabilités d'alignement de la LS vers la LC, et inversement.

3. les mots en LC sont alignés avec leur source par l'algorithme *Competitive Linking Algorithm* de [Melamed 1998].

Cette méthode d'alignement peut être utilisée pour d'autres applications, telles que la construction automatique de dictionnaires bilingues ou l'extraction automatique de termes spécifiques.



- Dessin 16 – *Les étapes de notre méthode d'alignement* -

Le prétraitement du corpus à aligner

Le corpus à aligner est constitué des versions française (la LS) et anglaise (la LC) du corpus Europarl, décrit au début de ce rapport (*section 2.2.1*). Avant l'application de la méthode d'alignement, nous avons procédé au marquage des expressions multi-termes des deux langues à l'aide de l'outil Unitex (lors de l'alignement, chacune de ces expressions multi-termes est considérée comme une seul unité). Nous avons également procédé à l'étiquetage morpho-syntaxique du corpus avec l'outil TreeTagger. L'alignement porte ensuite sur des types de la forme *lemme#étiquette_grammaticale*.

5.4. Évaluation de l'alignement

Nous présentons dans cette section une évaluation partielle de notre méthode d'alignement : notre algorithme a été entraîné sur 50 fichiers (sur les 600 fichiers du corpus), et seuls 30 fichiers ont été alignés mot à mot. À la section (6.4.1), la liste des traductions relevées manuellement (première colonne du tableau suivant) et la liste des alignements corrects (deuxième colonne du tableau) ne sont donc pas entièrement comparables.

Dans la colonne « Traductions relevées manuellement » :

- Les mots en turquoise sont les mots que nous avons relevés manuellement.et qui ont été alignés avec le mot concerné.
- Les mots en violet sont des mots alignés avec le mot concerné et qui sont des sous-unités de mots composés qui en sont la traduction.

Dans la colonne « Alignements corrects », les mots en rouge sont les alignements corrects que nous n'avons pas relevés manuellement.

Les erreurs que nous observons sont de deux types :

- Le mot est aligné avec un mot de la phrase en LC qui n'est pas sa traduction. La cause de l'erreur est l'algorithme d'alignement utilisé.

- Le mot est aligné avec une sous-unité d'un mot composé de la LC qui est sa traduction. Cette fois encore, la cause de l'erreur est l'algorithme d'alignement, qui ne permet pas les alignements multiples : un mot en LS avec plusieurs mots en LC ou, inversement, plusieurs mots en LS avec un mot en LC.

5.4.1. Évaluation qualitative de l'alignement des mots pleins

Traduction relevées manuellement	Alignements corrects	Nombre d'alignements corrects / nombre d'alignements	Alignements incertains
produit			
foods; *chemicals*; ***product***; *result*; ***produce***; *revenue*; ***good***; *originate*; *take* ***place***; *happened*	*Product*; *good*; *produce*; ***substance***; ***financial products***; *food*; ***crop***; ***agricultural products***; ***dairy products***; ***foodstuff***; ***animal feed***;	11 / 211	*Nutrition*; *cosmetic*; *duty*; *dangerous*; *substitute*; *animal*; *vegetable*; *compensation*; etc. (~ *product*)
cadre			
Within *the* ***scope*** *of*; *as part of*; ***framework***; *frame*; ***context***	*Framework*; *context*	4 / 450	*Scope*; *within* (*within the scope of*)

Traduction relevées manuellement	Alignements corrects	Nombre d'alignements corrects / nombre d'alignements	Alignements incertains
compte			
take into account ; *take account of*; *consider*; *be **aware** of*; *give **consideration** to*; *take into **consideration***; *realise*; *in the end*; *ultimately*; *__account__*; *be **accountable***; *accountability*; *__court of auditors__*	*Account*; *court of auditors*	5 / 327	*Auditor* (*court of auditors*); *aware* (*be aware*); *consideration* (*take into consideration*); *give* (*give consideration to*); *take* (*take into consideration*); *accountable* (*be accountable*)
saisir			
Grab; *__grasp__*; *__seize__*; *confiscate*; *grip*; *appeal*; *take*; *take up*; *welcome*; *propose*; *have recourse to*; *__sue__*	*Grasp*; *seize*; *__refer__*	3 / 36	

5.4.2. Évaluation qualitative de l'alignement des mots fonctionnels

Alignements corrects	Nombre d'alignements corrects / nombre d'alignements	Alignements incertains
le/la		
Him; *her*; *it*; *the*	4 / 1293	
vous		
You	1 / 410	*Yourself* (*vous-même*)
que		
Than; *that*	4 / 1060	
pour		
For; *to*	2 / 2060	*As*; *far* (*as far as*); purpose (*for the purpose of*); order (*in order to*); *per*; *cent* (*per cent*)
en		
In; *into*; *any*; *of*; *at*; *to*; *on*	6 / 2598	
alors		
Then; *so*; *while*; *when*	4 / 359	*Even*; *though* (*even though*)

5.5. Conclusion

Nous avons développée cette méthode d'alignement afin d'aligner mot à mot les versions française et anglaise du corpus EuroParl. Ceci doit nous permettre de construire automatiquement le lexique bilingue à partir duquel sont extraits d'EuroParl les sous-corpus des mots. Cette tâche, ainsi que l'évaluation du résultat de l'alignement n'ont pas encore été réalisés car l'alignement est en cours.

6. Conclusion

Nous avons comparé deux méthodes d'apprentissage dans le cadre de la désambiguïsation sémantique. Pour cela, nous avons sélectionné une vingtaine de mots fortement polysémiques et nous avons construit un sous-corpus d'exemples pour chacun de ces mots. Nous avons ensuite évalué et comparé les performances de ces deux méthodes. Puis, nous avons effectué une série de tests pour mesurer l'impact de différents divers types d'informations linguistiques sur les performances des classifieurs. Le but était de trouver la combinaison optimale de ces informations dans le cadre de la méthode de désambiguïsation que nous avons développée : nous en avons découvert douze.

Les résultats que nous avons obtenus sont satisfaisants, mais pourraient encore être améliorés de différentes manières :

- La qualité des classes de traductions représentant les différents usages des mots ambigus n'est pas satisfaisante. Ce problème pourrait être résolu par la prise en compte des cooccurrences du second ordre pour réaliser l'étape de désambiguïsation sémantique.
- D'autre part, des prétraitements plus précis du corpus d'apprentissage pourraient nous permettre d'améliorer les performances du module de désambiguïsation sémantique (et par là, celui de la sélection lexicale). Nous avons vu que lorsque les mots composés sont considérés en tant que tels, le résultat de la désambiguïsation est meilleur. Or, tous les types de mots composés n'ont pas été pris en compte dans cette expérience. Un traitement de ce type d'expressions plus sophistiqué (expressions multi-termes non-connexes, etc.) et de plus large couverture (noms, verbes, etc.) ne pourrait donc être que bénéfique pour la désambiguïsation.
- Du point de vue de la précision, l'approche « traduction multi-source » nous semble également être une piste intéressante qui pourrait être appliquée, en construisant un lexique multilingue, par exemple.
- La sélection lexicale, par ailleurs, pourrait être améliorée : on pourrait y intégrer des probabilités bidirectionnelles de traduction calculées à partir du corpus aligné mot à mot.
- Enfin, nous projetons d'élargir les sous-corpus des mots à l'aide d'un algorithme d'apprentissage.

Le passage au stade supérieur, à savoir la mise au point d'un véritable système de traduction automatique, nécessite un corpus aligné mot à mot. Nous avons, donc, développé une méthode d'alignement mot à mot que nous avons appliquée à notre corpus d'apprentissage. À partir de ce corpus, nous pourrons construire automatiquement un lexique bilingue à plus grande échelle. Auparavant, nous pourrions améliorer la qualité de notre méthode d'alignement, à au moins deux niveaux :

- La méthode d'initialisation de Hiemstra, implémentée mais non utilisée par manque de temps (elle ralentit considérablement la convergence des données), permet d'améliorer l'alignement des termes fonctionnels.
- Et une méthode plus sophistiquée de réduction des données, telle que l'AFC (Analyse Factorielle des Correspondances), devrait remplacer la méthode rudimentaire que nous avons appliquée.

Nous aimerions, enfin, appliquer notre méthode pour la traduction dans d'autres paires de langues que le français et l'anglais.

7. Annexes

7.1. Les vocables étudiés, leurs usages et leurs traductions relevées dans le corpus aligné

7.1.1. Les noms

Usages	Exemples/synonymes	TRADUCTIONS en LC
Article		
Partie d'une loi, d'une convention	L'article 125 du Code civil	*Rule; clause*
Texte formant un tout distinct	Article de journal	*Press report; article*
Marchandise		*article*
Barrage		
Ouvrage hydraulique	Barrage de retenue	*Dam; reservoir; flood-control dam; flood prevention barrier; flood defence; barrage; boom*
Loc. *faire barrage*		*Exclusion; stand in the way of; halt; stop; obstruct; oppose; stem; contrast; nip; fight; prevent; obstacle ; create a barrier to; block the rise of; lid*
Ce qui barre (abstrait)	Barrage de reproches	*Threshold clause; exclusion*
Ce qui barre (concret, momentané)	Barrage de rue	*Scatter-gun approach; street blockade; road blockade; road block; blockade; blockage; barrier; barrage; barricade; block(ing)*
Ce qui barre (concret, permanent)	Barrage de frontière	*Military barricade; checkpoint*
Cadre		
environnement	Contexte, milieu	*Frame; framework; context*
Loc. *dans le cadre de*		*Within the scope of; as part of*
Compte		
Fonds déposé dans un établissement financier	Compte épargne	*account*
Loc. *en fin de compte*	finalement	*In the end; ultimately*
Se rendre compte de, que	S'apercevoir de	*realise*
Rendre des comptes	Se justifier	*Be accountable*
- Tenir compte de - compte tenu de	Prendre en considération	*- Take into consideration; give consideration to; be aware of; consider; take account of; take into account* *- because; considering*

Usages	Exemples /synonymes	TRADUCTIONS en LC
E.N. Cour des Comptes		*Court of Auditors*
Conclusion		
Fin d'un discours	péroraison	*To conclude; conclusion; completion; outcome*
Action de conclure, accord final	Signature d'un accord	*conclusion*
Conséquence tirée d'un raisonnement		*Findings; outcome*
Loc. *En conclusion*	Ainsi; pour conclure	*In conclusion*
Conseil		
avis	Suivre un conseil	*advice*
Assemblée	Le Conseil constitutionnel	*Board; council*
Personne dont on prend avis	Avocat conseil	*consultancy*
Culture		
Travail de la terre	Culture de la soie	*Cultivation; agriculture*
Terres cultivées		*crop*
Ensemble des normes et activités propres à un groupe social	La culture française	*Culture; civilisation; identity*
Matière		
Loc. *en matière de*	En ce qui concerne	*In relation to; as regards; in terms of; in the area of*
Matières premières	Éléments bruts	*Raw materials*
Ce sur quoi on écrit, parle, travaille	Sujet; thème	*Matter; issue*
Passage		
Action, fait de passer dans un endroit	Le passage du train	*Visit to; to dock (bateau); running; flight; call at (bateau); passing; through-route;*
Moment où l'on passe à un endroit	Au passage du cortège	*Date; moment*
Fait de passer un court moment dans un endroit	Lors de mon passage à Londres	*Visiting; crossing point; come (v); time; visit*
Loc. passage en revue; action d'étudier	Passage en commission	*Survey; pass through; passage*
Condition indispensable à la suite d'un processus	Un passage obligé	*Port of call; route for (n)*
Loc. a*u passage*	En même temps	*by/on/along the way; incidentally; collateral; as an aside; in passing*

Usages	Exemples /synonymes	TRADUCTIONS en LC
fait de se rendre d'un endroit à un autre	Passage de frontières	*Cross; crossing; bridging*
fait de passer d'un degré à un autre	Passage en terminale	*Admission; transition*
Fait de passer d'un état à un autre	Passage à l'euro	*Change(-)over; increase; transition; move to; move on; move from-to; move into; movement; evolution; passage; switch; shift; transfer; conversion*
Fam. *passage à tabac*		
Court moment	Passage à vide	*Referral; period*
Fait de passer qqc. à qqn.	Passage du témoin	
Endroit par où l'on passe	Encombrer le passage	*route; crossing; passage*
Fragment d'une oeuvre orale ou écrite	Citer un court passage	*Area; section; paragraph; part; passage; point; wording; approach; statement*
Fait de se terminer	Passage du millénaire	*Turn of*
Produit		
Ce que rapporte une terre, une activité		*Revenue; produce; results*
Bien ou service		*Foods; chemicals; products; goods*
Pp. du verbe *produire*		*Materialise; originate; take place; occure; produce; happen*
Raison		
Loc. e*n raison de*		*Given; because (of); due to; in view (of)*
Sujet, cause, motif		*Reason; justification; factor*
Rapport		
Loc. Prép. *par rapport à*		*Regarding; with regard to*
Compte-rendu, exposé; témoignage, récit		*report*
Relation entre des personnes, groupes, États		*relation*
Réserve		
Quantité de choses accumulées pour être utilisées en cas de besoin		*Reserve; provision; funds*
Discrétion, retenue, prudence		*reservation*
Loc. *sans réserve*		*Unreservedly; unqualified; totally; without any reservations*

Usages	Exemples /synonymes	TRADUCTIONS en LC
Société		
Ensemble d'individus unis au sein d'un même groupe par des institutions, une culture, etc.		*Society*
La société civile		*Civil society*
Personne morale issue d'un contrat de société	entreprise	*Society; company; corporation*
Traitement		
Comportement, manière d'agir envers qqn.		*Treatment ; approach*
Manière de traiter une question, un problème		*Treatment ; approach ; processing*
Vol		
délit		*Theft; robbery*
Déplacement dans l'air		*Flight; flying*
Loc. *de haut vol*		*Hight flying; high level*
Loc. *à vol d'oiseau*		*As the crow flies*

7.1.2. Les verbes

Usages	Exemples /synonymes	TRADUCTIONS en LC
Lever		
Déplacer de bas en haut, redresser vers le haut		*Raise; rise; lift; get out*
Retirer, enlever		*Remove; get out*
Mettre fin à	clore	*End, put an end to, lift, raise,*
Gonfler, en parlant de la pate à fermentation		*Rise;*
Monter		
Se transporter dans un lieu plus haut	s'élever	*Climb; keep rising*
passer à un degré supérieur	augmenter	*Increase; boost; mount; blow up*
Prendre place dans un véhicule, sur une monture		*Board; go on board; come on board*

Usages	Exemples /synonymes	TRADUCTIONS en LC
Ajuster, assembler différentes parties pour former un tout		*Mount; assemble; set up; build up*
Porter		
Soutenir, maintenir, soulever un poids		*carry*
Avoir sur soi		*Wear; carry*
Prendre avec soi et mettre en un lieu déterminé		*Bear on*
Avoir pour objet, pour but		*Relate to;*
Amener, pousser à un degré supérieur		*Increase;*
Saisir		
Prendre, attraper		*Grab, grasp; grip; take; take up*
Opérer la saisie de	confisquer	*Grab; seize; confiscate*
Comprendre, sentir		*grasp*
Mettre immédiatement à profit		*welcome*
Avoir recours à, faire appel à		*Have recourse to; propose; sue; appeal*

7.2. Statistiques sur les vocables étudiés

7.2.1. Les noms

Vocable	Représentation des équivalents en LC (en nombre de segments de contexte)	Corpus	
		entraînement	test
article	*Article* (1000); *rule* (1000); *clause* (53); *press report* (7)	1647	413
barrage	*Large dam project* (1); *threshold clause* (1); *scatter-gun approach* (1); *reservoirs* (1); *road blockade* (1); *road block* (1); *flood prevention barrier* (1); *flood defence* (1); *military barricade* (1); *dam* (46); *exclusion* (1); *blockades* (6); *blockages* (1); *roadblocks* (2); *barrier* (2); *barrage* (4); *barriccade* (2); *block* (8); *boom* (1); *checkpoint* (2); *stand in the way of* (1); *halt* (2); *stop* (3); *obstruct* (1); *oppose* (1); *obstacle* (1); *stem* (1); *contrast* (1); *nip* (1); *fight* (3); *prevent* (5); *create a barrier to* (1); *lid* (1)	72	21
cadre	*Within the scope of* (52); *as part of* (330); *framework* (1000); *frame* (51); *context* (1000)	1945	488
compte	*take into account* (154); *take account of* (281); *consider* (56); *be aware of* (7); *give consideration to* (2); *take into consideration* (32); *realise* (4); *in the end* (91); *ultimately* (226); *account* (1000); *be accountable* (13); *accountability* (27); *court of auditors* (1000);	2308	385
conclusion	*in conclusion* (430); *conclusion* (1000); *completion* (31); *conclude* (459); *findings* (128); *outcome* (94)	1712	430
conseil	*Advice* (482); *board* (537); *consultancy* (11); *council* (1000)	1622	408
culture	*Agriculture* (*2); culture* (1000); *crop* (427); *civilisation* (43); *cultivation* (153); *identity* (80)	1362	343
matière	*in relation to* (110); *as regards* (205); *in terms of* (259); *raw materials* (29); *matters* (749); *in the area of* (179); *issue* (624)	1723	432
passage	*At the same time* (5); *as an aside* (1); *by the way* (3); *in passing* (7); *along the way* (2); *incidentally* (12); *collateral* (1); *changeover* (107); *passage* (91); *area* (56); *section* (21); *route* (18); *paragraph* (18); *visit* (14); *period* (32); *step* (26); *hand over* (4); *transition* (109); *cross* (43); *evolution* (3); *point* (82); *move* (82); *date* (10); *swith* (34); *wording* (25); *statement* (13); *change* (37); *way* (33); *shift* (19); *transit* (8); *component* (2); *condition* (18); *clause* (5); *tunnel* (3); *part* (42); *transfer* (12)	800	198
produit	*Foods* (230); *chemicals* (231); *product* (1000); *result* (297); *produce* (490); *revenue* (11); *good* (659); *originate* (52); *take place* (55); *happened* (68)	2471	942

Vocable	Représentation des équivalents en LC (en nombre de segments de contexte)	Corpus	
		entraînement	test
raison	*because of* (688); *due to* (184); *in view of* (29); *reason* (1000); *justification* (66); *factor* (114)	1663	418
rapport	*report* (1000); *relation* (1000); *with regard to* (69)	1655	414
réserve	*Unreservedly* (63); *unqualified* (2); *totally* (9); *without any reservations* (1); *reserve* (337); *reservation* (62); *fund* (38); *provision* (13)	417	108
société	*society* (1000); *corporation* (3); *company* (274)	1021	256
traitement	*Approach* (72); *processing* (166); *treatment* (1000)	989	249
vol	*flight time* (9); *international flight* (4); *short flight* (2); *long haul flight* (1); *high level* (1); *theft* (60); *robbery* (19); *flight* (429); *flying* (37)	412	150

7.2.2. Les verbes

Vocable	Représentation des équivalents en LC (en nombre de segments de contexte)	Corpus	
		entraînement	test
lever	*Raise*(33); *rise*(34); *lift*(213); *end*(29); *remove*(86)	314	81
monter	*Go on board* (1); *board* (9); *come on board* (2); *assemble* (1); *mount* (7); *set up* (8); *build up* (1); *increase* (36); *blow up* (3); *boost* (1)	50	14
porter	*Bear on* (5); *carry* (174); *increase* (202); *relate to* (237); *wear* (78)	555	141
saisir	*Grab* («); *grasp* (56); *seize* (192); *confiscate* (8); *grip* (10); *appeal* (18); *take* (514); *take up* (13); *welcome* (25); *propose* (36); *have recourse to* (3); *sue* (4)	691	178

7.3. KNN : Manuel d'utilisation

7.3.1. Utilisation du package KNN

Le package KNN est l'implémentation en Java de la méthode de désambiguïsation sémantique présentée à la section (4.2.3) de ce rapport.

a. La commande de lancement du programme

```
java -jar KNN.jar [arguments]
```

b. Les arguments de la ligne de commande

--word-space-source | -ws
fichier1 : la partie des sous-corpus source des mots à utiliser pour l'entraînement du module de classification; un fichier aux format et type décrits dans la sous-section (8.3.2) suivante

--test-source | v
fichier2 : la partie des sous-corpus source des mots à utiliser pour les tests du module de classification; un fichier aux format et type décrits dans la sous-section (8.3.2) suivante

--word-space-cible | -wc
fichier3 : la partie des sous-corpus cible des mots à utiliser pour l'entraînement du module de sélection lexicale; un fichier aux format et type décrits dans la sous-section (8.3.2) suivante

--test-cible | -tc
fichier4 : la partie des sous-corpus cible des mots à utiliser pour les test du module de sélection lexicale; un fichier aux format et type décrits dans la sous-section (8.3.2) suivante

--ls-test | -ls
fichier5 : les phrases en LS correspondant aux instances de -ts
une phrase par ligne, l'ordre des phrases correspond à l'ordre des instances de test dans -ts

--lc-test | -lc
fichier6 : les phrases en LC correspondant aux instances de -tc
une phrase par ligne, l'ordre des phrases correspond à l'ordre des instances de test dans -tc

--dictionnaire | -d
fichier7 : le lexique bilingue; un fichier au format décrit à la sous-section (8.3.3) suivante.

--contexte | -c
le type des fichiers 1 à 4 précédents
1 les fichiers sont du type 1 décrit à la section (8.3.2.a) suivante.
2 les fichiers sont du type 2 décrit à la section (8.3.2.b) suivante.

--reduction | -r
option de réduction des traits : les traits non significatifs (associés au moins une fois à chacune des classes de l'espace sémantique d'un mot) ne sont pas pris en compte dans les calculs de similarité entre classes

0 pas de reduction
1 les traits non significatifs ne sont pas pris en compte

--ptt | -p
les informations prises en compte pour calculer les similarités entre classes
1 tous les traits sont pris en compte
2 seuls les traits forts (voisins ou cooccurrents syntaxiques) sont pris en compte

--print | -v
mode d'affichage
0 aucun affichage
1 afficher les performances des modules de désambiguïsation et de sélection lexicale
2 afficher les usages (classes de traductions) découverts par le module de désambiguïsation
3 afficher les phrases de test en LS et en LC avec l'usage (classe de traductions) et la traduction qui lui ont été assignés, respectivement, par le module de désambiguïsation et le module de sélection lexicale
4 afficher les informations des options 1 à 3 avec le déroulement du programme

7.3.2. Format des fichiers d'entraînement et de test

Dans les fichiers d'entraînement et de test, le format d'une ligne peut être de deux types. Les indices sont des entiers, les poids peuvent être des entiers ou des nombres décimaux. Les traits sont triés par ordre croissant de leur indice. Le séparateur est l'espace.

classe est l'indice de la classe à laquelle est associé le vecteur de traits représenté à la ligne courante
indiceTraitX les cooccurrents du mot sont représentés par leur indice
poidsTraitX le poids de base est 1, les traits dont le poids est supérieur au poids de base (appelés traits forts) peuvent correspondre aux voisins du mot ou à ses cooccurrents syntaxiques directs.
PosTraitX la position du trait concerné par rapport au mot, dans le cas où l'on veut représenter une fenêtre d'occurrence du mot de taille *t*. Le vecteur de trait représenté par une ligne, de dimension *t**2, est une suite <-*t* -*t*-1 -*t*-2 ... -*t*-(*t*-1) 1 2 ... *t*>.

a. Type 1

<*classe*> [<*indiceTrait1*:*poidsTrait1*> <*indiceTrait2*:*poidsTrait2*> ... <*indiceTraitN*:*poidsTraitN*>]

b. Type 2

<*classe*> [<*posTrait1*:*indiceTrait1*> <*posTrait2*:*indiceTrait2*> ... <*posTraitN*:*indiceTraitN*>]

7.3.3. Format du fichier contenant le lexique bilingue

Les lignes du fichier contenant le lexique bilingue sont au format suivant :

<vocable>#<forme_LS>#<forme_LC>

avec :

vocable	un mot ambigu
forme_LS	une forme en LS du mot ambigu 'vocable' relevée dans un segment en LS du corpus aligné
forme_LC	la forme en LC du mot ambigu 'vocable' relevée dans le segment en LC du corpus aligné correspondant au segment en LS précédent

7.4. WordAlign : manuel d'utilisation

Le package WordAlign est l'implémentation en Python de la méthode d'alignement décrite à la section (6.3) de ce rapport.

a. La commande de lancement du programme

```
python Main.py [arguments]
```

b. Les arguments de la ligne de commande

--sl

Le corpus source étiqueté morpho-syntaxiquement (une ligne est au format '<*mot*>\t<*étiquette*>\t<*lemme*>')

--tl

Le corpus cible étiqueté morpho-syntaxiquement (une ligne est au format '<*mot*>\t<*étiquette*>\t<*lemme*>')

--exec

Mode de lancement du script (voir l'option -i suivante).

1 entraînement de l'algorithme sur les fichiers du corpus, un par un
2 calcul des probabilités bidirectionnelles de traduction
3 alignement des phrases

-i

L'entraînement de l'algorithme se fait fichier par fichier. Le programme peut être arrêté en cours d'entraînement, l'option courante permet de reprendre l'entraînement au fichier où il a été arrêté (les noms des fichiers 'appris' sont enregistrés dans le fichier /counts/tmplog.txt).

0 reprendre l'entraînement là où il a été arrêté
1 commencer l'entraînement au premier fichier du corpus

--words

1 le programme commence par construire les dictionnaires des mots des corpus source et cible (qui servent ensuite à représenter les couples de mots par des couple d'indices)
0 les dictionnaires des mots sont déjà construits. Ils se trouvent dans les fichiers /counts/slwords.txt (corpus source) et /counts/tlwords.txt (corpus cible)

--form

Le forme des éléments alignés.

1 les mots eux-mêmes
2 leurs lemmes

--converge

Le seuil de convergence.

--converge

Le type d'initialisation de l'algorithme EM.

1 initialisation par la matrice des fréquences observées de cooccurrence
2 initialisation par la matrice de l'option 1 transformée à l'aide de la formule de Hiemstra

c. Exemple

```
python Main.py --sl ~/europarl/aligned/fr-en/fr --tl ~/europarl/aligned/fr-en/en --exec 1 -i 0
--words 0 --form 2 --converge 0.01 --init 2
```

entraînement de l'algorithme (--exec 1), en reprenant l'entraînement au fichier où l'on s'est arrêté auparavant (-i 0), les dictionnaires des mots sont déjà enregistrés (--words 0), les mots sont de la forme 'lemme#étiquette' (--form 2), les itérations de l'algorithme IPFP s'arrêtent lorsque la différence maximale cellule à cellule entre les matrices globales de deux itérations consécutives est inférieure ou égale à 0.01 (--converge 0.01). Le mode d'initialisation utilisé est celui de Hiemstra.

8. Bibliographie

[Apidianaki 2008a]
M. Apidianaki. (2008). Translation-oriented Word Sense Induction based on Parallel Corpora. In *Proceedings of the 6th International Conference on Language ressources and Evaluation (LREC)*, p. 3269-3275, Marrakech, Maroc.

[Apidianaki 2008b]
M. Apidianaki. (2008). *Automatic induction for Word Sense Disambiguation in translation*. Présentation pour le Séminaire du CENTAL.

[Apidianaki 2009a]
M. Apidianaki. (2009). Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference on European Chapter of the ACL (EACL)*, pp. 77-85, Athènes, Grèce.

[Apidianaki 2009b]
M. Apidianaki. (2009). La place de la désambiguïsation lexicale dans la Traduction Automatique Statistique. TALN'09 (à paraître).

[Apidianaki 2006]
M. Apidianaki. (2006). Traitement de la polysémie lexicale dans un but de traduction. In *Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles* (TALN 2006), P. Mertens, C. Fairon, A. Dister et W. Patrick (éds.), Leuven, Belgique, 10-13 avril, 1:53-62.

[Baum 1972]
L. E. Baum. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. In *Inequalities*, 3:1-8.

[Besançon et Rajman 2002]
R. Besançon et M. Rajman. (2002). Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents. *Actes de TALN 2002*, Nancy, France, 24-27 juin.

[Bishop & al. 1975]
Y. M. M. Bishop, S. E. Fienberg, et P. W. Holland. (1975). *Discrete multivariate analysis : theory and practice*. MIT Press.

[Brown et al. 1993]
P. F. Brown, S. A. Pietra, V. J. D. Pietra & R. L. Mercer. (1993). The mathematics of machine translation : parameter estimation. In *Computational Linguistics*, 19(2).

[Brown, Lai et Mercer 1991]
P. F. Brown, J. C. Lai & R. L. Mercer. (1991). Aligning sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, pp. 169-176, juin.

[Cabezas et Resnik 2005]
C. Cabezas & P. Resnik. (2005). *Using WSD Techniques for Lexical Selection in Statistical Machine Translation.* Rapport interne LAMP-TR-124, CS-TR-4736, UMIACS-TR-2005-42, University of Maryland, College Park.

[Callison-Burch et al. 2008]
C. Callison-Burch, C. S. Fordyce, P. Koehn, C. Monz & J. Schroeder. (2008). Further meta-evaluation of machine translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pp. 70-106, Columbus, Ohia.

[Carl 2003]
M. Carl. (2003). *Introduction à la traduction guidée par l'exemple (traduction par analogie)*. Tutoriel, TALN'03, Batz-sur-mer.

[Carpuat et Wu 2007]
M. Carpuat & D. Wu. (2007). Word sense disambiguation vs. statistical machine translation. In *Proceedings of 43rd Annual Meeting of the ACL*, pp. 387-394, Ann Arbor, Michigan.

[Chiang 2005]
D. Chiang. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pp. 263-270, Ann Arbor, Michigan.

[Church 1993]
K. W. Church. (1993). A program for aligning parallel texts at the character level. In *Proceedings of ACL-93*, Columbus OH.

[Crego et al. 2009]
J. M. Crego, A. Max et F. Yvon. (2009). Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multi-source par renforcement lexical. *TALN'09*, 24-26 juin, Senlis, France (à paraître).

[Debili et Sammouda 1992]
F. Debili & E. Sammouda. (1992). Appariement de phrases de textes bilingues français-anglais et français-arabes. In *Actes de COLING-92*, Nantes, pp. 524-528.

[Debili et Zribi 1996]
F. Debili et A. Zribi. (1996). Les dépendances syntaxiques au service de l'appariemment des mots. *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA '96)*.

[Dempster et al. 1977]
A.P. Dempster, N.M. Laird et D.B. Rubin. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39-1:1-38.

[Firth 1957]
J. Firth. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, Philological Society, Oxford. Reprinted in F. Palmer. (1969). *Selected Papers of J. R. Firth, 1952-59*, pp. 168-205, London: Longmans.

[Fox 2002]
H. J. Fox. (2002). Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP-02*, pp. 304-311.

[Gale et Church 1991]
W. A. Gale & K. W. Church. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of ACL-91*, Berkeley CA

[Gaussier 2005]
E. Gaussier. (1995). Contributions à l'accès à l'information documentaire. HDR thesis, Université Joseph Fourier, Grenoble, France.

[Gaussier, Hull et Aït-Mokhtar 2000]
E. Gaussier, D. Hull & S. Aït-Mokhtar. (2000). Term alignment in use : machine-aided human translation. Parallel text processing. In J. Veronis (ed.). *Parallel Text processing. Alignment and use of Translation Corpora*. Dordrecht : Kluwer Academic Publishers, pp. 253-274.

[Grefenstette 1994]
G. Grefenstette. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston/Dordrecht/London: Kluwer Academic Publishers.

[Habert et al. 1997]
B. Habert, A. Nazarenko & A. Salem. (1997). *Les linguistiques de corpus*, Armand Colin/Masson, Paris.

[Harris 1988]
Z. Harris. (1988). Are you bi-textual ? *Language Technology*, 7:41-45

[Harris 1968]
Z. Harris. (1968). *Mathematical structures of language*. New York: Interscience Publishers.

[Haruno et Yamazaki 1997]
M. Haruno & T. Yamazaki. (1997). High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa-Cruz, California, pp. 131-138.

[Hiemstra 1996]
D. Hiemstra. (1996). *Using statistical methods to create a bilingual dictionary.* Master's thesis, University of Twente.

[Hutchins 2007]
W. J. Hutchins. (2007). Machine translation : a concise history. In C. S. Wai (ed.). (2007). *Computer aided translation : theory and practice*. China : Chinese University of Hong Kong.

[Johansson, Ebeling et Hofland, 1996]
S. Johansson, J. Ebeling & K. Hofland. (1996). Coding and aligning the English-Norwegian Parallel Corpus. In M. Kay & M. Röscheisen. (1993). Text-translation alignment. *Computational Linguistics,* 19(1):121-142.

[Kay et Röscheisen 1993]
M. Kay & M. Röscheisen. (1993). Text -Translation Alignment. In *Computational Linguistics*, 19(1):121-142.

[Kilgarriff 1997]
A. Kilgarriff. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135-155.

[Koehn 2004]
P. Koehn. (2004). Pharaoh : A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 115-124, Washington, DC.

[Koehn 2003]
P. Koehn. (2005). *Europarl : A parallel corpus for Statistical Machine Translation*. MT Summit.

[Koehn et al. 2003]
P. Koehn, F. J. Och & D. Marcu. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology and North American ACL Conference (HLT/NAACL)*, pp. 48-54, Edmonton, Canada.

[Kruithof 1937]
J. Kruithof. (1937). Calculation of telephone trafic. *De Ingenieur*, 52:E15-E25.

[Langlais et El-Beze 1997]
P. Langlais & M. El-Bèze. (1997). Alignement de corpus bilingues : algorithmes et évaluation. *1ères JST FRANCIL (Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF)*, Avignon, pp. 191-197.

[Larousse 2009]
Dictionnaire Hachette, édition 2009, Paris : Hachette.

[Lemaire 2008]
B. Lemaire. (2008). Limites de la lemmatisation pour l'extraction de significations. *9e Journées internationales d'Analyse Statistique des Données Textuelles*, Lyon, France.

[Léon à paraître]
J. Léon. (à paraître). La traduction automatique I : les premières tentatives jusqu'au rapport ALPAC, in *History of the Language sciences*, Berlin : Walter de Gruyter and co., vol. 3, *Histoire des Sciences du Langage*, pp. 2774-2780.

[Léon 2002]
J. Léon. (2002). Le CNRS et les débuts de la traduction automatique en France, *La Revue pour l'histoire du CNRS*, 6:6-24.

[Leusch et al. 2009]
G. Leusch, E. Matusov & H. Ney. (2009). The RWTH system combination for WMT 2009. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pp. 51-55, Athènes, Grèce.

[Lin et Cherry 2003]
D. Lin & C. Cherry. (2003). Linguistic heuristics in word alignment. In *Proceedings of PACLing-2003*.

[Loffler-Laurian 1994]
A. M. Loffler-Laurian. (1994). La traduction automatique : son utilisation par le grand public. In *Langages*, 116:87-94.

[Martinet 1970]
André Martinet. (1970). *Éléments de linguistique générale.* Paris : Armand Colin

[McEnery et Oakes 1995]
A. M. McEnery & M. P. Oakes. (1995). Sentence and word alignment in the CRATER project : methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop*, Dublin.

[Meillet 1926]
A. Meillet. (1926). *Linguistique historique et linguistique générale*. Vol. 1, Paris : Honoré Champion.

[Melamed 1998]
D. Melamed. (1998). A word-to-word model of translation equivalence.*Technical report IRCS Technical report*, 98:08, Université de Pennsylvanie.

[Nomoto 2004]
T. Nomoto. (2004). Multi-engine machine translation with voted language model. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 494-501, Barcelone, Espagne.

[Och et Ney 2004]
F. J. Och & H. Ney. (2001). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417-449.

[Och & Ney 2001]
F. J. Och & H. Ney. (2001). Statistical multi-source translation. In *Proceedings of MT Summit*, Santiago de Compostela, Spain.

[Ozdowska 2008]
S. Ozdowska. (2008). Cross-corpus evaluation of word alignment. In *Proceedings of the 6th International Conference on Language ressources and Evaluation, LREC'08*, Marrakech, Maroc.

[Ozdowska 2004]
S. Ozdowska. (2004). Appariement de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *Actes des 8èmes Rencontres des Étudiants et Jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues, RECITAL'04*, Fès, Maroc.

[Rastier 2009]
F. Rastier. (2009). *Sémantique interprétative*. Paris : Presses Universitaires de France, coll. « Formes sémiotiques ».

[Rastier 2001]
F. Rastier. (2001). *Sémantique et recherches cognitives*. Paris : PUF.

[Resnik 2007]
P. Resnik. (2007). WSD in NLP applications. In E. Agirre and P. Edmonds. Eds. *Word Sense Disambiguation : Algorithms and Applications*, pp. 299-337, Springer.

[Riegel, Pellat et Rioul 1994]
M. Riegel, J.-C. Pellat & R. Rioul. (1994). *Grammaire méthodique du français*. Paris : Presses Universitaires de France, coll. « Linguistique nouvelle ».

[Sapir 2001]
E. Sapir. (2001). *Le langage. Introduction à l'étude de la parole*. Paris : Petite Bibliothèque Payot.

[Schmid 1995]
H. Schmid. (1994). Probabilistic part-of-speech tagging using Decision Trees. *Actes de la Conférence Internationale sur les Nouvelles Méthodes en Traitement du Langage*, septembre.

[Simard, Foster et Isabelle 1992]
M. Simard, G. Foster & P. Isabelle. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of TMI-92*, Montréal, Canada.

[Schütze 1998]
H. Schütze. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97-123.

[Schwartz 2008]
L. Schwartz. (2008). Multi-source translation methods. In *MT at work : Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pp. 279-288, Waikiki, Hawaï.

[Sparck Jone 1964]
K. Sparck Jones. (1964). Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press.

[Vapnik 1999]
V.N. Vapnik. (1999). *Statistical learning theory*. Wiley Interscience.

[Véronis 2000]
J. Véronis. (2000). From the Rosetta stone to the information society. A survey of parallel text processing. In J. Véronis (éd.), *Parallel Text Processing : alignment and use of translation corpora*, Dordrecht : Kluwer Academic Publishers, pp. 1-24.

[Vickrey et al. 2005]
D. Vickrey, L. Biewald, M. Teyssier and D. Koller. (2005). Word-Sense Disambiguation for Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing* (HLT-EMNLP), Vancouver, Canade, 771-778.

[Weaver 1955]
W. Weaver. (1955). The Mathematic of Information. *Automatic Control*, New York, NY : Simon and Schuster.

[Wilks 2009]
Y. Wilks. (2009). *Machine Translation : its scope and limits*. Springer.

[Wu 2000]
D. Wu. (2000). Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars. In J. Véronis (ed.), *Parallel text processing : alignment and use of translation corpora*, Dordrecht : Kluwer Academic Publishers, pp. 139-167.